A Group Listening Experiment on Authorship Discrimination in AI-Generated and Human-Produced Music Across Three Instrumental Genres

Dr. Joseph Longardner

Sichuan Conservatory of Music
longardnerjj@gmail.com

Abstract

This study examines whether authorship discrimination between AI-generated and human-produced music remains statistically meaningful when listening occurs as a group classroom experiment over shared loudspeakers rather than in controlled, individual headphone sessions. Building on Longardner (2026), undergraduate and graduate-level music students at the Sichuan Conservatory of Music (N = 107) completed a forced-choice authorship task on 30 instrumental excerpts spanning Lo-Fi, Hard Rock, and Swing Jazz. The stimulus set included 18 AI-generated tracks produced with Suno, Udio, and Mureka, as well as 12 human-produced tracks drawn from studio-quality releases. Across 3,210 responses, overall identification accuracy was 48.16%, statistically indistinguishable from chance, indicating that participants could not reliably differentiate AI-generated from human-produced recordings under group listening conditions. These findings suggest that listening context meaningfully constrains perceptual discriminability and that classroom-based AI literacy activities should not assume consistent authorship-identification performance in the absence of controlled playback conditions.
Keywords: AI-Generated Music; Authorship Perception; Classroom Listening; Group Listening; Playback Conditions; Suno; Udio; Mureka

1 INTRODUCTION

Generative artificial intelligence now participates in everyday music production as a compositional aid (İmik, 2026), a mixing and arrangement assistant (Clemens & Marasović, 2025), and, increasingly, as an end-to-end content generator (Cao et al., 2024). As these systems improve, authorship becomes a perceptual judgment as much as a technical question as listeners infer agency, intention, competence, and influence from cues that can also be produced by contemporary models (Sun et al., 2026). Recent work on generative music systems has shown that listeners' authorship judgments are often driven less by explicit knowledge of model capabilities than by inferences drawn from production cues such as mix clarity, timbral consistency, phrasing regularity, and the perceived presence of human intention (Bagratuni et al., 2025; Hernandez-Olivan & Beltran, 2021; van Schaik, 2024; White et al., 2025). As commercially available tools produce increasingly polished outputs (Hasan, 2025), the discriminative task shifts from detecting obvious artefacts to interpreting ambiguous signals (Lecamwasam & Ray, 2017) that can plausibly arise from either human production workflows or contemporary models (Wang, 2025). More broadly, listeners do not rely solely on abstract compositional structure when making sense of music in real time; they also orient themselves through local perceptual cues, especially texture, which can guide expectations, formal navigation, and judgments about musical development (Schwitzgebel, 2025). As generative music systems have become more sophisticated, distinguishing AI-produced music from human-produced music has become increasingly difficult, highlighting the importance of rigorous methodology in studies of perceptual identification, aesthetic judgment, and long-term audience response. Continued research is therefore needed to determine whether repeated exposure reduces initial skepticism, increases acceptance, and reshapes how listeners evaluate music as AI becomes more common in creative practice (White et al., 2025).

One practical question is whether people can still distinguish AI-generated recordings from human-produced recordings in everyday listening contexts. Several studies on AI-generated or AI-attributed music have used controlled or partly controlled listening designs, including laboratory, lab-in-field, and headphone-standardized online experiments (Dunham et al., 2025; Fišer et al., 2025; Shank et al., 2023; Tigre Moura & Maw, 2021; Wu & Holmes, 2026). Perception of musical features such as dynamics, rhythm, timbre, and tonality (Shank et al., 2023) is likely to be more acute in headphone-based sessions conducted in quiet rooms, where playback can be more tightly standardized at the level of individual listening conditions and frequency-response control (Olive, 2025). By contrast, shared loudspeaker playback introduces room reflections, seating-distance effects, frequency-response constraints, and intermittent noise that can collectively mask fine-grained cues. Research on loudspeaker-room-listener interaction further suggests that as the acoustic field becomes more complex, listeners' ability to "listen through" the room is reduced, and imperfections in reproduced sound become harder to detect (Toole, 2025). In applied settings, these factors, including spectral smearing caused by reverberation, can compress perceptual differences that may be detectable under controlled laboratory conditions (Eurich et al., 2019). In everyday settings, music is more often heard over shared loudspeakers in acoustically variable spaces. Room

reflections, seating distance, background noise, and speaker frequency response can obscure micro-timing, articulation, and spectral detail that might otherwise support authorship discrimination (Toole, 2025). Group listening adds a further layer of ecological complexity, as social listening contexts can alter listeners' emotional experience, attentional focus, and mind-wandering, making them especially relevant to classroom and public-facing experimental settings (Egermann et al., 2011; Swarbrick et al., 2024). The implication is both methodological and pedagogical: a discrimination task that performs well under ideal listening conditions may fall to chance in the classroom, even among trained musicians. For AI literacy curricula and assessment design, it is therefore critical to test whether authorship identification remains above chance when evaluation occurs under typical classroom listening conditions rather than idealized individual sessions.

The present study follows Longardner (2026), who evaluated authorship identification under quiet, individual listening conditions using high-quality headphones. While this quiet listening protocol yielded reliable above-chance performance, it also likely established an upper bound on discriminability by minimizing environmental masking. The current experiment intentionally shifts toward greater ecological validity by using a group listening design in standard classrooms, shared loudspeakers, and single-exposure judgments. The core question is whether the perceptual boundary between AI-generated and human-produced instrumental music remains detectable when listening conditions are less than ideal.

1.1 Research Questions

This study addresses four questions concerning group listening conditions and authorship judgements:

1. Listening context: Under classroom speaker playback, do participants identify authorship above chance (0.50) in a forced-choice task?
2. Style dependence: Does identification accuracy vary across the selected styles (Lo-Fi, Hard Rock, and Swing Jazz) when all stimuli are instrumental and production-matched?
3. Error asymmetry: Are errors symmetric, or do participants more often label human-produced tracks as AI-generated than AI-generated tracks as human-produced?
4. Platform differences: Do outputs from Suno, Udio, and Mureka yield different accuracy patterns under group listening conditions?

1.2 Hypotheses

Grounded in the role of playback conditions in auditory perception, the study advances four hypotheses:

1. H1: Overall accuracy under classroom speaker playback will be statistically lower, or indistinguishable, from chance and lower than performance reported under controlled headphone sessions.
2. H2: Style-level differences will be modest because room acoustics and speaker playback attenuate fine-grained timbral and timing cues across all styles.
3. H3: Errors will be asymmetric, with a higher rate of labeling human-produced tracks as AI-generated than labeling AI-generated tracks as human-produced.
4. H4: Between-platform differences will be small relative to the overall chance-level baseline, although track-level variability may be higher for some models.

As AI-generated music becomes more common in contemporary listening culture, authorship identification must be understood not only as a technical issue but also as a context-dependent perceptual one. By examining judgment under shared classroom playback rather than ideal headphone conditions, this study extends current research toward greater ecological validity and pedagogical relevance. Its broader significance lies in clarifying how the listening environment shapes discrimination, and in helping educators and researchers better understand how AI-generated music is perceived in real-world settings.

2 METHODOLOGY

2.1 Participants and Setting

The participant cohort consisted of 107 undergraduate and graduate-level students from the Sichuan Conservatory of Music (Chengdu, China), drawn primarily from jazz and popular music programs. All participants had passed the Chinese Ministry of Education's Yikao (Arts College Entrance Examination), which certifies "robust musical competency" (Lin & Weatherly, 2024). All listening took place during regular class sessions on the Xindu campus between September 22 and 26, 2025. Compared to the Longardner (2026) study, there was no overlap of participants on the individual level. All participants were selected from different classes, but at the same school, all within the jazz and popular music programs.

2.1.1 Participant Demographics and Background

The sample was aged 19 to 25 and reflects the cohort distribution typical of the participating programs. Participants were 19 years old (19.63%, n = 21), 20 (39.25%, n = 40), 21 (31.78%, n = 34), 22 (5.61%, n = 6), and 23 to 25 (3.74%, n = 4). Gender distribution was 56.07% male (n = 60) and 43.93% female (n = 47). Participants were also asked about perfect pitch, but no participants reported having perfect pitch, and this variable was not analyzed further. Graphs for this section are omitted by design.

2.2 Stimuli and Materials

Audio samples were generated using three AI platforms: Suno, Udio, and Mureka, selected because they are among the most widely discussed and frequently analyzed AI music-generation systems in recent academic literature (Bown, 2025; Casini et al., 2025; Fiorino, 2025; Nayar, 2025; Rahman et al., 2024; Riedl, 2025; Salganik et al., 2026; Solak et al., 2025; Vila et al., 2025; White et al., 2025).

The evaluation set comprised 30 instrumental excerpts spanning three styles: Lo-Fi, Hard Rock, and Swing Jazz, and all AI-generated audio was downloaded between September 19 and 21, 2025, to ensure consistency and reliability. The stimuli included 18 AI-generated tracks, with six from each platform, created using Suno v4.5+, Udio v1.5 Allegro, and Mureka O1, as well as 12 human-produced tracks, with four from each style, sourced from studio-quality releases. These styles were selected to balance everyday exposure (Lo-Fi), genre-specific

production texture (Hard Rock), and formal training relevance for the participant cohort (Swing Jazz).

2.2.1 Recording and Output Standardization

All audio was captured using loopback recording to standardize extraction across platforms, as informed by Berger & Neitsch (2023), and then exported as stereo MP3 files (44.1 kHz, 320 kbps) without changes to gain, dynamics, panning, compression, noise reduction, or normalization. Each track was edited to a 60-second excerpt with a 2-second fade-in and a 5-second fade-out to focus attention on local timbre, groove, and production characteristics rather than large-scale formal development.

2.2.2 AI Model Configurations

All AI generations were configured as instrumental outputs to avoid linguistic cues and concentrate attention on production texture and timbral realism. Within each platform, settings were held constant to reduce parameter-driven variance. Suno generations used the instrumental mode with no inspiration features. Udio generations used instrumental mode with prompt strength fixed at 50% and clarity fixed at 25% (Ultra quality). Mureka generations used the O1 model without reference-song upload, lyric optimization, or melody-idea assistance.

2.2.3 Human-Produced Music Selection and Prompt Matching

Since Spotify's recommendation system may occasionally use AI-generated material within curated playlists (Kratochvil, 2025), each selected track underwent an additional verification process to confirm human authorship. For every track in the dataset, the author recorded the title, artist, album, and corresponding Spotify playlist, as listed in Appendix A. A total of nine playlists were used, chosen for their stylistic fit with the six genres examined in this study and for their low likelihood of familiarity among participants. Highly recognizable or canonized recordings were removed in order to limit recognition bias. Once the human reference tracks had been finalized, AI prompts were developed to approximate the instrumentation, groove profile, harmonic density, and production texture of the three styles under investigation.

2.3 Procedure: Group Classroom Listening

Participants completed a blind, forced-choice authorship identification task (AI-generated vs. human-produced) for each excerpt. Playback occurred in standard classrooms using shared speakers (Medsome R10000TC) and VLC Media Player with all audio enhancements disabled. Peer discussion was not permitted during playback, and re-listening was not allowed. Participants recorded responses on printed questionnaires, ensuring that each judgement reflected a single exposure under typical classroom acoustics.

2.4 Methodology for Statistical Analysis

All analyses were conducted at the trial level, with participants making a binary judgement about whether each excerpt was AI-generated or human-produced. The outcome measure was identification accuracy, defined as the proportion of correct responses within prespecified analytical cells, including production category, model, genre, and participant subgroup. Accuracy was reported with two-sided 95% binomial confidence intervals and evaluated against a 0.50 chance benchmark. Within-cell departures from chance were tested using two-sided exact one-sample binomial tests. Relationships between accuracy, model, and genre were assessed with Pearson chi-square tests of independence on contingency tables, and Cramér's V was reported as the associated effect size. Error patterns were examined through confusion matrices that distinguished AI→Human from Human→AI misattributions by genre. All tests used an alpha level of 0.05, and the figures and tables report counts, proportions, confidence intervals, and p-value annotations derived from these procedures alone. All statistical analyses were carried out in Python.

3 RESULTS

Figures 1 to 6 and Tables 1 to 9 summarize aggregate performance, style-level patterns, model-specific results, misattribution direction, song-level variability, and pooled participant-level dispersion. Across the dataset, overall identification remained close to chance in several broad categories, but meaningful variation emerged across production methods, styles, and individual excerpts.

3.1 Aggregate Results by Production Category

Fig. 1 and Tab. 1 present the aggregate identification results by production category. Human-produced excerpts were identified below the 0.50 benchmark, whereas pooled AI remained near chance overall. Among the three AI systems, Udio yielded the highest aggregate identification accuracy, Suno fell below chance, and Mureka remained statistically indistinguishable from chance.

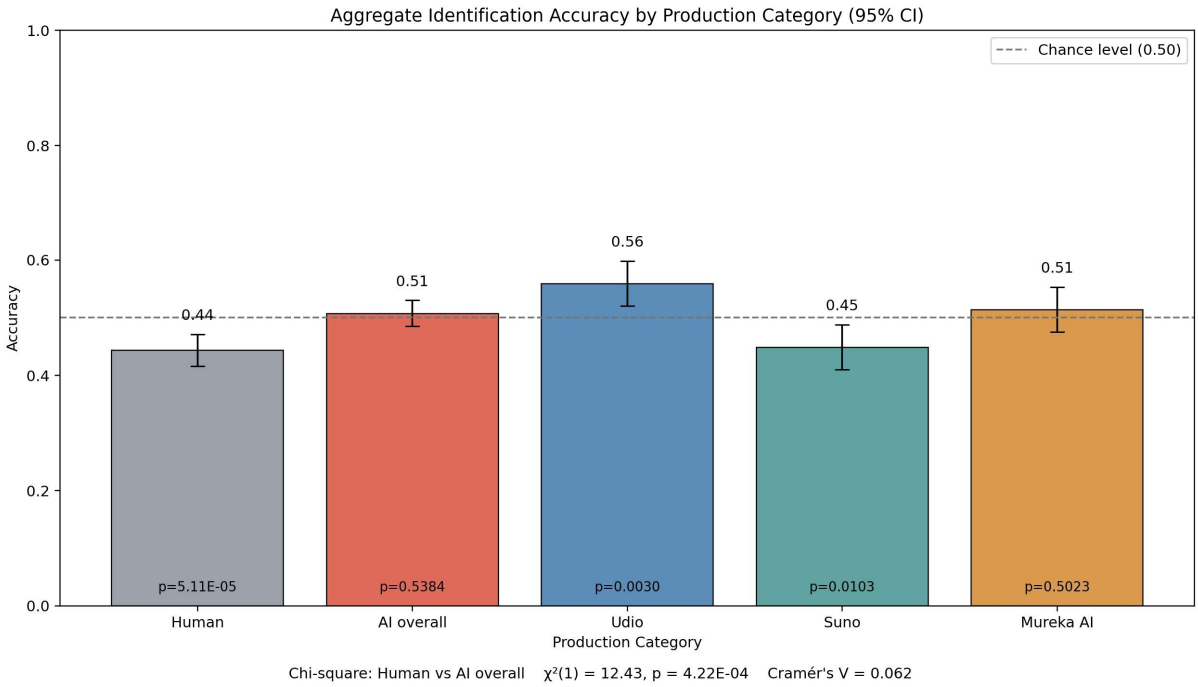


Figure 1: Aggregate identification accuracy by production category.

Table 1: Aggregate identification accuracy by production category.

Production Method	Correct	Total	Accuracy	95% CI	p-value (vs 0.50)
Human	569	1284	44.31%	[0.416, 0.471]	5.11E-05
AI overall	977	1926	50.73%	[0.485, 0.530]	0.5384
Udio	359	642	55.92%	[0.520, 0.598]	0.0030
Suno	288	642	44.86%	[0.410, 0.488]	0.0103
Mureka AI	330	642	51.40%	[0.475, 0.553]	0.5023
Chi-square: Human vs AI overall $\chi^2(1) = 12.43$, p = 4.22E-04, Cramér's V = 0.062.					

3.2 Broad Overview of Results by Music Genre

Fig. 2 and Tab. 2 show an overview of results by music genre. Lo-Fi remained statistically indistinguishable from chance, Hard Rock fell below chance with a reliable deviation, and Swing Jazz was likewise near chance. At the broad style level, no genre produced a robust above-chance identification advantage under the classroom listening condition.

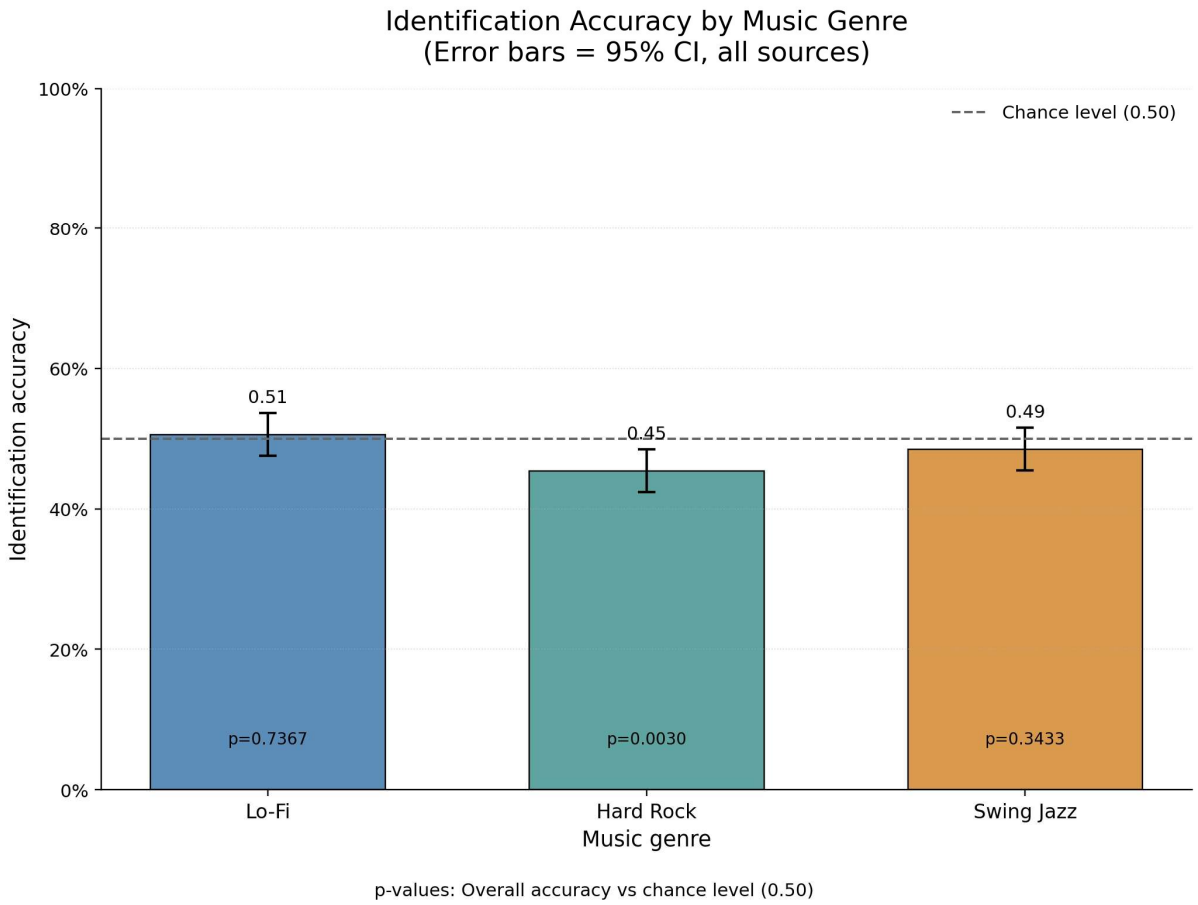


Figure 2: Identification accuracy by music genre.

Table 2: Detailed results by music genre.

Style	N	Correct	Incorrect	Accuracy	95% CI	p-value
Lo-Fi	1070	541	529	0.5056	[0.475, 0.536]	0.7367
Hard Rock	1070	486	584	0.4542	[0.424, 0.485]	0.0030
Swing Jazz	1070	519	551	0.4850	[0.455, 0.515]	0.3433

3.3 Identification Accuracy by Style and Production Method

Fig. 3 and Tab. 3 disaggregate identification accuracy by style and production method. Model-specific differences were evident within styles. In Lo-Fi, Mureka produced the highest accuracy, while Udio and Suno both fell clearly below chance. In Hard Rock, all AI systems clustered near chance, whereas Human-produced excerpts were identified least accurately. In Swing Jazz, Udio showed the weakest performance, while Suno, Mureka, and Human Produced remained close to chance. The model × genre interaction was statistically significant, indicating that performance varied across style-platform combinations rather than by production method alone.

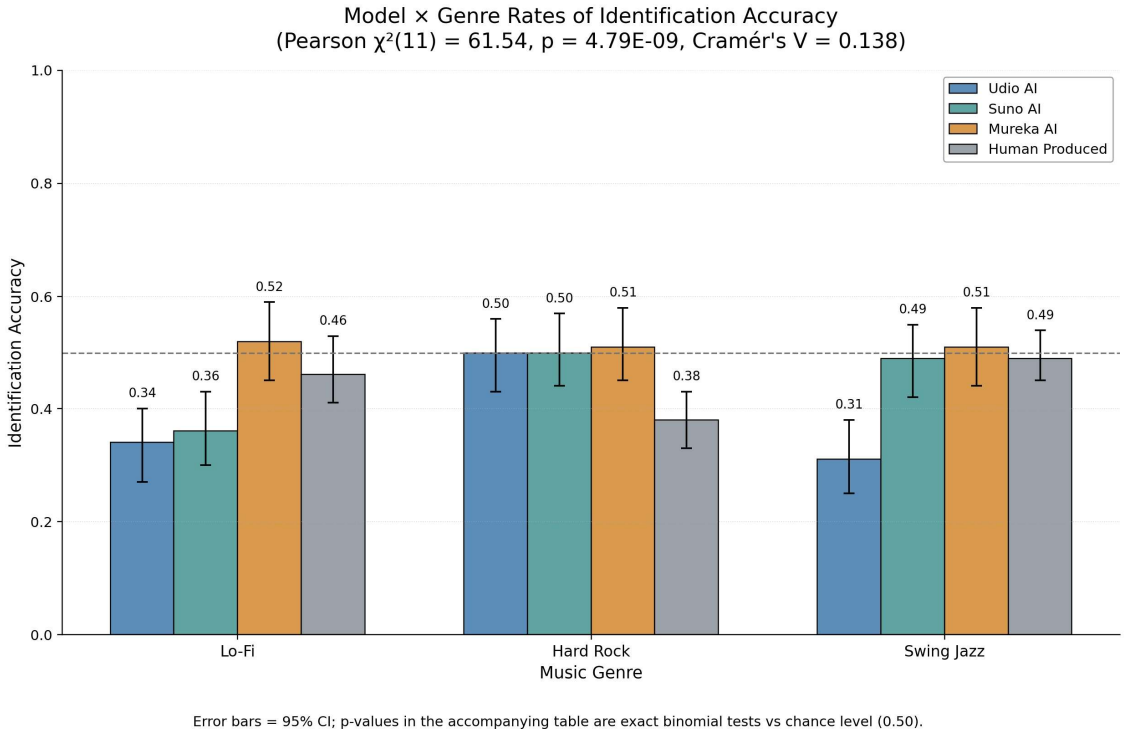


Figure 3: Model × genre rates of identification accuracy.

Table 3: Detailed results by genre and production method.

Udio AI						
Genre	N	Correct	Incorrect	Accuracy	95% CI	p-value
Lo-Fi	214	72	142	33.64%	[0.273, 0.404]	1.96E-06
Hard Rock	214	106	108	49.53%	[0.426, 0.564]	0.9455
Swing Jazz	214	67	147	31.31%	[0.252, 0.380]	4.69E-08
Suno AI						
Genre	N	Correct	Incorrect	Accuracy	95% CI	p-value
Lo-Fi	214	77	137	35.98%	[0.296, 0.428]	4.95E-05
Hard Rock	214	107	107	50%	[0.431, 0.569]	1.0000
Swing Jazz	214	104	110	48.60%	[0.417, 0.555]	0.7326
Mureka AI						
Genre	N	Correct	Incorrect	Accuracy	95% CI	p-value
Lo-Fi	214	111	103	51.87%	[0.450, 0.587]	0.6324
Hard Rock	214	110	104	51.40%	[0.445, 0.583]	0.7326
Swing Jazz	214	109	105	50.93%	[0.440, 0.578]	0.8376
Human Produced						
Genre	N	Correct	Incorrect	Accuracy	95% CI	p-value
Lo-Fi	428	195	233	45.56%	[0.408, 0.504]	0.0736
Hard Rock	428	163	265	38.08%	[0.335, 0.429]	9.38E-07
Swing Jazz	428	211	217	49.30%	[0.445, 0.541]	0.8091

Note: p-values are exact binomial tests of overall accuracy vs chance level (0.50); 95% CI are exact binomial confidence intervals. Model × Genre Interaction in Identification Accuracy: Pearson $\chi^2(11) = 61.54$, p = 4.79E-09, Cramér's V = 0.138.

3.4 Misattribution Direction by Genre

Fig. 4 and Tab. 4 examine the direction of misattribution by genre. When expressed as rates relative to source-specific totals, Human → AI misattributions exceeded AI → Human in Lo-Fi and Hard Rock, whereas Swing Jazz showed a slight reversal, with AI → Human marginally higher. In absolute counts, however, AI → Human remained numerically larger across all three styles because the pooled AI pool contained more items than the human pool.

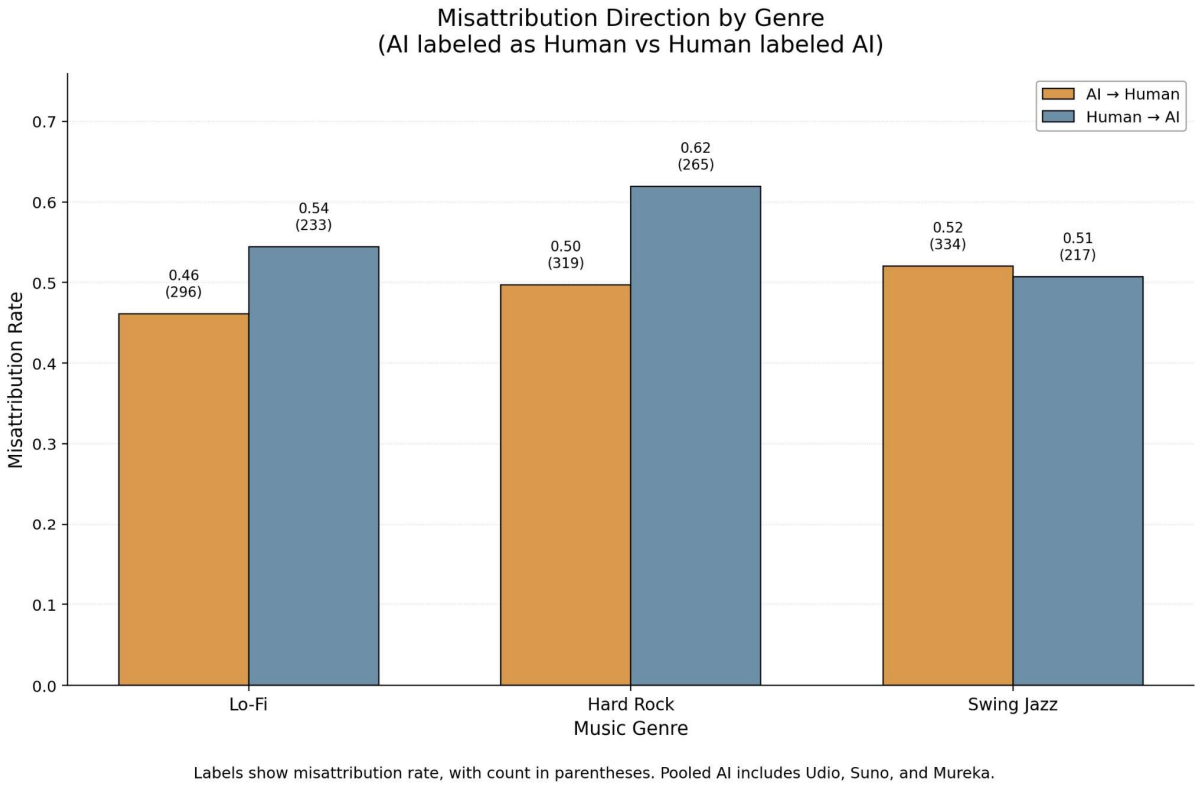


Figure 4: Misattribution direction by genre.

Table 4: Detailed results of misattribution direction by genre.

Genre	AI→AI (correct)	AI→ Human	Human→Human (correct)	Human→AI	Delta (AI→H minus H→AI)	Total	χ^2	p-value	Cramér's V
Lo-Fi	346	296	195	233	63	1070	6.8047	0.0091	0.0797
Hard Rock	323	319	163	265	54	1070	14.9982	1.08E-04	0.1184
Swing Jazz	308	334	211	217	117	1070	0.1311	0.7173	0.0111
TOTAL	977	949	569	715	234	3210			

3.5 Song-Level Identification Accuracy by Style and Production Method

Fig. 5 and Tabs. 5 to 7 extend the analysis to the individual-song level. Considerable within-style variability was evident. Lo-Fi ranged from 0.29 to 0.80, Hard Rock from 0.34 to 0.53, and Swing Jazz from 0.26 to 0.63. Style-level summaries, therefore, masked meaningful track-level spread, especially in Lo-Fi and Swing Jazz, where both highly confusable and relatively discriminable excerpts appeared within the same style. At the song-summary level, Udio produced the highest mean of song means, while Human Produced yielded the lowest.

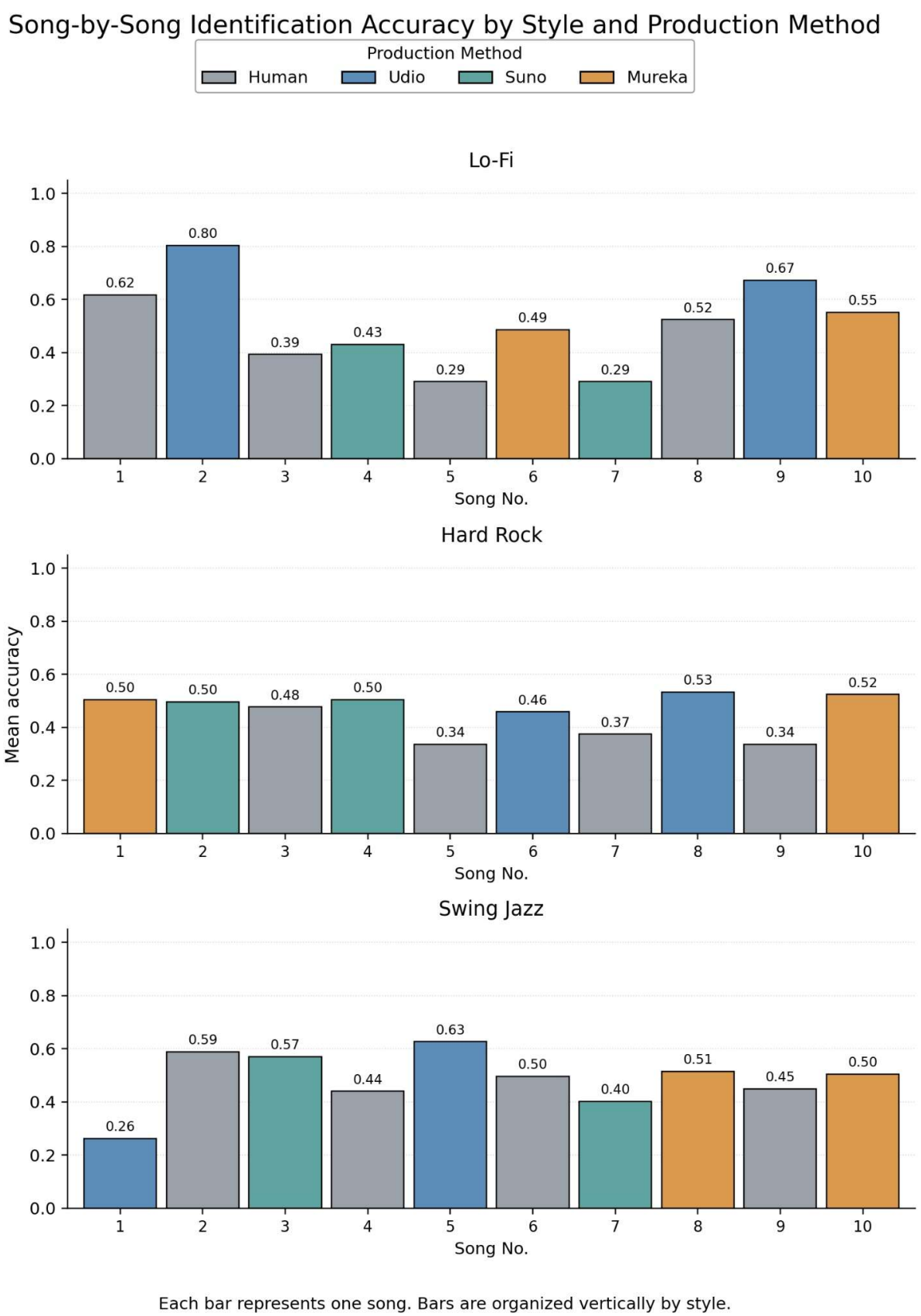


Figure 5: Song-by-song identification accuracy by style and production method.

Table 5: Song-by-Song SD.

Style	Song No	Production Method	Song Label	Participants	Correct	Incorrect	Mean Accuracy	SD	Variance	SE
Lo-Fi	1	Human	Human – Socially Awkward - Kiefer	107	66	41	61.68%	0.4884	0.2386	0.0472
Lo-Fi	2	Udio	Udio – Digital Jazz Vibes	107	86	21	80.37%	0.3990	0.1592	0.0386
Lo-Fi	3	Human	Human – Retainer - Alfa Mist	107	42	65	39.25%	0.4906	0.2407	0.0474
Lo-Fi	4	Suno	Suno – City Lights, Neon Nights	107	46	61	42.99%	0.4974	0.2474	0.0481
Lo-Fi	5	Human	Human – Of Course That's Happening	107	31	76	28.97%	0.4558	0.2077	0.0441
Lo-Fi	6	Mureka	Mureka Lo-Fi 2	107	52	55	48.60%	0.5022	0.2522	0.0485
Lo-Fi	7	Suno	Suno – City Lights Serenade	107	31	76	28.97%	0.4558	0.2077	0.0441
Lo-Fi	8	Human	Human – Makaya McCraven - Boom Bapped	107	56	51	52.34%	0.5018	0.2518	0.0485
Lo-Fi	9	Udio	Udio – Rhythmic Improv	107	72	35	67.29%	0.4714	0.2222	0.0456
Lo-Fi	10	Mureka	Mureka Lo-Fi 3	107	59	48	55.14%	0.4997	0.2497	0.0483
Hard Rock	1	Mureka	Murerka – Rock 3	107	54	53	50.47%	0.5023	0.2523	0.0486
Hard Rock	2	Suno	Suno – Fracture Glass	107	53	54	49.53%	0.5023	0.2523	0.0486
Hard Rock	3	Human	Human – Enough This Time - Rockin' For Decades	107	51	56	47.66%	0.5018	0.2518	0.0485
Hard Rock	4	Suno	Suno – Shatter the Silence	107	54	53	50.47%	0.5023	0.2523	0.0486
Hard Rock	5	Human	Human – My Mother Raised a Fool - Deathkite	107	36	71	33.64%	0.4747	0.2254	0.0459
Hard Rock	6	Udio	Udio – Rhythmic Fury	107	49	58	45.79%	0.5006	0.2506	0.0484
Hard Rock	7	Human	Human – Never Get to Me - Deaf Election	107	40	67	37.38%	0.4861	0.2363	0.0470
Hard Rock	8	Udio	Udio – Riff Revolution	107	57	50	53.27%	0.5013	0.2513	0.0485
Hard Rock	9	Human	Human – Riding the Storm - Deaf Election	107	36	71	33.64%	0.4747	0.2254	0.0459
Hard Rock	10	Mureka	Murerka – Rock 4	107	56	51	52.34%	0.5018	0.2518	0.0485
Swing Jazz	1	Udio	Udio – Deep Swing Vibes	107	28	79	26.17%	0.4416	0.1950	0.0427
Swing Jazz	2	Human	Human – Baby Soda Jazz Band – Swing That Music	107	63	44	58.88%	0.4944	0.2444	0.0478
Swing Jazz	3	Suno	Suno – Swingin' On a Whisper	107	61	46	57.01%	0.4974	0.2474	0.0481
Swing Jazz	4	Human	Human – Bright Lights Late Nights - the Speakeasies' Swing Band!	107	47	60	43.93%	0.4986	0.2486	0.0482
Swing Jazz	5	Udio	Udio – Swingin' Deep	107	67	40	62.62%	0.4861	0.2363	0.0470
Swing Jazz	6	Human	Human – Diga Diga Doo - Odeur de Swing	107	53	54	49.53%	0.5023	0.2523	0.0486
Swing Jazz	7	Suno	Suno – Swingin' Through the Night	107	43	64	40.19%	0.4926	0.2426	0.0476
Swing Jazz	8	Mureka	Murerka – Jazz 1	107	55	52	51.40%	0.5022	0.2522	0.0485
Swing Jazz	9	Human	Human – Lizzy & the Triggermen - Someday	107	48	59	44.86%	0.4997	0.2497	0.0483
Swing Jazz	10	Mureka	Murerka – Jazz 2	107	54	53	50.47%	0.5023	0.2523	0.0486

Table 6: Style Summaries.

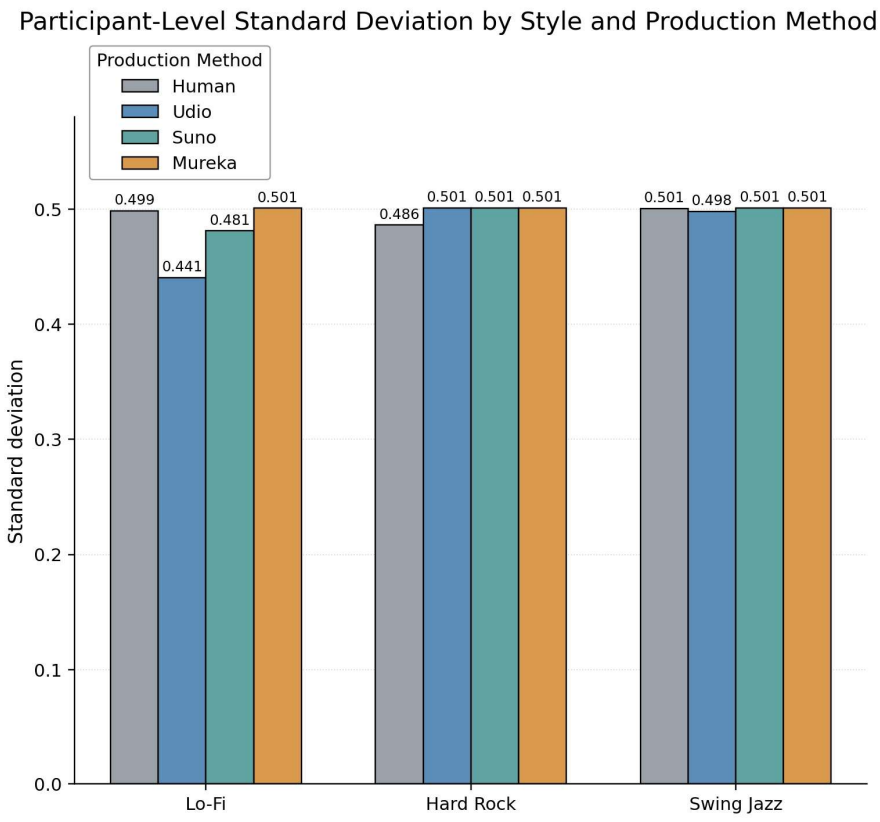
Style	Songs	Mean of Song Means	Mean SD Across Songs	Min Song Accuracy	Max Song Accuracy
Lo-Fi	10	50.56%	47.62%	28.97%	80.37%
Hard Rock	10	45.42%	49.48%	33.64%	53.27%
Swing Jazz	10	48.50%	49.17%	26.17%	62.62%

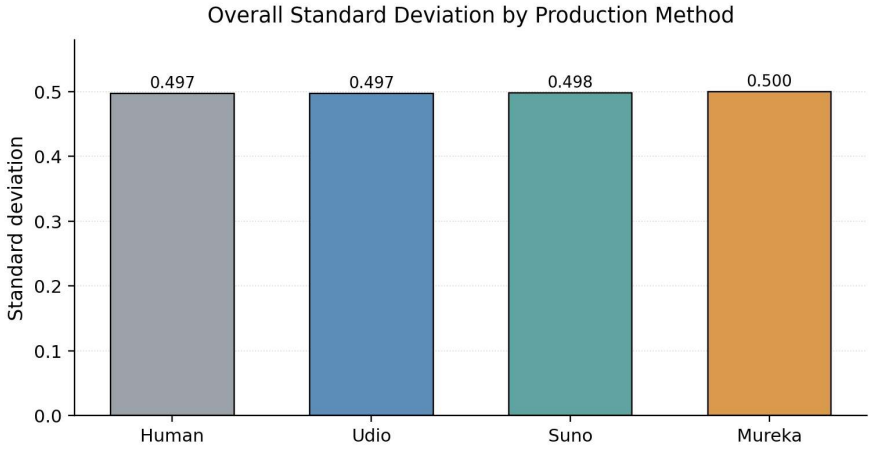
Table 7: Method Summaries.

Production Method	Songs	Mean of Song Means	Mean SD Across Songs	Min Song Accuracy	Max Song Accuracy
Human	12	44.31%	48.91%	28.97%	61.68%
Udio	6	55.92%	46.67%	26.17%	80.37%
Suno	6	44.86%	49.13%	28.97%	57.01%
Mureka	6	51.40%	50.17%	48.60%	55.14%

3.6 Participant-Level Standard Deviation by Style and Production Method

Fig. 6 and Tabs. 8 to 9 summarize participant-level standard deviations pooled within style × production-method groups and by production method overall. Dispersion was remarkably stable across most groups, clustering near 0.50, which is consistent with highly mixed binary response patterns under near-chance conditions. The clearest exception appeared in Lo-Fi Udio, where responses were modestly less dispersed because identification accuracy was higher than in most other groups. At the overall level, all four production methods remained tightly grouped, reinforcing the interpretation that the classroom listening condition produced broadly unstable, near-balanced response behavior across methods.





Standard deviations are computed from pooled participant 0/1 responses within each requested group.

Table 8: By Style and Method.

Style	Production Method	Songs in Group	Trials	Correct	Incorrect	Mean Accuracy	SD	Variance	SE
Lo-Fi	Human	4	428	195	233	45.56%	0.4986	0.2486	0.0241
Lo-Fi	Udio	2	214	158	56	73.83%	0.4406	0.1941	0.0301
Lo-Fi	Suno	2	214	77	137	35.98%	0.4811	0.2314	0.0329
Lo-Fi	Mureka	2	214	111	103	51.87%	0.5008	0.2508	0.0342
Hard Rock	Human	4	428	163	265	38.08%	0.4862	0.2364	0.0235
Hard Rock	Udio	2	214	106	108	49.53%	0.5012	0.2512	0.0343
Hard Rock	Suno	2	214	107	107	50.00%	0.5012	0.2512	0.0343
Hard Rock	Mureka	2	214	110	104	51.40%	0.5010	0.2510	0.0342
Swing Jazz	Human	4	428	211	217	49.30%	0.5005	0.2505	0.0242
Swing Jazz	Udio	2	214	95	119	44.39%	0.4980	0.2480	0.0340
Swing Jazz	Suno	2	214	104	110	48.60%	0.5010	0.2510	0.0342
Swing Jazz	Mureka	2	214	109	105	50.93%	0.5011	0.2511	0.0343

Table 9: By Production Overall.

Production Method	Songs in Group	Styles Covered	Trials	Correct	Incorrect	Mean Accuracy	SD	Variance	SE
Human	12	3	1284	569	715	44.31%	0.4970	0.2470	0.0139
Udio	6	3	642	359	283	55.92%	0.4969	0.2469	0.0196
Suno	6	3	642	288	354	44.86%	0.4977	0.2477	0.0196
Mureka	6	3	642	330	312	51.40%	0.5002	0.2502	0.0197

4 DISCUSSION

4.1 Listening Context and the Limits of Discrimination

The primary finding is that authorship discrimination collapsed to chance under group classroom playback. In contrast to the headphone-based listening sessions reported in Longardner (2026), the present protocol exposed participants to something closer to real-world conditions: shared speakers, spectral smearing caused by reverberation, and variable listening perspectives based on position in the space. These factors plausibly reduce access to micro-timing deviations, transient detail, and subtle spectral artefacts that might otherwise inform authorship judgments. The result is not simply a reflection of participant ability, but it suggests that the conditions of listening play a central role in whether listeners can meaningfully distinguish between AI-generated and human-produced music.

4.1.1 Comparison to the 2026 study

Methodologically, the present experiment is stronger than Longardner (2026) in ecological validity because it tests authorship discrimination under ordinary classroom speaker playback, single-exposure listening, and a contemporary three-platform comparison that includes Mureka alongside Suno and Udio, thereby capturing a more realistic pedagogical setting than the earlier headphone-based design. At the same time, it is weaker in internal control and inferential precision, as the published study used quieter and more standardized listening conditions, a larger and more stylistically diverse musical corpus, broader genre coverage, and a design better suited to isolating perceptual differences from playback-related masking. Statistically, the current study remains rigorous at the trial level. Although the number of songs was smaller than in the earlier experiment, the design yielded many more responses per song and therefore a greater concentration of observations at the excerpt level, strengthening the statistical basis of the analysis. This also permits analysis of standard deviation at the track level, which helps assess the stability of individual songs and compare outputs generated from similar prompts across different models. By contrast, in Longardner (2026), each track was heard only once by a single participant, which limited track-level comparison and reduced the possibility of evaluating variability across parallel model outputs. Nevertheless, the near-chance classroom results should be interpreted as context-dependent rather than as a direct replacement for the earlier findings, because uncontrolled acoustic variation, fixed classroom playback, and the absence of a within-sample headphone comparison limit the extent to which differences can be attributed solely to the stimuli themselves.

4.1.2 Cramér's V Indicates Modest Practical Effects Across Key Comparisons

Across the present study, the Cramér's V values indicate that the statistically significant associations were generally small in a practical scale. The model × genre interaction reached significance, but the corresponding effect size remained modest (Cramér's V = 0.138), suggesting that differences across platforms and styles were detectable yet not strong enough to indicate clear separation among conditions. A similar pattern appears in the misattribution analysis: Lo-Fi (V = 0.0797) and Hard Rock (V = 0.1184) show only weak asymmetries in error direction, while Swing Jazz was effectively negligible (V = 0.0111). Taken together, these values suggest that although some response patterns were statistically reliable, the underlying associations were small, consistent with a listening context in which participants' judgements remained highly mixed and only weakly structured by platform or style.

4.2 Implications for Classroom AI Literacy and Assessment

For teaching, the findings caution against classroom exercises that assume students can reliably "hear the difference" between AI-generated and human-produced recordings under ordinary playback. If discrimination accuracy is near 50% in speaker-based group listening, then identification tasks may measure room acoustics, attention, and playback quality as much as they measure perceptual skill. Classroom AI literacy may therefore be better served by (a) emphasizing process transparency and documentation, (b) teaching critical listening for production decisions rather than binary authorship classification, and (c) using controlled headphone playback when an identification task is pedagogically required.

4.3 Platform Parity and Cross-System Convergence

Mureka performed comparably to Suno and Udio at the aggregate level, suggesting perceptual convergence between a Chinese platform and widely used Western systems for the instrumental styles examined under the conditions of this study. This does not imply identical training data or inductive biases, but it does suggest that multiple large-scale systems can now produce short, polished instrumental outputs that approach the perceptual boundary of studio recordings in everyday listening contexts. However, Mureka remains less extensively studied than Suno and Udio, and additional experiments conducted under different listening conditions and across other stylistic settings are necessary before concluding that the system achieves consistent parity with these more established AI models.

4.4 Limitations

The study is limited by its homogeneous participant pool, drawn from a single conservatory, the use of short instrumental excerpts, and the lack of a within-sample

headphone comparison condition. Classroom acoustics were not formally measured, and seating position was not recorded, limiting fine-grained inference about which acoustic variables most strongly predict performance. The study also did not include a qualitative component, such as a survey or interview, that might have captured more nuanced judgements, interpretive reasoning, and feature-level perceptions. Finally, the forced-choice design may compress these subtleties into a binary response format, masking partial confidence and more complex evaluative responses.

4.4.1 Study Fatigue and Cognitive Load of Participants

The author observed that the 30-minute experiment introduced cognitive fatigue during the listening session. Participants' focus appeared to decrease as they progressed through the excerpts, which may have reduced identification accuracy in the later stages. However, Swing Jazz, which was always presented last, did not show lower overall identification accuracy than Lo-Fi or Hard Rock, apart from a slightly different pattern in misattribution direction by genre that brought both directions closer to chance. This relative consistency suggests that fatigue may not have been a major confounding factor. Future research could reduce potential cognitive-load effects by varying genre order during the experiment and shortening excerpts to approximately 45 seconds in order to sustain engagement and accuracy.

4.4.2 Temporal Separation and Potential Model Improvement Between AI Generations

Because the AI-generated tracks in Longardner (2026) were produced in February 2025, whereas the tracks in the present study were generated in September 2025, it cannot be fully ruled out that improved model capabilities contributed to the near-chance results. This separation complicates direct comparison between the two studies, because any reduction in discriminability may reflect not only differences in playback conditions but also advances in generation quality across contemporary AI systems. Although the present results remain statistically significant, the rapid pace of AI model development means that it is plausible that later-generation outputs would have produced similarly low discrimination rates even under headphone-based listening conditions. The current design, therefore, cannot isolate playback context from temporal improvements in model performance, and this possibility should be treated as an important limitation when comparing the present findings with those of Longardner (2026).

4.5 Future Directions

Future work should directly compare playback conditions within the same participant cohort (headphones vs. speakers) and record seating location to model acoustic and attentional effects. Longer excerpts, vocal materials in multiple languages, and re-listening conditions may also change discriminability. More generally, experimental designs that combine perceptual judgements with feature-level prompts (e.g., timing realism, instrumental articulation, mix balance) may better capture how listeners form authorship judgments under varied listening conditions.

5 CONCLUSION

In a group classroom listening setting using shared speakers, undergraduate and graduate-level music students could not reliably distinguish AI-generated from human-produced instrumental music. Overall accuracy was near chance, style effects were modest, and platform differences were small relative to the baseline. Compared with the controlled headphone protocol reported in Longardner (2026), these findings indicate that less-than-ideal playback conditions can erase perceptual separability. For pedagogy and policy, the implication is straightforward: classroom demonstrations should not treat authorship identification as a stable skill unless listening conditions are explicitly controlled.

AUTHOR DECLARATIONS

This paper was developed with the assistance of generative artificial intelligence tools, including ChatGPT 5.4 and Grammarly, for brainstorming, language revision, proofreading, and the generation of Python code that assisted with data analysis, graph, and table creation.

APPENDIX A: DATA AND MATERIALS

All data, basic analyses, and tracks are available at this link: https://drive.google.com/file/d/1aQgNIQnHyIx2-ar1W3LRrbglecGdMeCH/view?usp=sharing

APPENDIX B: PRE-LISTENING PARTICIPANT BACKGROUND QUESTIONNAIRE (TRANSLATED FROM CHINESE)

Age: _____
Gender:
Male
Female
Do you have perfect pitch?
Yes
No

REFERENCES

Bagratuni, M., Müller, P., & Planing, P. (2025). Innovation in tune: An empirical investigation of user acceptance of artificial intelligence-generated music. Computers in Human Behavior Reports, 18, 100660. https://doi.org/10.1016/j.chbr.2025.100660

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

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

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P., & Sun, L. (2024). A survey of AI-generated content (AIGC). ACM Computing Surveys, 57(5). https://doi.org/10.1145/3704262

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

Clemens, M., & Marasović, A. (2025). MixAssist: An audio-language dataset for co-creative AI assistance in music mixing. ArXiv.org. https://doi.org/10.48550/arXiv.2507.06329

Dunham, R. L., van Kleef, G. A., & Stamkou, E. (2025). The threat of synthetic harmony: The effects of AI vs. human origin beliefs on listeners' cognitive, emotional, and physiological responses to music. Computers in Human Behavior: Artificial Humans, 6, 100205. https://doi.org/10.1016/j.chbah.2025.100205

Egermann, H., Sutherland, M. E., Grewe, O., Nagel, F., Kopiez, R., & Altenmüller, E. (2011). Does music listening in a social context alter experience? A physiological and psychological perspective on emotion. Musicae Scientiae, 15(3), 307-323. https://doi.org/10.1177/1029864911399497

Eurich, B., Klenzner, T., & Oehler, M. (2019). Impact of room acoustic parameters on speech and music perception among participants with cochlear implants. Hearing Research, 377, 122-132. https://doi.org/10.1016/j.heares.2019.03.012

Fišer, N., Martín-Pascual, M. Á., & Andreu-Sánchez, C. (2025). Emotional impact of AI-generated vs. human-composed music in audiovisual media: A biometric and self-report study. PLOS ONE, 20(6), e0326498. https://doi.org/10.1371/journal.pone.0326498

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/academic/grade/thesis/Tesis_de_Licenciatura_-_Santiago_Fiorino.pdf

Frid, E., Gomes, C., & Jin, Z. (2020). Music creation by example. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3313831.3376514

Hasan, M. R. (2025). The impact of artificial intelligence on music composition and performance. International Journal of Technology Management & Humanities, 11(1). https://doi.org/10.21590/ijtmh.2025.v11.i01.02

Hernandez-Olivan, C., & Beltran, J. R. (2021, August 27). Music composition with deep learning: A review [Preprint]. arXiv. https://arxiv.org/abs/2108.12290

İmik, O. H. (2026). AI-assisted music production and the transformation of production processes: Possibilities and limitations. Turkish Online Journal of Design Art and Communication, 16(2), 1226-1237. https://doi.org/10.7456/tojdac.1858350

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

Lecamwasam, K., & Ray, T. (2017). Exploring listeners' perceptions of AI-generated and human-composed music for functional emotional applications. ArXiv.org. https://arxiv.org/html/2506.02856v1

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

Longardner, J. (2026). A multi-genre study of identification and style bias of AI-generated music. Journal of Creative Music Systems, 10(1). https://doi.org/10.5920/jcms.1704

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

Olive, S. (2025). Headphones: The perception and measurement of their sound quality. In Sound Reproduction (pp. 469-525). Focal Press. https://doi.org/10.4324/9781003477495-16

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

Salganik, R., Tu, T., Chen, F.-Y., Liu, X., Lu, K., Luvisia, E., Duan, Z., Salha-Galvan, G., Kahng, A., Ma, Y., & Kang, J. (2026). MusicSem: A semantically rich language-audio dataset of natural music descriptions. ArXiv.org. https://doi.org/10.48550/arXiv.2602.17769

Schwitzgebel, E. (2025). "Cueing" your playlist: Texture and teleology in post-millennial pop. Music Theory Online, 31(1). https://doi.org/10.30535/mto.31.1.5

Shank, D. B., Stefanik, C., Stuhlsatz, C., Kacirek, K., & Belfi, A. M. (2022). AI composer bias: Listeners like music less when they think it was composed by an AI. Journal of Experimental Psychology: Applied, 29(3). https://doi.org/10.1037/xap0000447.supp

Solak, A., Grötschla, F., Wattenhofer, R., & Lanzendörfer, L. A. (2025). Bias beyond borders: Global inequalities in AI-generated music. ArXiv.org. https://doi.org/10.48550/arXiv.2510.01963

Sun, Y., Fan, Z., Sheng, D., Zhang, X., & Xu, F. (2026). Interactive perceptions and persuasion: How AI-generated content influences users. The Electronic Library, 44(1), 109-135. https://doi.org/10.1108/el-05-2025-0172

Swarbrick, D., Martin, R., Høffding, S., Nielsen, N., & Vuoskoski, J. K. (2024). Audience musical absorption: Exploring attention and affect in the live concert setting. Music & Science, 7. https://doi.org/10.1177/20592043241263461

Tigre Moura, F., & Maw, C. (2021). Artificial intelligence became Beethoven: How do listeners and music professionals perceive artificially composed music? Journal of Consumer Marketing, 38(2), 137-146. https://doi.org/10.1108/jcm-02-2020-3671

Toole, F. (2025). Loudspeakers, rooms, recordings and adaptation: Key factors in what and how we hear. In Sound Reproduction (pp. 127-162). Focal Press. https://doi.org/10.4324/9781003477495-6

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

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. Proceedings of the 32nd Journées d'Informatique Musicale, 21-30. https://hal.science/hal-05102298/document

Wei, L., Yu, Y., Qin, Y., & Zhang, S. (2025). From tools to creators: A review on the development and application of artificial intelligence music generation. Information, 16(8), 656. https://doi.org/10.3390/info16080656

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

Wu, S. H., & Holmes, K. J. (2026). Is there a "mind" behind the music? Attributing music to AI can suppress narrative meaning-making. Cognitive Research: Principles and Implications, 11(1). https://doi.org/10.1186/s41235-026-00715-z