Humanness Percept: A Study on Human Perception in AI-Generated Music

1st Henrique Daniel de Sousa
Departmento de Ciência da Computação (DCC)
Universidade de Minas Gerais (UFMG)
Belo Horizonte, Brazil
henrique.daniel@dcc.ufmg.br

2nd Flavio Figueiredo
Departmento de Ciência da Computação (DCC)
Universidade de Minas Gerais (UFMG)
Belo Horizonte, Brazil
flaviovdf@dcc.ufmg.br

***Abstract*—Recent advances in AI music (AIM) generation services are currently transforming the music industry. Given these advances, understanding how humans perceive AIM is crucial both to educate users on identifying AIM songs, and, conversely, to improve current models. We present results from a listener-focused experiment aimed at understanding how humans perceive AIM. In a blind, Turing-like test, participants were asked to distinguish, from a pair, the AIM and human-made song. We contrast with other studies by utilizing a randomized controlled crossover trial that controls for pairwise similarity and allows for a causal interpretation. We are also the first study to employ a novel, author-uncontrolled dataset of AIM songs from real-world usage of commercial models (i.e., Suno). We establish that listeners' reliability in distinguishing AIM causally increases when pairs are similar. Then, we conduct a mixed-methods content analysis of listeners' free-form feedback, revealing a focus on vocal and technical cues in their judgments. Lastly, we conducted a comparative analysis of the vocal tracks, which revealed distinct differences between the human and AI performances.**

***Index Terms*—AI, Music Information Retrieval, Human Computer Interaction**

I. Introduction

The advancement of generative artificial intelligence (AI) technologies on the last years has redefined creative production across multiple domains on art [1], [2]. Among these, music has become a particularly prominent area where AI-generated compositions are increasingly more similar to those created by humans in terms of structure, harmony, and even emotions. Commercial models such as Suno[1] and Udio[2] demonstrate how convincingly AI can generate creative melodies, harmonies and even full arrangements, often confounding human and machine creation. Although, since art is is intrinsically moved by subjectivity, we no longer have a exact score of correctness [19]. Therefore, this motivates a philosophical question of what differentiates authentic (human) from artificial (AI).

As these technologies emerge, a important question arise: can listeners distinguish between AI-generated music (AIM) and human-composed music? This question brings concerns about authenticity, artistic value, and the evolving role of human creativity. These worries affects not only artists and audiences but also music streaming platforms and the music industry, which now deal with the implications of content authenticity, copyright, and trust.

Although some prior research has explored human interaction with AI in creative domains [10], [29], relatively few studies have empirically studied aspects that could influence in how humans percept AI music and detect them. This gap limits our understanding of human-AI interaction in musical contexts and motivates the present study.

Importantly, this work does not aim to definitively determine whether humans can effectively distinguish AIM from human music. Instead, our focus is on the listener and their perceptions, covering the reasons and the factors that guide their decisions when differentiating between human and AI-created music and to identify perception limitations on human and machine creations and to assess the factors that may influence detection accuracy. Specifically, we test the following hypothesis:

Human listeners rely on contextually grounded cues (e.g., as repetitive structure or synthetic-sounding vocals) that help discern whether a piece of music is AIM or human-made.

To address this, we designed a perception experiment, based on a listener-focused Turing-like [] test in which participants were asked to listen to five pairs of music excerpts. After listening an specific pair, they judged their origin: AI or human. For each pair, participants could choose one of five options: (1) both tracks were human-composed; (2) the first track was AI-generated; (3) the second track was AI-generated; (4) both were AI-generated; or (5) they could not state. Additionally, to remove recognition bias of the responses, participants were asked whether they recognized any of the tracks.

Furthermore, for each pair, participants were also presented with an optional free-text input field to share any opinions about the songs or explain their choice. Finally, we also internally logged how long participants spent on each pair To explore potential demographic influences, we also collected information on participant's age, native language, and musical experience, divided into practical (e.g., instrument playing, composing), formal (e.g., conservatory, musical schools, college) experience and if they had any experience with AIM.

After the responses were collected, to study the how the

[1] https://suno.ai

[2] https://udio.com/

context influenced participants to detect AIM and what are the most important studied variables for this matter, we used a Randomized Controlled Cross Trial (RCCT) [4]. For instance, we used pairs formed by random music, working as the control group and pairs preformed using music similarity, working as the study group. Listeners evaluated multiple pairs, being them disposed in a random order for each listener. Also, with the provided comments, we used from a mixed-methods [5] grounded-theory content analysis [5] on the text feedback provided on why they made their choices.

Our findings reveal that when songs are paired randomly, listeners are unable to distinguish between AIM and human music. However, when pairs are intentionally similar, listeners show a better ability to make this distinction. We also find that practical musical experience (such as playing an instrument) and prior familiarity with AI music increase the likelihood of identification. Furthermore, our content analysis observed that listeners often rely on vocal and technical cues when making their judgments. These insights have important implications both for improving the human-likeness of AI-generated music and for developing strategies to educate audiences on how to recognize it.

II. Related Work

The intersection of artificial intelligence and human creativity has drawn significant attention with the rise of generative models []. Since our work focuses on musical content, we will focus primarily on music-related studies. Nonetheless, it is worth noting that perception research in other creative domains has a long history. On the 1960s, a study showed that participants often struggled to distinguish between human-made and computer-generated images [6]. More recent researches report similar findings in the context of generative AI for images [7]–[9] and poetry [10]. Focusing, on the human perception instead on accuracy, Candello et al. [11] demonstrated that even design elements such as typefaces can be judged by humans as appearing more or less "machine-like."

These studies, however, often guide participants toward evaluating specific characteristics, such as beauty or novelty. In contrast, our approach employs a randomized controlled crossover trial (RCCT) [4], which allows us to demonstrate that participants can distinguish between AI-generated and human-made content when the material is closely related, with causality. Moreover, unlike prior work, we do not direct participants toward any predefined cues. Instead, we invite them to describe the reasoning behind their choices in open-text responses, which we then analyze through a mixed-methods [5] content analysis.

In the music domain, a few works have explored public perception of AI-generated compositions. Lecamwasam and Chaudhuri [12] conducted a perception study assessing how listeners respond the emotions on AI-created music, showing that while participants often failed to correctly identify AI-generated pieces, their emotional reactions were not significantly different from responses to human compositions. Sarmento et al. [13] analyzed how contextual framing (e.g., disclosing the music's source) influences the perceived quality and authenticity of AI-generated progressive metal music, being the artificial music generated as a symbolic composition. Grötschla et al. [14] benchmarks human preference over the quality of different music generators and datasets, such as MTG-Jamendo [15], which uses human music, and Suno for AI music, as this work does.

Also, works using voluntary participants are important to fundament the proposal and the validity of this research. Therefore, Santy et al. [16] arguments that research using voluntary work can help sustain high-quality data. So, a related method was used on [17] on video and social media domain. This work investigated how people perceive video popularity by presenting users with pairs of videos and asking them to judge which would be more popular. This study used a similar interface as implemented in our study.

III. Materials and Methods

A. AI-Generated Music Dataset

AI-generated tracks were obtained by web scraping posts from the Suno Reddit's community `r/Suno` posts. We developed a custom web scraper that parsed Reddit threads, extracted relevant media links, and downloaded publicly shared AI-generated music samples. Only posts with complete audio files and clear user confirmation that the content was generated using Suno were included. Posts may be accompanied by textual tags, known as *flairs*, that indicate the content of the post: `Song`, `Song - Audio Upload`, `Song - Human Written Lyrics`, `and/or Song - Meme`. From these, we ignored the Meme songs due to their comedic and distinctive nature.

After scrapping 33,626 postsJ dated from July 21, 2023, up to February 25, 2025, it resulted on a dataset with 12,483 audio files generated by Suno, posted by random users on Reddit. From the other flairs, audios were downloaded by following Suno (4,059 songs) and/or YouTube (8,315) links. These songs were paired with ones from MTG-Jamendo. From these music, only 5,244 had their genres predicted, with the genres distribution as follows in Figure 1.

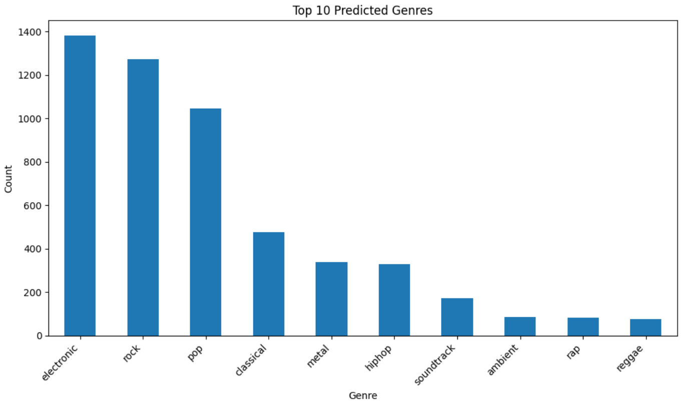

Fig. 1: Collected music dataset genres distribution

B. Human-Composed Music Dataset

Human-composed tracks were selected from the MTG-Jamendo dataset [15], an open-source music collection of songs distributed under Creative Commons licenses. This dataset was chosen for its rich genre diversity, metadata and genre annotations, and certainty that all tracks were created by human artists.

C. Pairing AI and human music

To create meaningful comparisons for the perception task, each AI-generated music sample was paired with a human-composed counterpart based on genre, duration, and overall similarity. Moreover, we also made sure to select songs with durations ranging from 1.5 minutes to 4 minutes.

We applied the pre-trained genre classification model provided by Essentia [20], developed using the MTG-Jamendo dataset, to predict the musical genre of both the AI-generated (Suno AI) and human-composed tracks. This allowed us to assign genre labels across both sources, with a mapped accuracy, despite the lack of structured metadata in the Reddit-sourced AI tracks, when it treats to genre.

To compute audio similarity, we extracted high-dimensional embeddings from all selected tracks using the CLAP (Contrastive Language-Audio Pretraining) model [18], which is designed for cross-modal representation learning of audio and text, from the Hugging Face[3] library. These embeddings capture semantic and acoustic features beyond surface-level attributes, making them suitable for perceptual pairing. Therefore, we computed cosine similarity between each AI-human pair using the cosine similarity of the embeddings of the two audios.

To improve the fairness and quality of comparisons, we used the following pairing constraints:

- Cosine similarity between embeddings ≥ 0.85
- Duration difference less than 30 seconds
- Genre classification confidence score ≥ 0.4 for both tracks

This pairing strategy ensured that human and AI tracks in each pair were aligned in genre and embedding similarity, isolating genre and other musical features and allowing to study the AI bias of the music excerpts, only.

As result, the pairs obtained followed the genre distribution showed on 2. Although, of all the formed pairs, we choose 10 to be part of the experiment. Also, we choose, among the music from both datasets, 5 artificial and 5 human music, not present on the previously selected pairs, to form pairs of random matching, always with one human and one AI track.

D. The Hummanness Perception Study

Upon entering the study, participants provided an email address, which was immediately converted to a one-way hash to anonymously track their responses. The core task required participants to evaluate five pairs of songs.

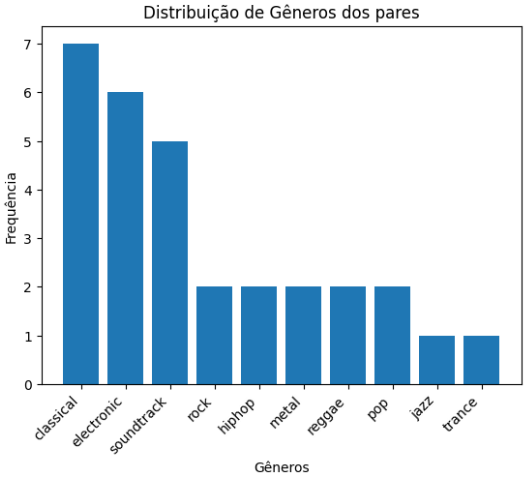

Fig. 2: Pairs genre distribution

The first four pairs were randomized for each participant. Two pairs were drawn from a *random* pool and two from a *similar* pool of predefined pairs. To prevent ordering effects, these four pairs, and the order of the two songs within each pair, were fully randomized. Songs were never repeated across pairs for a given participant.

The fifth and final pair was a fixed attention check (or "trap"). It featured the iconic introduction to Beethoven's Symphony No. 5 in C minor, Op. 67, paired against an AI-generated song explicitly designed for the task, which opened with the lyric: "This is not a human song, I'll say it right away."

The study was presented as a web form where each song pair appeared on its own page. Participants could not skip pairs or return to change previous answers. Song titles were not displayed. For each pair, participants were required to answer two questions:

1) **Source Identification:** A multiple-choice question asking them to classify the songs' origins:
   A. The first song was generated by AI.
   B. The second song was generated by AI.
   C. Neither song was generated by AI.
   D. Both songs were generated by AI.
   E. I cannot state if these songs were generated by AI or by humans.
2) **Familiarity:** A multiple-choice question about prior exposure to the songs:
   A. I have heard the first song before.
   B. I have heard the second song before.
   C. I have heard neither song before.
   D. I have heard both songs before.

Additionally, an optional text field allowed participants to provide comments or explain their choices. The time spent evaluating each pair was automatically logged.

After evaluating all five pairs, participants were invited to complete an optional survey collecting demographic and

[3] https://huggingface.co/laion/clap-htsat-unfused

background information. This included their age, native language, and familiarity with AI music (AIM) services. We also inquired about their musical background using binned categories: *Less than one year, one to five years, five to ten years,* or *over ten years* for both formal musical education and practical experience (i.e., playing an instrument).

We recruited participants from two distinct populations: online volunteers and paid crowd-workers. The **volunteer pool** was sourced primarily from social media posts by the Computer Science and Music Departments of a major Brazilian university. The study link was also shared on the university's website[4] and featured in local news outlets. Participants were also encouraged to share it further.

The **crowd-worker pool** consisted of 100 English-speaking participants recruited through the Prolific[5] platform, who were paid £2 for their time. This dual-recruitment strategy provided a more diverse demographic sample and allowed us to compare responses from intrinsically motivated volunteers with those from extrinsically compensated crowd-workers. Our pairs, responses, and source code are available at: https://github.com/uai-ufmg/hp-study.

IV. Results

Our analysis is based on a refined dataset, obtained by meticulously filtering the initial pool of participants to ensure data reliability and relevance. From June 6th to July 30th, 2025, a total of 653 participants accessed our study's website. To isolate a cohort of highly engaged and valid participants, we applied a two-stage filtering process based on criteria designed to focus on participants who demonstrated sustained attention and had no prior knowledge of the experimental stimuli.

This process yielded a final cohort of **308 participants**, which provided a total of **1,232 valid answers** for the core analysis of the initial four pairs (308×4).

The filtering process is summarized in detail in Table I.

TABLE I: Participant Filtering Flow

Stage	Description	Participants
Initial Pool	All participants who logged into the study's website.	653
Filter 1	Participants who reached the fifth pair and correctly identified the well-known Beethoven piece, demonstrating sustained attention and engagement.	337 (out of 504)
Filter 2	Participants from the previous stage who had no prior knowledge of any songs in the first four pairs. This ensured that knowledge of a song was not a confounding factor.	308 (out of 337)

The first filter was crucial for removing participants who did not complete the study or were not fully engaged. Of the 504 participants who reached the final pair, 337 successfully identified the Beethoven piece, which corresponds to a success rate of approximately 66% ($337/504$). The second filter ensured that our results were not skewed by prior musical knowledge, which could compromise the integrity of the study's findings. This rigorous filtering resulted in a final, robust dataset for our analysis.

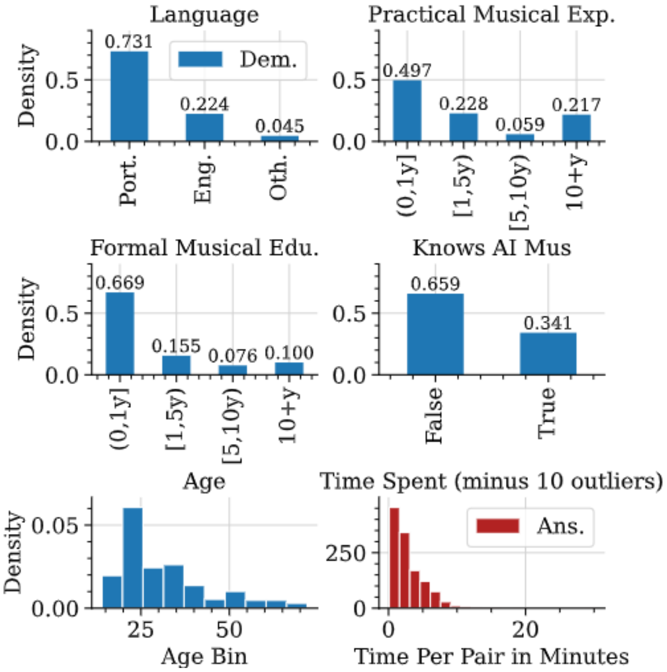

Fig. 3: Demographic and Answer Variables

Of the 308 valid participants, 290 completed the demographic survey, with the results summarized in 3. The majority of participants were native Portuguese speakers (73%), followed by English speakers (22%) and other languages (5%). Regarding musical background, most participants reported having no formal musical education (67%) and no practical experience (50%). Furthermore, 34% of participants indicated knowledge of AIM. The mean age of the cohort was 31 years (SD = 13), with a median of 27 years. After excluding 10 outliers (values greater than 30 minutes), the average time participants spent on each pair was 2.98 minutes.

A. *RCCT analysis*

In our RCCT, participants were asked to identify which song in each pair—A or B—was AI-generated (AIM). Correct responses accounted for 57% of all answers. In contrast, responses indicating "both" (20%), "neither" (16%), or "can't decide" (7%) were classified as incorrect. Because of the randomized trial design, these response patterns were influenced by the pair type (*similar* or *random*). Additional factors may also have affected outcomes, including the specific $\{A, B\}$ pairing, listener demographics, time spent evaluating each pair, the pair's presentation order (from one to four), and other unobserved participant characteristics. In the following sections, we present results both with and without controls for these endogenous factors.

We first compared listeners' success rates in distinguishing between song A and song B. With two possible choices,

[4] https://hpc.uai.science

[5] https://www.prolific.com/

random guessing would yield an expected success rate of $\mathbb{E}[s] = 0.5$. Across all pair types, the overall observed rate was $\hat{s}_o = 0.60$, indicating that participants performed above chance on average. When examined by the context, i.e., which experiment set the response referred to, the *random* set yielded a success rate of $\hat{s}_r = 0.53$, which a binomial test confirmed was not significantly different from chance ($p > 0.05$). In contrast, the *similar* set produced a higher success rate, resulting in $\hat{s}_s = 0.66$ of accuracy, significantly exceeding both random guessing ($p < 10^{-9}$) and the overall mean performance ($p < 10^{-2}$).

To assess the robustness of these effects, we repeated the analysis under two restricted conditions. First, when considering only songs containing lyrics, the results were consistent: $\hat{s}_r = 0.53$ ($p > 0.05$) and $\hat{s}_s = 0.75$ ($p < 10^{-6}$), suggesting that the presence of vocals enhanced the discriminability of AI-generated music. Second, when excluding the *classical* and *ambient* genres, which appeared exclusively in the *similar* set, the success rates remained nearly unchanged ($\hat{s}_s = 0.66$, $p > 0.05$; $\hat{s}_r = 0.53$, $p < 10^{-7}$). These consistent patterns across subsets indicate that the observed differences are unlikely to be confounded or biased by genre distribution or lyrical content, strengthening the conclusion that song contextual similarity plays a key role in listeners' ability to detect AI-generated music.

To examine the influence of our control variables, we employed a Mixed Effects Logistic Model [21], summarized in Table II, with the goal of serving as a explanatory model to study the variables which were more relevant while discriminating the pairs. In fact, this model's was designed as an estimator for success rate or accuracy, but in this work perspective, it served mainly as an explanatory model. The model incorporated pair-specific intercepts to account for systematic differences in pair difficulty, as well as nested random intercepts for participant IDs and pair order, thereby controlling for individual differences in performance and potential ordering effects within each participant's session.

The model achieved a McFadden's $R^2 = 0.44$, which represents a relatively high level of explanatory power for behavioral data and is considered more than acceptable for models of this kind [22]. This suggests that a substantial portion of the variability in listeners' accuracy can be explained by the included random effects and control variables. In particular, the inclusion of pair-level and participant-level terms indicates that both the inherent difficulty of each pair and individual differences among listeners played a meaningful role in predicting success. Overall, the model confirms that accuracy was not solely driven by chance or pair type, but was systematically influenced by contextual and participant-specific factors.

Full Hierarchical Model

		Estimate	Pr(¿|z|)	Sig.
	Intercept	-24.26	0.9968	
↑	Similar Pair	0.61	0.0999	*
	Choice: Song A or B	22.29	0.9971	
	Choice: Both Songs	0.82	0.9999	
	Choice: Neither Song	4.94	0.9994	
↑	$\log_{10}$(TimeSpent+1)	0.49	0.0611	*
	Lang. Port.	-0.07	0.7621	
	Prac. Exp. 1 to 5 y	0.42	0.1157	
↑	Prac. Exp. 5 to 10 y	0.92	0.0995	*
↑	Prac. Exp. Over 10 y	1.25	0.0009	***
	Formal Edu. 1 to 5 y	-0.22	0.4803	
↓	Formal Edu. 5 to 10 y	-1.30	0.0086	***
	Formal Edu. Over 10 y	-0.82	0.0614	*
↑	Knowledge on AIM	0.89	0.00005	***
↓	Participants' Age	-0.03	0.0009	***

TABLE II: Covariates *$p < .1$, **$< .05$, ***$< .01$

Non-Hier. Without Practical Exp.

		Estimate	Pr(¿|z|)	Sig.
	Intercept	-22.01	0.9851	
↑	Similar Pair	0.58	0.0005	***
	Question Order	0.00	0.9864	
	Choice: Song A or B	19.89	0.9865	
	Choice: Both Songs	-0.14	0.9999	
	Choice: Neither Song	-0.27	0.9998	
↑	$\log_{10}$(TimeSpent+1)	0.55	0.0181	*
	Lang. Port.	-0.12	0.5645	
	Formal Edu. 1 to 5 y	0.15	0.5256	
	Formal Edu. 5 to 10 y	-0.22	0.4574	
	Formal Edu. Over 10 y	0.11	0.6996	
↑	Knowledge on AIM	0.86	0.000003	***
↓	Participants' Age	-0.03	0.0003	***

TABLE III: Covariates *$p < .1$, **$< .05$, ***$< .01$

Using our model, we observed that participants with more than five years of *Practical Experience* in music exhibited a positive and significant effect on accuracy. This pattern was also found for those reporting prior knowledge of AI-generated music (AIM), suggesting that familiarity with the domain or with generative tools enhances one's ability to detect synthetic compositions. In contrast, *age* showed a negative association, indicating that older participants were less likely to correctly identify AI-generated songs. Interestingly, having a *Formal Education* in music of 5 to 10 years appeared with a negative coefficient. However, this effect should be interpreted with caution, as *Practical Experience* and *Formal Education* are highly correlated. Indeed, when we estimated an alternative model excluding the Practical Experience factors (Table **??**), the apparent negative effect of education disappeared, suggesting multicollinearity rather than a true suppressor effect.

The model further indicates that the treatment effect of being exposed to a *similar* pair, as opposed to a *random* pair, corresponds to a 13% improvement in accuracy ($\hat{s}_s - \hat{s}_r = 0.13$). This difference remains consistent across robustness checks, highlighting that perceptual similarity between songs substantially increases the likelihood of correctly identifying the AIM one. Importantly, this treatment effect is not uniform across all listeners—it is moderated by individual characteristics such as experience, prior knowledge, and age. Together, these results provide evidence not only for *when* users differentiate between human and AI-generated music, but also *who* is more capable of doing so and under what contextual conditions.

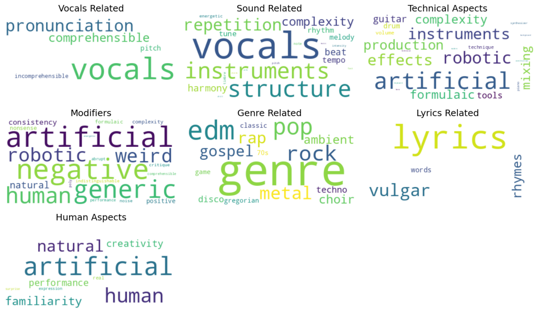

Fig. 4: Topics and tags. Word size is proportional to usage within topic. Top-7 overall frequency: vocals (369), lyrics (247), negative (231), artificial (224), generic (174), human (130), robotic (112)

B. Mixed Methods Analysis of Feedback

Redirecting our attention to the analysis of the free-text responses collected in the survey, we obtained a total of 317 comments from 140 distinct participants. To examine this qualitative data, we employed a manual open-coding approach conducted over two iterative sessions. Each session involved three independent coders who received a uniquely randomized list of responses to minimize shared bias and ensure a diverse reading order.

During the first coding round, the three coders independently annotated 100 responses, freely assigning descriptive textual labels that captured salient themes. After one week, the coders met for a calibration session to consolidate their observations and establish a shared coding framework. This process resulted in the definition of seven overarching thematic categories: *vocals related*, *sound related*, *technical aspects*, *human aspects*, *modifiers*, *genre*, and *lyrics related*. For each of these topics, a set of non-exclusive tags was defined to allow multiple dimensions to coexist within a single response (Figure 10).

In the second round, conducted over the following week, all 317 responses were coded by all three coders using the finalized tagging scheme. Additionally, coders specified when a tag referred to a specific song within the presented pair, enabling a finer-grained mapping between textual justifications and pair-level perceptions. Our subsequent quantitative analysis focuses on responses with full coder agreement (3/3; $n = 289$), ensuring high reliability in the extracted categories. A complementary qualitative discussion of these themes, including representative excerpts and edge cases, is provided in Appendix A.

The high level of inter-coder agreement suggests that participants' reasoning patterns were relatively consistent and that the identified categories capture meaningful aspects of how listeners differentiate AI-generated from human-made music. These qualitative insights enrich the quantitative findings, offering a deeper understanding of the perceptual cues and interpretive strategies underlying participants' judgments.

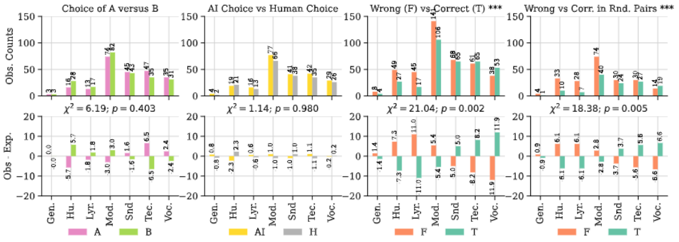

Fig. 5: Observed Topic Frequencies and Differences Towards the Expected. $***p < .01$

We initially observe from Figure 10 that participants frequently cite vocal and lyrical aspects to justify their decisions. To quantify the impact of these cues, Figure 5 analyzes topic distribution across four distinct contexts, employing χ^2-tests for significance: (1) arbitrary song selection (Song A vs. B); (2) attribution (AIM vs. human-made); (3) classification accuracy (correct vs. incorrect); and (4) accuracy specifically within the challenging *random* subset.

Given the randomized song order, we observed no statistical difference in topic usage based on the choice of one pair over the other. Similarly, the choice between AIM and human-made attribution yielded no significant disparities. This suggests that participants employ consistent justifications regardless of whether they label the song as AI or human. However, statistical significance emerges when conditioning on accuracy. In comparisons of correct versus incorrect responses—particularly within the harder *random* setting—we find that contextually grounded cues (*sound*, *technical*, and *vocal* elements) are instrumental in successfully distinguishing AIM.

These findings advance our understanding of human perception regarding AI-generated content. Identifying the cues listeners employ to distinguish AIM can inform strategies for improving the naturalness of generative models. Conversely, these insights may guide the development of educational initiatives designed to help users better recognize AI-generated media.

C. Instrumental and Vocal analysis

Motivated by the finding that participants rely heavily on *sound*, *technical*, and *vocal* cues, we investigated the objective feature differences between AI and human-generated audio components.

To isolate vocal features, we performed source separation using the Demucs model [24]. We then refined the dataset by extracting high-confidence vocal segments using a Voice Activity Detection (VAD) model [25], restricting the duration to a range of 2–15 seconds. This methodology aligns with the preprocessing steps employed in [23]. As detailed in Appendix B, the resulting AI segments averaged 13 seconds in length, compared to 7.8 seconds for human vocals. We subsequently extracted 512-dimensional embeddings for each segment using YAMNet [26].

To account for stylistic variations, we stratified our analysis by genre using the Essentia [20] classifications obtained previ-

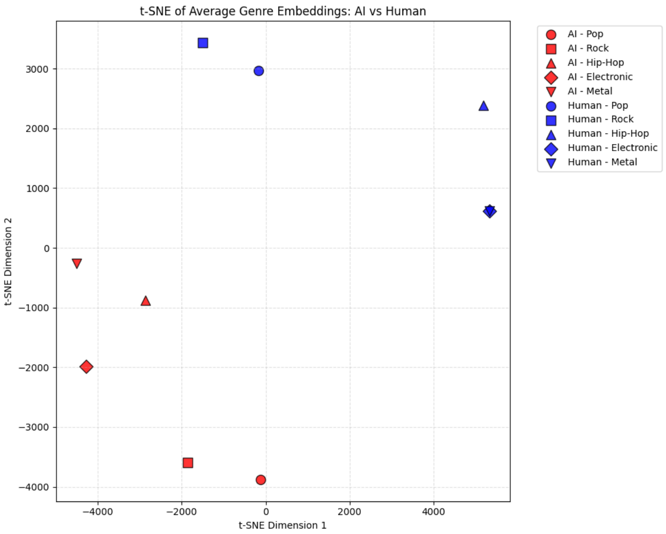

Fig. 6: t-SNE plot of average embeddings for vocals

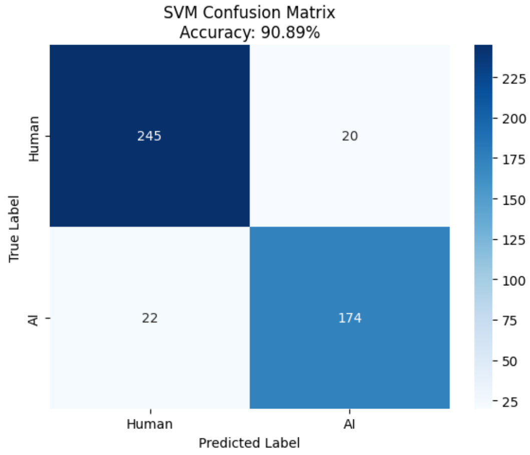

Fig. 7: SVM confusion matrix for AI vs Human predictions on test split

ously (pop, rock, hiphop, electronic, and metal). We computed the mean embedding vectors (centroids) for each genre and visualized their distribution in two dimensions using t-SNE [27]. These projections are illustrated in Figure 6.

The results demonstrate that AI vocals and Human vocals form different clusters, suggesting that, in average, for each genre, they differ.

The results are enforced by the SVM results achieved on Fig. 7. It achieved an accuracy of 0.9 on deciding if a vocal is human or not, when comparing the overall vocals embeddings. This SVM model was trained with 979 AI samples and 1324 Human samples, with a train-test split proportion of 0.8 to train and 0.2 to test.

To identify the concrete signal characteristics driving this distinction, we performed a Fourier spectral analysis [28] to examine the frequency distribution of the segments. Figure 8

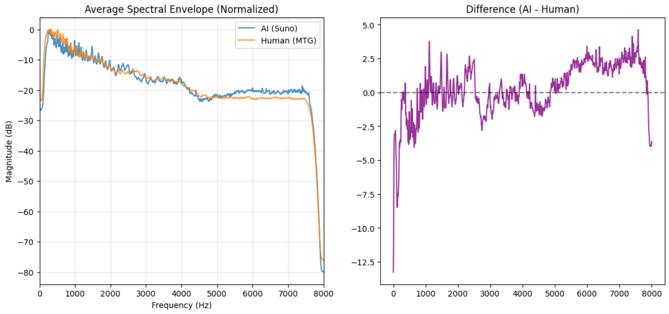

Fig. 8: SVM confusion matrix for AI vs Human predictions

reveals a distinct spectral inversion: AI-generated vocals exhibit a marked energy surplus in the high-frequency range (approximately 5–8 kHz), whereas human vocals maintain dominance in the lower fundamental frequencies (< 1 kHz).

This spectral envelope discrepancy offers an objective basis for the subjective feedback reported by participants. The lack of low-end energy likely contributes to a perceived lack of ”warmth” or ”body” in the AI vocals, while the high-frequency content aligns with descriptions of a ”robotic” or ”metallic” timbre. This phenomenon is often associated with high-frequency quantization noise in generative models.

V. Limitations

We acknowledge that we had no control over the prompts used to generate the AIM songs, nor over the recording and mixing processes of the Jamendo tracks. Consequently, prompt efficacy and recording fidelity remain uncontrolled variables that could theoretically influence our results; however, we aimed to mitigate these effects by examining a diverse selection of song pairs. Furthermore, we utilized the highest audio quality available for download: 48 kHz, 192 kbps (stereo) for AIM, and 44.1 kHz, 320 kbps (stereo) for the MTG-Jamendo dataset. To minimize potential processing bias, we strictly avoided re-encoding the audio files, utilizing them exactly as they were hosted on their respective source platforms. We further posit that the marginal difference in bitrates across datasets had no impact on our findings, a conclusion supported by the fact that participants in the random experiment performed no better than random chance in identifying AIM tracks.

VI. Conclusion

The increasing usage of AI-generated content poses several challenges to the music industry, particularly regarding the need to educate users on identifying artificial content. Through a randomized controlled crossover trial and a mixed-methods coding study, we demonstrate that the ability to distinguish AI-generated music (AIM) from human compositions is causally linked to contextual similarity, with detection rates significantly higher when comparing similar tracks rather than random pairings. We unveil that participants primarily rely on vocal anomalies and technical artifacts to identify

AIM, a subjective observation corroborated by our spectral analysis, which revealed distinct high-frequency energy in AI models. Furthermore, demographic analysis indicates that practical musical experience and familiarity with AI tools improve detection accuracy, while older age is associated with lower performance. These findings highlight the need for technical advancements in vocal synthesis to improve fidelity and underscore the importance of educational initiatives to foster media literacy. As future work, we aim to extend our study to music beyond the Western, educated, industrialized, rich, and democratic (WEIRD) domain and develop similar initiatives for other media types, such as text, images, and video.

References

[1] Heigl, R. Generative artificial intelligence in creative contexts: a systematic review and future research agenda. Manag Rev Q (2025). https://doi.org/10.1007/s11301-025-00494-9

[2] Nantheera Anantrasirichai and David Bull. 2022. Artificial intelligence in the creative industries: a review. Artif. Intell. Rev. 55, 1 (Jan 2022), 589–656. https://doi.org/10.1007/s10462-021-10039-7

[3] Y. Zang, Y. Zhang, M. Heydari and Z. Duan, ”SingFake: Singing Voice Deepfake Detection,” ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, 10.1109/ICASSP48485.2024.10448184.

[4] B. Jones and M. G. Kenward. Design and analysis of cross-over trials. Chapman and Hall/CRC, 2003.

[5] L. Doyle, A.-M. Brady, and G. Byrne. An overview of mixed methods research. Journal of research in nursing, 14(2), 2009.

[6] A. M. Noll. Human or machine: A subjective comparison of piet mondrian’s “composition with lines” (1917) and a computer-generated picture. The Psychological Record, 16(1), 1966.

[7] A. Elgammal, B. Liu, M. Elhoseiny, and M. Mazzone. CAN: Creative adversarial networks generating “art” by learning about styles and deviating from style norms. In Proc. ICCC., 2017.

[8] M. Ragot, N. Martin, and S. Cojean. Ai-generated vs. human artworks. a perception bias towards artificial intelligence? In Proc. CHI. Extended Abstracts, 2020.

[9] A. Xu, S. Fang, H. Yang, S. Hosio, and K. Yatani. Examining human perception of generative content replacement in image privacy protection. In Proc. CHI, 2024.

[10] N. Köbis and L. D. Mossink. Artificial intelligence versus maya angelou: Experimental evidence that people cannot differentiate ai-generated from human-written poetry. Computers in Human Behavior, 114:106553, 2021.

[11] H. Candello, C. Pinhanez, and F. Figueiredo. Typefaces and the perception of humanness in natural language chatbots. In Proc. CHI, 2017.

[12] K. Lecamwasam and T. R. Chaudhuri. Exploring listeners’ perceptions of ai-generated and human-composed music for functional emotional applications. arXiv preprint arXiv:2506.02856, 2025.

[13] P. Sarmento, J. Loth, and M. Barthet. Between the ai and me: Analysing listeners’ perspectives on ai-and human-composed progressive metal music. In Proc. ISMIR., 2024.

[14] F. Grötschla, A. Solak, L. A. Lanzendörfer, and R. Wattenhofer. Benchmarking music generation models and metrics via human preference studies. In Proc. ICASSP. IEEE, 2025.

[15] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra. The mtg-jamendo dataset for automatic music tagging. In Proc. ML4MD at ICML., 2019.

[16] Santy, S., Bhattacharya, P., Ribeiro, M. H., Allen, K., and Oh, S. (2025). When incentives backfire, data stops being human.

[17] Figueiredo, F., Almeida, J. M., Benevenuto, F., and Gummadi, K. P. (2014). Does content determine information popularity in social media? A case study of youtube videos’ content and their popularity. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI ’14, page 979–982, New York, NY, USA. Association for Computing Machinery.

[18] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In Proc. ICASSP, pages 1–5, 2023.

[19] A. Lerch, C. Arthur, N. Bryan-Kinns, C. Ford, Q. Sun, and A. Vinay. Survey on the evaluation of generative models in music. arXiv preprint arXiv:2506.05104, 2025.

[20] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José Zapata, and Xavier Serra. 2013. ESSENTIA: an open-source library for sound and music analysis. In Proceedings of the 21st ACM international conference on Multimedia (MM ’13). Association for Computing Machinery, New York, NY, USA, 855–858. https://doi.org/10.1145/2502081.2502229

[21] Vermunt, J. K. (2005). Mixed-Effects Logistic Regression Models for Indirectly Observed Discrete Outcome Variables. Multivariate Behavioral Research, 40(3), 281–301.

[22] McFadden, D. (1974) Conditional Logit Analysis of Qualitative Choice Behavior. Frontiers in Econometrics, 105-142.

[23] Y. Zang, Y. Zhang, M. Heydari and Z. Duan, ”SingFake: Singing Voice Deepfake Detection,” ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 12156-12160, doi: 10.1109/ICASSP48485.2024.10448184.

[24] Défossez, A., Usunier, N., Bottou, L., Bach, F. (2019). Music Source Separation in the Waveform Domain.

[25] Hervé Bredin (2023), pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe

[26] S. Lachenani, H. Kheddar and M. Ouldzmirli, ”Improving Pretrained YAMNet for Enhanced Speech Command Detection via Transfer Learning,” 2024 International Conference on Telecommunications and Intelligent Systems (ICTIS), Djelfa, Algeria, 2024, pp. 1-6, doi: 10.1109/ICTIS62692.2024.10894266.

[27] Laurens van der Maaten and Geoffrey Hinton, (2008), Visualizing Data using t-SNE, Journal of Machine Learning Research

[28] Afchar, D., Meseguer-Brocal, G., Akesbi, K., & Hennequin, R. (2025). A Fourier Explanation of AI-music Artifacts. ISMIR 2025

[29] Miguel Civit, Javier Civit-Masot, Francisco Cuadrado, and Maria J. Escalona. 2022. A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends. Expert Syst. Appl. 209, C (Dec 2022). https://doi.org/10.1016/j.eswa.2022.118190

Appendix A
Qualitative Analysis of Textual Feedback

In this appendix, we present a brief qualitative exploration of some of the answers provided by participants. We group these answers by topics, that were defined on our main text.

We begin with an exploration of answers that were **Sound Related** and **Technical Aspects Related**. Regarding sound aspects, we observe that the listener’s perception on the musical features plays an important role for the task of differentiating AIM from human-made. On the technical side, listeners presented their perception on the audio quality, the effects, the production and the tools used to create the track, from their point of view.

(translated) *[..] In general, both songs may have been produced using samples, synthesizers, or software that wouldn’t be classified as AI but aren’t exactly human either..*

Audio quality on track 2 makes it sound like it was generated

Track 2 immediately seemed more human cause it made better use of stereo sound.

Vocals Related: When considering vocal aspects, participants took into account cues such as pronunciation, technical performance, singing quality, as well as others. Some examples are:

"The first song felt awkward, the singer's voice failed sometimes. The second one was more smooth, I could hear the air in the microphone, in the p's and b's". Notice the focus on pronunciation

(translated) "The voices on the second track sounded a bit robotic to me, as if they were artificial, created by AI, and I couldn't pick up on any singing strategies (like head and chest voices) throughout the song.". Here the focus is on singing performance. this is similar to the example below:

"In the beginning, I guessed number 1 was AI-generated, but after listening number 2, where the voice is not fluid as expected, I decided that the track 2 is AI-generated, instead of 1.".

Some cases focus on whether the voice was robotic or not: *"The singer's voice in track 2 sounds more robotic"*

Or if it sounds like a voice recorded live: *"Sounds like a live recording, especially because the lead vocal is quieter than it should be [...]."*.

Human vs Artificial Aspects. Listeners considered factors such as creativity, expression, performance, naturalness, in order to present human or artificial aspects of a specific track. Some comments brought their intuition on judging how "natural" a track seemed, with an attempt to justify this perception.

(translated) "AI-generated audio, from videos like those that have been trending, has a noise at the end of each word. And the first audio also has... That's why it's believed to be AI.". Observe the focus on pronunciation. This is also an example of a feedback that focused on vocal aspects.

[...] the singer's voice sounds unnatural, recognizably unnatural, something that's not quite humanly natural. [...] the #2 track sounds ok, creative like humans would do.. Focus on AI versus human creativity.

[...] It's a stage feeling to describe but it sounds very human in the way it sound less like other human songs you hear normally. Also it sounds very good. The second it sound very artificial and in my opinion kinda uninspired too [...]. Focus on AI versus human performance.

The first song seems very Human made, the pronounciation was really human like. The vocal intonations felt real. Also the deeper voice in some parts seems like a human touch. [...]. Focus on AI versus human feeling.

Some of the vocal resonances and the way the vocals were false made me think both tracks were artificially created. Same as above.

Lyrics Related: Listeners heavily consider the lyric's quality when evaluating songs. Participants tend to classify as AI what they consider as incoherence in the text.

"Track 2 lyrics don't make any sense but, the song sounds good. That is why I believe track 2 is AI generated.".

Additionally, the content of the lyrics is viewed as a tip for some participants.

"The lyrics in track 1 were what made me think it was AI, they were a bit silly [...]".

Also, the poetic structure of lyrics (rhymes) is pointed out as a characteristic that some users employ to differ artificial from human musics.

"Track 1 is the first one I hear that I think current AI models couldn't replicate. The flow and the rhymes are too complex to be an AI song [...].".

"[...] both lyrics sound like made by AI, like the more average pop songs ever created."

"Track 1 is the first one I hear that I think current AI models couldn't replicate. The flow and the rhymes are too complex to be an AI song (I'm not saying AI will never do such songs). I'm almost certain it's human. [...]"

Genre Related: Some users highlight the commonness of elements present in the song, correlating them with the usual aspects seen on frequently distributed musics of the genre.

"Track 1 could be a solo for any 80's hard rock band [...]".

In contrast, a number of listeners specify which of these elements they notice as **not** being common; observe that, in this case, usually the listener ranges at different topics to conclude about the origin of the music.

(translated) "The instrumentation of the first song doesn't really match the theme, which seems to be somewhat gospel. [...]". Also, a few participants pointed out their previous familiarity on AIM models capacities to create convincing songs on various genres.

"[...] I just think its easier for AI to make a rap song than a heavy metal song without messing up too much.".

Furthermore, the match between lyrics and genre was also emphasized.

"The lyrics to genre match don't seem to make sense in the first track."

Modifiers. Modifiers were used to provide adjectives to the coders. These could be used in conjunction with another tag or as to describe the feedback as well.

(translated) I think the first one is strange, usually rappers don't repeat the same thing so many times. (repetitive)

Again, for me it comes to things being generic. I know there are some awfully written songs made by humans, but even then you can hear some distinctions in their voices in different moments (which is not the case on track 2). (generic)

Something about Track 1 sounded too perfect, too contrived to be Human. I was pretty confident with saying Track 2 was humans, by all the layers sounds and dimensions it had. (perfection)

The first has a weird noise on the voice. The second the vocalist seams that don't breath, is odd (weirdness)

I hope track 1 is AI generated, since it sounds idiotic. Track 2 lyrics sounds like something that AI could create trying to emulate rap music (critique)

The first one seems to have nonsensical lyrics, and the second one seems to have strange lyrics as well. (nonsense)

[...] Nonetheless, they were both very good recordings and hard to tell the difference. (positive)

Appendix B
Segment Duration distribution

This appendix shows the duration distribution of the vocal segments extracted from both Suno and MTG-Jamendo datasets, used in our vocal analysis, in order to add reliability to the study.

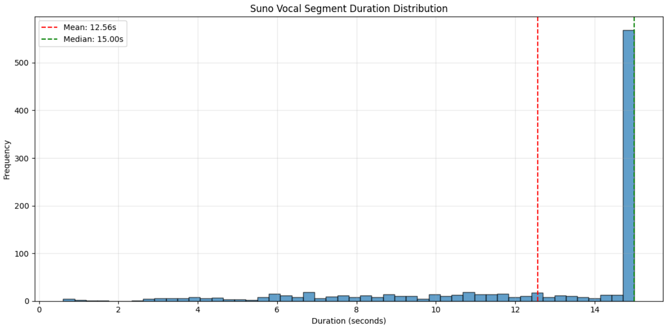

Fig. 9: Suno dataset vocal segments duration distribution

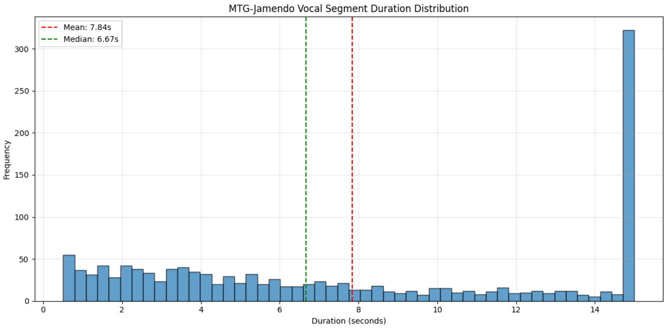

Fig. 10: MTG-Jamendo vocal segments duration distribution

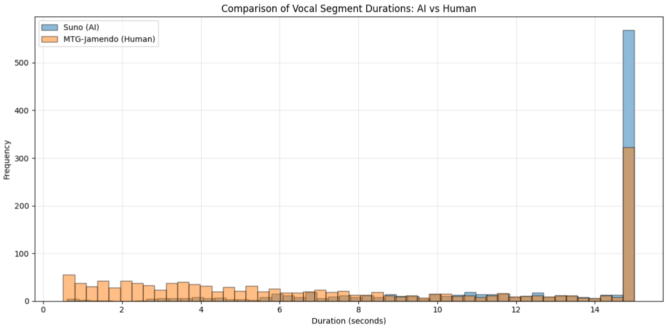

Fig. 11: Comparison of the MTG-Jamendo and Suno segments duration