Help That Hurts: The Creativity Cost of Generative AI Ideation in Music Composition

Jose Oros[1], Richard Randall[2], and Rahul Telang[1]

[1]Heinz College, Carnegie Mellon University
[2]School of Music, Carnegie Mellon University

Abstract

Generative AI is transforming creative industries, particularly music, and raising questions about artistic practice and creator roles. Text-to-music systems can produce full songs from prompts, yet their effect on human creativity and quality is unclear. This paper analyzes how AI assistance shapes the creative process and outcomes in music. We conduct a randomized controlled trial with musically trained individuals who compose an original 15-second piano melody, assigning them to either an AI-ideation group or a no-AI control. External evaluators of varied musical experience rate each melody on perceived creativity. On average, we find that listeners rate the melodies generated under the treatment condition as less creative. Furthermore, access to the AI changes stylistic choices, which are perceived as less creative by evaluators.

1 Introduction

Generative Artificial Intelligence (AI) is driving profound transformations across society by enhancing productivity, reshaping industries, and redefining the nature of human work. A growing body of research has examined the productivity effects of generative AI across various contexts. For instance, Brynjolfsson et al. (2025) find that generative AI assistance improves call-center agent productivity by 14%, while a recent study by the Federal Reserve Bank (Bick et al., 2024) reports an aggregate productivity increase of 1.1% following generative AI adoption. However, the majority of existing studies have focused on text-based generative systems, such as ChatGPT, and emphasize productivity, leaving other creative domains, particularly music, relatively understudied.

Recent advances in generative AI enable the production of sophisticated, human-like artworks. In music, generative AI is likely to impact the economics of music creation, industry workflows, cost structures, and revenue opportunities (Epstein et al., 2023). By automating aspects of composition, production, and performance, AI can reduce barriers to entry for creators and enable new business models that may disrupt traditional music production and distribution. Text-to-music systems can now produce complete songs from textual prompts, fundamentally altering artistic workflows and raising questions about authorship, originality, and the future role of human artists.

Beyond productivity, creativity and quality are central outcomes in music. If generative AI tools can produce songs that match human compositions in perceived quality and creativity, this raises questions about the future role of artists. It is also possible that, while generative AI can replicate patterns and generate novel sound combinations, it may lack the interpersonal experiences and cultural context that humans care about. Existing empirical work on text-based generative AI and creativity has focused almost entirely on text and images, with mixed findings on whether AI-generated or AI-assisted outputs are more creative. This literature also finds evidence of output homogenization. To our knowledge, there is no causal evidence on how access to generative AI affects creativity in music composition.

This paper analyzes how generative AI ideation shapes the creative process and outcomes in music composition. In our first experiment, we conduct a randomized controlled trial in which musically trained participants are asked to compose an original 15-second melody on a piano. Participants are randomly assigned to one of two conditions. In the AI-ideation group, they have access to Udio[1], a text-to-music generative AI platform, during composition. In the control group, they compose music without AI access. In both conditions, participants record their own melody on a piano; the AI system is used only as an ideation tool, not as a direct source of the final audio. In a second experiment, external evaluators of varying musical backgrounds, recruited online, listen to the resulting melodies and rate each one on perceived creativity.

One would expect that ideation can either improve creativity or have no impact. Generative AI models in music can generate ideas, continuations, or even full songs, expanding the set of readily available examples that composers can draw on during ideation. Composers may selectively incorporate elements they find appealing that they would not otherwise explore, or they may treat the tool as a reference and ultimately ignore its suggestions; either way, access could increase creativity ratings or leave them unchanged. However, exposure to Udio's polished AI outputs may anchor composers' choices and induce fixation, narrowing exploration and steering melodies toward common, high-probability patterns, which could lower perceived creativity ratings.

Our analysis explores this in detail and addresses three questions: (1) How does access to a text-to-music generative AI tool during the ideation phase affect the perceived creativity of the melodies that composers produce? (2) How does AI access change the musical features of these melodies (for example, their temporal density, pitch range, and tonal characteristics)? (3) Which musical features are most predictive of creativity ratings, and do structural changes induced by AI access align with dimensions that listeners reward or penalize?

We find that access to Udio during ideation does not increase creativity ratings; instead, it leads to lower perceived creativity. Assignment to the AI condition reduces average creativity ratings by about 0.15 points on a 1–5 scale. Interestingly, self-reported enjoyment and satisfaction with one's own melody do not differ by treatment arm, suggesting a gap between composers' internal evaluations and how their music is judged by external listeners.

In addition, we document systematic differences in compositional choices between the treatment

[1] `https://www.udio.com`

and control groups. We use the symbolic representation of each melody to extract measures of note density, timing, pitch register and range, contour, and tonality. Relative to the control group, melodies composed with AI access contain fewer notes, a narrower pitch range, and longer average note durations and inter-onset intervals. We then relate these features to creativity ratings. We find that timing and density measures explain a substantial share of the variance in creativity scores. Melodies with more notes and faster pacing tend to be rated as more creative, while less rhythmically active, slower melodies receive lower creativity scores.

We make the following contributions to the literature. First, we provide causal evidence on how access to a generative text-to-music system affects perceived creativity in human-composed music. Second, we show that AI access systematically changes the musical features of compositions and that these structural shifts are strongly associated with how listeners evaluate the melodies. These findings suggest that in creative domains such as music, generative AI assistance can expand the range of available ideas without a clear gain in perceived creativity and may nudge outputs toward patterns that listeners perceive as less creative, highlighting a potential trade-off between convenience and quality relevant to creators and platform designers. Our study also sheds light on the current debate about training generative AI models on copyright-protected content. The fair use doctrine permits the use of such material if it can generate significant innovation. If creativity is the criterion for innovation, we find that the use of generative AI tools for music composition diminishes it, raising questions about the extent to which such training enhances innovation.

2 Conceptual Background and Related Work

2.1 Creativity and generative AI

Creativity is commonly defined as the ability to produce something novel and valuable, such as an idea, solution, or product that departs from expected responses while remaining appropriate to its context (Marrone et al., 2024). Creative work often involves combining existing knowledge in new ways, exploring alternative possibilities, and iteratively refining ideas. In this sense, creativity has both a divergent component (i.e., generating multiple possibilities) and a convergent component (evaluating, selecting, and refining them).

A central challenge in studying creativity is how to measure it. Different tests have been developed depending on the approach and definition of creativity adopted by researchers (Said-Metwaly et al., 2017). Traditional tests such as the Torrance Tests of Creative Thinking (TTCT; Torrance, 1966, 2008) or the Alternative Uses Task (Guilford, 1967) capture divergent thinking by scoring responses on originality, fluency, and flexibility, while convergent tasks such as the Remote Associates Test (RAT; Mednick, 1962, 1968) focus on insights and problem-solving. These tests are useful for comparing individual traits or cognitive processes, but are less suitable for evaluating concrete creative products, such as pieces of music, in context. In contrast, product-based approaches such as Amabile's Consensual Assessment Technique (CAT; Amabile, 1982) assess creativity by asking knowledgeable observers to rate real outputs, on the premise that creativity is best judged

by those familiar with the relevant domain.

Generative AI systems can change the environment in which creative work occurs. Large Language Models (LLMs) and related systems can generate ideas, continuations, or even full artifacts, expanding the set of readily available options for creators. This suggests that such ideation can positively impact creativity. Recent work has examined the impact of generative AI systems on creativity. However, this literature has largely examined general-purpose LLMs (e.g., ChatGPT) and reports mixed effects. Some studies report that generative AI improves individual performance in some dimensions while increasing similarity and reducing collective novelty. In writing tasks, for example, AI-suggested ideas can increase ratings of creativity, quality, and enjoyment, but can also make stories more similar across authors (Doshi and Hauser, 2024). In poetry, Porter and Machery (2024) find that non-expert readers do not distinguish human-written from AI-generated poems, and tend to rate AI-generated poems more favorably. In that study, AI is not used for ideation; poems are created entirely by LLM, suggesting that fully AI-generated products can be experienced as highly creative by some consumers.

For visual products, Zhou and Lee (2024) analyze over four million artworks by more than 50,000 users on an online platform and show that, after the adoption of text-to-image generative AI, creative productivity rises by about 25% and favorites per view by roughly 50%. At the same time, average content novelty and average visual novelty decline, generating a more homogeneous idea space. This homogenization effect on creative results is also observed in other settings where technology is used to assist creative tasks. Oppenheimer and Patterson (2025) show that the use of a search engine leaves individual creativity unchanged while reducing collective creativity. Kumar et al. (2025) show that when an LLM is used as an assisted guide or as a direct answer provider in a creative writing task, creativity decreases relative to a condition without AI assistance. In summary, there is mixed evidence on whether AI-generated or AI-ideated output is more creative. However, a common theme in these studies is that these technologies lead to homogenization and lower average novelty, highlighting the potential downstream consequences of the extensive adoption of these technologies in creative tasks.

A key mechanism often mentioned in creativity research and psychology is fixation. Psychological research shows that exposure to examples can anchor people's mental representations and constrain their search for alternatives. Therefore, exposure to external ideas can actually diminish creativity. When participants are given examples of solutions to problems, they tend to be influenced by those answers and explore a narrower region of the idea space (Smith, 1995; Smith et al., 1993). Similarly, working in groups has been shown to inhibit creativity. Indeed, studies often show that the number of ideas generated by people brainstorming in actual groups is lower than the number of ideas generated by nominal groups (i.e., aggregating the ideas of people brainstorming by themselves; Diehl and Stroebe, 1987; Nijstad et al., 2003; Taylor et al., 1958; Kohn and Smith, 2011).

Generative AI can amplify these fixation dynamics. LLMs and related generative technologies sample from a list of possible next tokens. Unless directed otherwise, outputs tend to reflect central tendencies of the training distribution. Users can, in principle, push models toward more unusual

regions by prompting, but the initial model suggestions serve as strong priors. Once an AI system proposes a particular direction, users may fixate in that direction, refining around it rather than exploring more distant alternatives. Recent work on human-computer interaction describes this as a "creative fixation" on LLM outputs, particularly in more complex tasks (Xusen and Zhang, 2025). Related evidence from creative ad-copy writing shows that using an LLM as a ghostwriter can generate anchoring effects that reduce performance for expert users (Chen and Chan, 2024). This suggests that AI assistance during ideation may, in some settings, steer users toward particular directions and reduce the likelihood of highly creative and novel outcomes, even when users remain in the loop.

To our knowledge, there is no causal evidence on the impact of large language models or related generative AI systems in the context of music. Given that music composition experiments are hard to run at scale, no study has measured the effects of generative AI on musical creativity. This is particularly salient because creativity is an important component of both music creation and music enjoyment. Our study bridges this gap by examining how access to text-to-music models during ideation influences and potentially steers composers' stylistic choices. We then measure how listeners judge those choices in terms of creativity.

2.2 Generative AI models for music

AI models for music can be broadly categorized by the musical representation on which they are trained and the form of their output: symbolic models and audio (text-to-music) models. Symbolic models work with discrete encodings of musical structure, such as scores or Musical Instrument Digital Interface (MIDI) sequences that represent notes, rhythms, and harmonies. These systems output new symbolic representations that can be edited and refined. For example, Hookpad Aria (Donahue et al., 2025) is a symbolic generative AI model designed for Western pop songwriting. The model is intended for the composition of lead sheets (symbolic music scores that describe melody and harmony) and supports non-sequential workflows. Users can generate left-to-right continuations of their existing material, fill in missing spans in the middle, and generate harmony from melody and vice versa.

Audio-based models operate directly on sound, often in a text-to-music framework. Systems such as Udio and Suno, two leading commercial generative AI music products, are trained on large collections of audio recordings and associated text, and generate fully instrumented or vocal tracks in response to text prompts. In these systems, users describe their desired genre, moods, instrumentation, or reference artists, and the models generate complete audio clips. Although powerful, these tools typically offer less granular control over the musical structure of the output. The generated songs are finished-sounding tracks, and editing their structure requires external audio or MIDI tools. Furthermore, current commercial models are often "one-shot" and stateless. Each prompt generates new tracks without a persistent chat history (in contrast to systems like ChatGPT) or an explicit model of the user's evolving ideas.

Given the difference in output, the degree to which these tools can be incorporated into creative

workflows also differs. Symbolic models operate at the same level as many composers, support non-sequential workflows, and can "fill in the blanks" or extend existing material, making them well-suited for iterative refinement of a musical piece. Text-to-music models, in contrast, generate audio files that resemble finished songs and are harder to structurally modify. Because music composition is typically an iterative process, the one-shot nature and limited memory of such tools may inhibit iteration, diminish a sense of agency, or steer creators toward a narrower set of stylistic options. At the same time, text-to-music models may expand the idea space by providing quick, ready-made musical examples across a wide range of styles, potentially helping users overcome creative blocks or explore directions they would not have considered on their own, thereby primarily assisting in the divergent phase of the creative task.

2.3 Udio

In our study, we use Udio, an audio-based text-to-music model, in the second class described above. Udio is a web-based generative AI tool developed in 2023 by a team of former Google DeepMind researchers. The main feature of the platform is to create songs from text prompts that specify the desired genre, lyrics, mood, or reference artists. For each text prompt, the program generates two songs between 32 seconds and 2:10 minutes in length (chosen by the user), along with descriptive tags and track names. Users can adjust parameters such as output length, a "quality of generation" setting, and the model's adherence to the prompt. Users can also choose whether to include lyrics or generate purely instrumental music. Udio's prompt system does not provide a chat-like experience. Users cannot have a conversation with the agent, ask for modifications to a given output, or rely on a memory of previous requests. Instead, each generation is initiated by a new textual prompt, and the resulting audio tracks are self-contained. Udio also does not expose symbolic-level controls over pitch, rhythm, or harmony; any structural editing must be done outside the platform. We show the Udio user interface in Figure 1.

In a creative workflow, Udio is therefore most likely to be used as an ideation tool. Composers prompt the system, listen to the resulting tracks, and use them as references, sources of motifs, or high-level inspiration for their own work. These design features matter for our setting. Given that Udio generates polished audio, its output is likely to serve as strong anchors for tempo, density, contour, and other stylistic dimensions. For composers, the convergent stage now includes evaluating not only their own music but also Udio's output. This shifts the creator's role toward that of a critic or editor. Composers may select elements from Udio's tracks, internalize them, and reproduce these patterns, consciously or not, in their own melodies. At the same time, Udio expands the set of initial ideas available, providing more options during the divergent phase of the composition.

As discussed previously, music-based generative AI models are trained on large datasets of existing music. The generated melodies, while new to the user, may resemble known patterns, reducing the true novelty. On the other hand, AI systems can also produce odd combinations or less conventional sequences, potentially increasing surprise. Whether an AI-influenced output is ultimately judged as more creative by listeners will depend on the prompts composers use and on

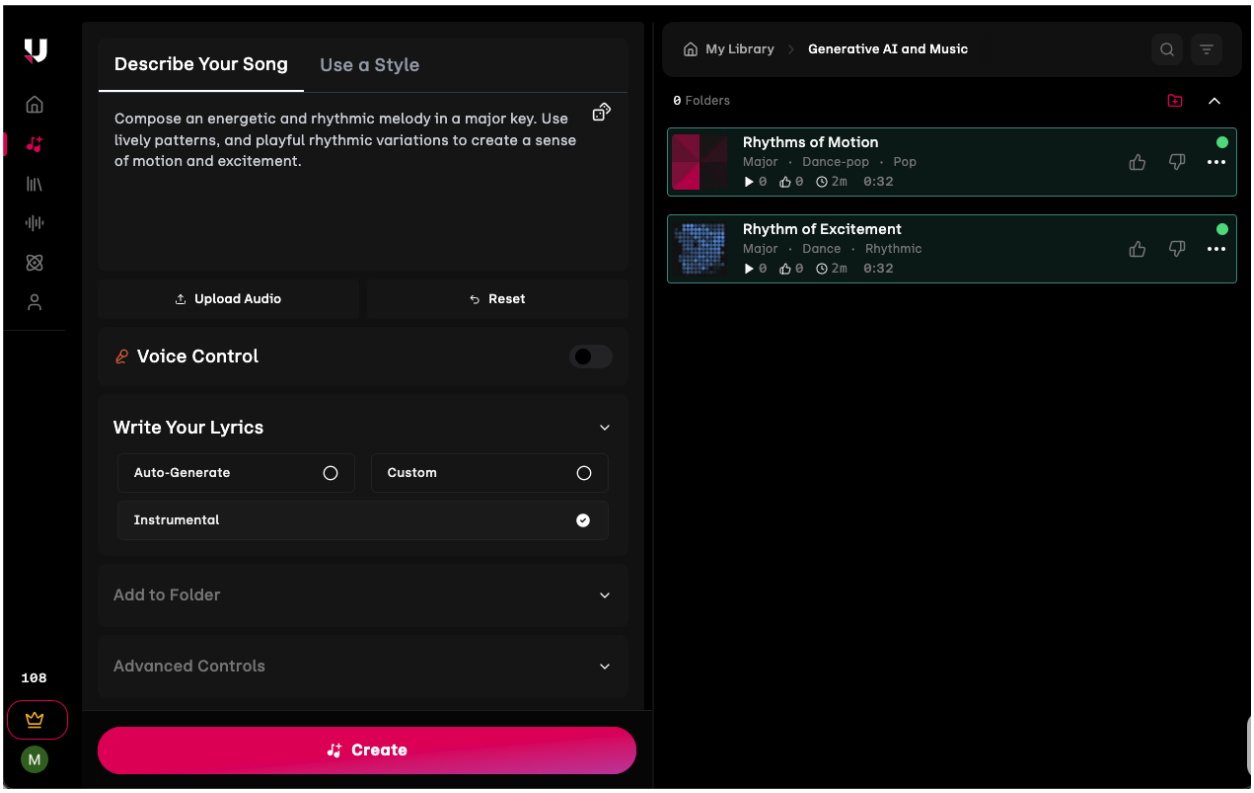


Figure 1: Udio User Interface for Creating Songs Based on a Song Description

how these factors balance out in practice.

Our experiment uses this setting to measure the net effect of providing access to Udio during ideation by comparing the melodies composed with and without the tool, their musical features, and how listeners evaluate them for creativity.

3 Methods

This section describes the design of our two experiments. Experiment 1 studies the composition of the melody under AI-assisted and control conditions, and Experiment 2 measures how external evaluators perceive the creativity of the resulting melodies.

3.1 Experiment 1 Procedure

We recruited 140 musically trained participants from Carnegie Mellon University (CMU) and the surrounding Pittsburgh area. The study received approval from the University's Institutional Review Board (IRB). Recruitment took place between March and August 2025 through several channels: we placed posters on the CMU, University of Pittsburgh, and Duquesne University campuses and in nearby neighborhoods; an email was circulated through the CMU School of Music mailing list; and a posting was created on CMU's online research participation pool website. All communication

informed prospective participants about the subject of the study, participation criteria, required tasks, compensation, and contact information, and included a link to an online screening survey.

The online screening survey asked prospective participants about their main instrument and years of formal training on that instrument, years of piano training, and whether they felt comfortable composing a melody on a keyboard. It also collected basic contact information. Based on these responses, we invited only those individuals who reported prior piano experience and indicated that they felt comfortable composing a melody on a keyboard. In total, 181 prospective participants signed up and completed the pre-screening survey. Of these, 170 were eligible to participate and were invited to the study: 144 participated, 5 were scheduled but later canceled, and 21 did not respond to further communication. Four participants were excluded from the final sample, two because they did not compose a monophonic piece and the other two because of technical issues with the recorded melody.

The invited participants scheduled an in-person session, with one participant per session and no interaction between participants. Sessions were held on weekdays in 50-minute time slots between 9 AM and 5 PM. Each session was randomly assigned to either the AI-ideation or control condition. Upon arrival, participants were informed about study procedures and were given a written consent form to sign. Each participant received a fixed payment of $20 USD for completing the study, regardless of treatment assignment or session outcome.

3.1.1 Experiment 1 Design

The experimental task was a constrained composition exercise in which each participant composed and recorded an original 15-second melody using a piano keyboard. To minimize the role of performance skill, we required that all melodies be monophonic: without accompaniment, harmony, or chords. This design isolates ideation from instrumental proficiency.

Participants were randomly assigned to one of two conditions: an AI-ideation condition or a control condition. In the AI-ideation condition, participants were given access to Udio, a web-based text-to-music generative AI platform, and were told they could use it during the composition process. In the control condition, participants composed and recorded the melody without access to any AI music tool. In both conditions, the final outcome of the session was a 15-second piano recording produced by the participant. The AI-generated songs from Udio were used only for ideation in the treatment condition. Of the 140 participants, 73 were randomly assigned to the AI-ideation condition and 67 to the control condition.

The sessions were conducted in person, and the melodies were captured using a 25-key Musical Instrument Digital Interface (MIDI) controller with two octaves (C–C). The keyboard was set to operate with octave shift equal to 0 and transpose equal to 0 semitones, so it recorded unshifted MIDI note numbers (with middle C equal to MIDI 60). The recordings were made using Cakewalk, a digital audio workstation (DAW), with the instrument sound set to a “Dark Grand” piano. No other actions were allowed in the DAW beyond recording and deleting whole takes; for example, participants could not manually edit notes or change instrumentation.

We limited the melody to 15 seconds. Previous studies on musical creativity use musical examples of varying length. Geringer and Johnson (2007) use 12 seconds as a lower bound, and Priest (2001) uses 20 seconds as a lower bound. Our choice of 15 seconds is between these two values. The length of a melody to be evaluated for creativity should be fixed or strongly limited, as length is known to modulate creativity ratings, and longer examples are typically rated higher than shorter examples (Loui et al., 2024; Wapnick et al., 2005; Geringer and Johnson, 2007).

All participants received the same initial textual prompt:

> "Compose an energetic and rhythmic melody in a major key. Use lively patterns and playful rhythmic variations to create a sense of motion and excitement."

They were told that they could use this prompt as an inspiration or ignore it if they preferred.

Each session followed a standardized sequence. First, participants were given details about the session and asked to complete the pre-task survey (described below). Then, instructions for the composition task were given along with the initial prompt. Participants were allowed up to 30 minutes for rehearsal and recording, and could re-record as many times as they wished within that window. In the AI-ideation condition, the computer display showed the DAW and the Udio web interface side by side, as shown in Figure 2. Participants in the treatment condition were instructed to enter the initial prompt into Udio and to listen at least once to the two songs generated. After this baseline interaction, they were free to create as many prompts as they desired and to incorporate ideas from the generated songs into their composition as they wished. Udio was configured to generate two 32-second instrumental (no vocals) audio clips per prompt. Control participants saw only the DAW and the prompt and did not interact with Udio at any point. After recording and submitting their final melody, all participants completed a post-task survey, were debriefed, and then paid.

3.1.2 Experiment 1 Self-Reported Surveys and Udio Data

We collected self-reported measures from participants before and after the composition task. All surveys were administered on the lab computer. In the pre-task survey, we collected basic demographic information and a baseline measure of self-perceived creativity. Participants reported their age, gender, current status (e.g., full-time student, employed, or other), whether they were currently enrolled in the School of Music at CMU, their major, and their current academic year, when applicable. The survey also included a baseline self-assessment of creativity, where we asked: "How creative do you feel you are relative to your peers?" on a 1–5 scale.

In the post-task survey, we captured participants' evaluations of the melody they had just composed, their experience with the task, and additional information about their musical background. All participants, regardless of treatment condition, answered the following questions on a scale of 1–5: "How much do you like the melody you composed?", "How satisfied are you with the final melody you composed?", "Compared to your peers, how creative do you consider your final melody to be?" and "How easy or difficult was it to come up with ideas for this melody?". They

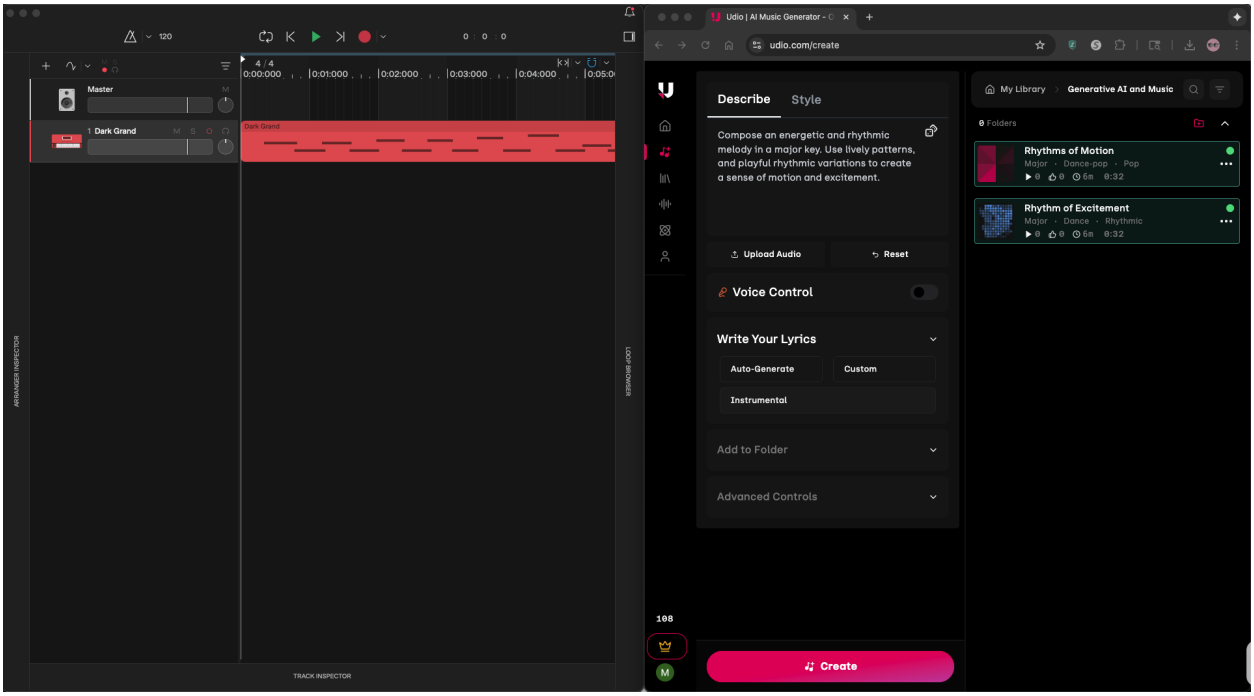


Figure 2: Split screen setup for treatment condition participants

also reported approximately how many minutes they spent working on their melody. In addition, participants were asked to briefly explain how they arrived at their melody and describe any feelings they experienced during the composition process.

The post-task survey also asked about musical experience and previous exposure to AI-based music tools. Participants reported their primary instrument (the instrument they play the most), the number of years of formal training with that instrument, and the number of years of experience playing the piano. They also rated, on a 1–5 scale, how skilled they felt in composition relative to their peers and how often they composed music. Participants were asked to rate their exposure to non-Western music. Additionally, they were asked whether they had ever used Udio before this task and whether they had used any other AI-based music tools (for example, Suno, MUREKA, Stability AI), with yes/no responses.

Participants in the AI-ideation condition answered additional questions about their interaction with Udio. On a scale of 1–5, they rated the following: "To what extent did you utilize the AI tool during your composition process? (1 = None, only used my own ideas; 5 = Extensive, heavily guided by AI suggestions)", "How helpful was the AI tool in generating ideas or composing your melody? (1 = Not at all helpful; 5 = Extremely helpful)", "How well did the AI tool's suggestions match the kind of melody or style you had in mind? (1 = Not at all aligned; 5 = Completely aligned)", "How do you feel the AI influenced your creativity? (1 = No influence; 5 = Significantly enhanced creativity)", "How much control did you feel you had over the final melody, as opposed to the AI controlling it? (1 = I felt I had almost no control; 5 = I felt I had complete control)", and "Overall, how would you rate the quality of the AI's musical suggestions? (1 = Very poor; 5 =

Excellent)". They also indicated whether they would use AI again for music composition and could provide additional comments about their experience composing with AI.

For participants in the AI-ideation condition, we also collected interaction logs from Udio. For each session, we recorded the text prompts entered into the system, the tracks generated in response, the descriptive tags Udio produced for each track, and the number of times each track was played by the participant. These logs summarize how participants engaged with the AI tool during ideation and provide measures of AI usage intensity and diversity of prompts.

3.2 Experiment 2 Procedure: Creativity Measurement

To measure how listeners perceive the melodies produced by participants in the composition task, we conducted a separate online evaluation study. We recruited 500 evaluators using Prolific[2] and collected 9,000 ratings. Each melody received at least 63 evaluations, with a median of 64 and a maximum of 68. The assessors were required to be at least 18 years old, fluent in English, have completed secondary or higher education, and report no hearing difficulties. All evaluators were located in the United States. We report summary statistics for the demographic characteristics of the evaluators in Appendix Table A1. Although musical experience was not required to participate in the evaluation, 54% of participants reported some musical training, with 34% reporting four or more years of training on their primary instrument. Furthermore, 34% reported some piano experience, with 18% reporting four or more years. The most common primary instrument was piano (20%), followed by guitar (11%) and woodwind (7%). Additionally, 22% of the evaluators reported performing some level of music composition, with 6% composing at least a few times a month. Our sample is 49% female and is predominantly white (71%), and 92% were born in the United States. Median age is 40, and the majority have some employment (50% full time, 17% part time) and a college degree (61%).

Each evaluator completed a listening task in which they rated a sample of the melodies generated in the composition task. On each survey, evaluators were randomly assigned 18 distinct melodies from the pool of 140, without replacement within evaluator and without any information about the source of the melodies. The order of melodies was randomized for each evaluator, and melodies were presented sequentially. For every melody, participants were instructed to listen at least once, with the option to replay if desired. The survey interface was designed so that ratings could only be submitted after the melody had been played in its entirety at least once. After listening, participants answered the following question on a 5-point Likert scale: "How creative do you find this melody?". Responses ranged from 1 ("Not at all") to 5 ("Extremely"). Evaluators were instructed to use the full scale in their assessment. Each evaluator received a fixed payment of $3 USD for completing the survey.

Our measurement of creativity follows a product-based, consensual assessment approach. Following Amabile's Consensual Assessment Technique (CAT), we operationalize creativity as the degree to which appropriate observers independently agree that an output is creative in its context (e.g.,

[2] `https://www.prolific.com`

Amabile, 1996). Instead of relying exclusively on domain experts, we intentionally use a sample of typical listeners, since our focus is on perceived creativity as experienced by a general audience. Prior work using CAT indicates that undergraduate and other non-expert judges can provide reliable and valid ratings of creative products, and that their judgments often align with those of experts, especially when the task is to compare relative creativity within a set of stimuli rather than to enforce a narrow domain-theoretic definition of creativity (e.g., Amabile, 1983; Sternberg and Lubart, 1995; Chen et al., 2005).

For our main specification, we use the mean rating for each melody as the dependent variable. Each melody is rated by between 63 and 68 evaluators (mean = 64.3, median = 64), which yields reasonably precise estimates of its mean creativity score. The typical within-melody standard deviation of individual creativity ratings is about 1.06 on the 1–5 scale. This implies an average standard error of a melody's mean of roughly $1.06/\sqrt{64.3} \approx 0.13$, which corresponds to the expected sampling noise around the mean for a single melody. A corresponding 95% interval for the mean of a melody is therefore approximately the mean ± 0.26. In contrast, the cross-melody distribution of mean creativity scores has a standard deviation of about 0.47, indicating substantive between-melody variation beyond sampling error. Consistent with this, a linear mixed-effects model with random intercepts for melodies and evaluators yields an intraclass correlation for the mean creativity score of 0.93 given an average of 64.3 ratings per melody, indicating high reliability of the aggregated score.

For each melody, we construct outcome measures by averaging the ratings across all evaluators who heard that melody. Our main dependent variable is the mean creativity score for each melody. Treating the 5-point scales as approximately continuous is standard in this literature and is justified by the number of raters per melody and the observed reliability of the ratings.

3.3 Musical Features

Each melody generated by participants during the composition task is stored in MIDI format, a symbolic representation of music. A melody consists of a sequence of notes, and each note is represented by its pitch (MIDI note number), onset time, duration, and velocity (intensity on a 0–127 scale). Using this symbolic representation, we summarize each melody with a set of musical features computed with the Python library `music21` (Cuthbert and Ariza, 2010). These features describe the melody along several musical dimensions: density and timing, pitch and register, melodic contour, and tonal strength.

We include density and timing features such as the total number of notes, the mean and standard deviation of note durations, and the mean and standard deviation of inter-onset intervals (the time between successive note onsets). We also calculate an entropy measure over note durations to capture rhythmic variety. These features describe how dense or sparse the melody is in time, and how regular or varied its rhythmic surface is.

We extract pitch and register features, including the mean MIDI pitch (overall register) and the pitch range in semitones (maximum minus minimum pitch). These measures capture whether a

melody sits higher or lower on the keyboard and the range of pitches it covers.

We also include measures of contour and interval that summarize the shape of melodic motion. These capture melodic intervals between successive notes (for example, the mean absolute interval size) and the rate at which the contour moves in small steps versus larger leaps.

Tonal strength features measure how strongly the melody supports a stable key. In particular, we use a scalar measure of tonal strength, which describes the degree to which the distribution of pitches aligns with a clear major or minor key.

A complete list of the musical features, along with definitions and summary statistics by treatment condition, is provided in Table 7. We use these features both to compare melodies composed with and without AI access and to understand how they are associated with creativity ratings.

3.4 Empirical Strategy

We analyze our data in four steps. First, we estimate intent-to-treat (ITT) effects of AI access on listeners' ratings. For each average melody rating Y_i, we estimate

$$Y_i = \alpha_0 + \alpha_1 T_i + \mathbf{X}_i^\top \boldsymbol{\alpha} + \epsilon_i, \tag{1}$$

where $T_i = 1$ if melody i was produced by a participant assigned to the AI-ideation (Udio) condition and $T_i = 0$ otherwise, and $\mathbf{X}_i$ is a vector of composer characteristics (musical background, prior AI exposure, demographics, and session timing). The coefficient α_1 identifies the causal intent-to-treat effect of AI access on the average rating of the melody, since the composers were randomly assigned to the experimental condition. We estimate equation (1) by ordinary least squares (OLS).

Second, we estimate local average treatment effects (LATE) of engaging with the AI tool beyond the required baseline prompt. Participants in the treatment condition can prompt Udio as much or as little as they want during their composition process. As a result, some participants choose not to engage with Udio beyond the initial required prompt, and the decision to engage is endogenous. We therefore use treatment assignment as an instrumental variable to estimate local average treatment effects of AI usage on outcomes. We consider two measures of AI usage U_i: (1) a binary indicator equal to 1 if the composer generated at least one prompt in addition to the initial baseline prompt and 0 otherwise, and (2) a self-reported usage-intensity measure based on the question "To what extent did you utilize the AI tool during your composition process?". For each usage measure, we estimate a two-stage least squares model,

$$U_i = \pi_0 + \pi_1 T_i + \mathbf{X}_i^\top \boldsymbol{\pi} + v_i, \tag{2}$$

$$Y_i = \beta_0 + \beta_1 U_i + \mathbf{X}_i^\top \boldsymbol{\beta} + \varepsilon_i. \tag{3}$$

The first regression serves as the first stage of our instrumental variable estimation, and the second equation as the outcome equation. The coefficient β_1 identifies a local average treatment effect, which we interpret as the average causal effect on Y_i of engaging with Udio beyond the baseline for compliers, that is, participants who engaged with Udio due to the treatment assignment.

Third, we examine how AI access changes the musical features of the melodies. For each feature Z_{ki} described in Section 3.3, we estimate a separate OLS regression of that feature on the treatment indicator,

$$Z_{ki} = \delta_{0k} + \delta_{1k} T_i + \nu_{ki} \tag{4}$$

and report treatment and control means, treatment-control differences, and HC3-robust standard errors. Again, random assignment of treatment ensures that δ_{1k} is an unbiased estimate of the ITT effect of AI access on feature k. We control the false discovery rate at $q = 0.05$ using the Benjamini–Hochberg procedure, which is standard under positive dependence among tests.

Finally, we relate musical features to listener ratings by regressing their melody-level means on the vector of standardized musical features. For each outcome, we estimate

$$Y_i = \gamma_0 + \mathbf{Z}_i^\top \boldsymbol{\gamma} + \eta_i, \tag{5}$$

where Y_i is the mean rating of melody i, and $\mathbf{Z}_i$ is the vector of standardized musical features for melody i. The coefficients $\boldsymbol{\gamma}$ describe which aspects of timing, density, pitch, contour, and tonal strength are most strongly associated with higher or lower ratings on each dimension.

4 Results

4.1 Summary Statistics

In Table 1, we report summary statistics for baseline covariates by treatment condition for participants in Experiment 1. These variables measure musical background (piano and primary-instrument experience, self-reported composition skill, and whether the participant is a CMU School of Music major), prior exposure to AI for music (Udio and other tools), demographics (gender, age, work/student status), and session timing (morning vs. afternoon; day of week). Treatment and control are well balanced on these characteristics, and we find no statistically significant differences by treatment arm on any covariate (all p-values > 0.15).

Most participants list piano or keyboard as their main instrument (about 54%). Over half report at least eight years of piano experience (about 55%), and about two thirds report eight or more years of experience on their primary instrument (about 64%). About 54% of participants report a high exposure to non-Western music. Prior Udio use is uncommon (about 7%), and around 20% report experience using other AI music tools. The sample is predominantly composed of students (about 82%), and most are between ages 18 and 24 (about 69%). Sessions occurred on weekdays and were mostly scheduled in the afternoon (12–5 pm), with about one third beginning in the morning (9–11 am).

Variable	Control (N=67)	Treatment (N=73)	All (N=140)	p
Self-reported Outcomes				
Composition task duration	19.82 (8.29)	21.32 (7.20)	20.60 (7.75)	0.262
Self-reported Melody Enjoyment	3.85 (0.74)	3.81 (0.76)	3.83 (0.75)	0.740
Self-reported Melody Satisfaction	3.79 (0.71)	3.71 (0.75)	3.75 (0.73)	0.528
Self-reported Melody Creation Difficulty	3.94 (1.00)	3.99 (1.03)	3.96 (1.01)	0.791
Self-reported pre-task creativity	3.57 (0.81)	3.52 (0.85)	3.54 (0.83)	0.743
Self-reported post-task creativity	3.06 (0.83)	3.08 (0.86)	3.07 (0.85)	0.876
Covariates				
Main Instrument: Piano/keyboard	0.57 (0.50)	0.52 (0.50)	0.54 (0.50)	0.586
Main Instrument: Bowed strings	0.15 (0.36)	0.12 (0.33)	0.14 (0.34)	0.660
Main Instrument: Woodwind	0.04 (0.21)	0.11 (0.31)	0.08 (0.27)	0.153
Main Instrument: Other	0.24 (0.43)	0.25 (0.43)	0.24 (0.43)	0.916
Piano experience 1–3 yrs	0.31 (0.47)	0.29 (0.46)	0.30 (0.46)	0.744
Piano experience 4–7 yrs	0.16 (0.37)	0.14 (0.35)	0.15 (0.36)	0.659
Piano experience $\geq$8 yrs	0.52 (0.50)	0.58 (0.50)	0.55 (0.50)	0.536
Main Instrument experience 1–3 yrs	0.16 (0.37)	0.11 (0.31)	0.14 (0.34)	0.357
Main Instrument experience 4–7 yrs	0.18 (0.39)	0.26 (0.44)	0.22 (0.42)	0.252
Main Instrument experience $\geq$8 yrs	0.66 (0.48)	0.63 (0.49)	0.64 (0.48)	0.747
Self-reported Composition skill (self)	2.93 (1.15)	3.04 (1.18)	2.99 (1.16)	0.561
High non-Western music exposure	0.48 (0.50)	0.59 (0.50)	0.54 (0.50)	0.193
CMU School of Music major	0.24 (0.43)	0.26 (0.44)	0.25 (0.43)	0.773
Used Udio (past)	0.09 (0.29)	0.05 (0.23)	0.07 (0.26)	0.436
Used other AI for music (past)	0.19 (0.40)	0.21 (0.41)	0.20 (0.40)	0.868
Female	0.43 (0.50)	0.48 (0.50)	0.46 (0.50)	0.586
Age 18–24	0.67 (0.47)	0.71 (0.46)	0.69 (0.46)	0.608
Age 25–34	0.24 (0.43)	0.19 (0.40)	0.21 (0.41)	0.506
Age 35–44	0.03 (0.17)	0.04 (0.20)	0.04 (0.19)	0.723
Student	0.81 (0.40)	0.84 (0.37)	0.82 (0.38)	0.653
Full-time worker	0.07 (0.26)	0.11 (0.31)	0.09 (0.29)	0.480
Part-time worker	0.04 (0.21)	0.04 (0.20)	0.04 (0.20)	0.916
Task start period (Morning=1/Afternoon=0)	0.30 (0.46)	0.34 (0.48)	0.32 (0.47)	0.583
Task day of week	1.88 (1.45)	1.78 (1.48)	1.83 (1.46)	0.690

Table 1: Self-reported Outcomes and Demographic Balance by Arm (HC3 robust).

4.2 Self-reported Creative Process and Composition Experience

As described in Section 3.1.2, participants in Experiment 1 completed a short survey about their experience during the composition task and their own evaluation of the melody. In particular, we asked them to rate, on a 5-point Likert scale: how much they enjoyed the melody they composed, how satisfied they were with their melody, how creative they considered their final melody to be, and how easy it was to come up with ideas for the melody. Table 1 reports the average scores by treatment status. We find no statistically significant differences between treatment and control.

Self-reported melody enjoyment is 3.81 (SD = 0.76) in treatment vs. 3.85 (0.74) in control. Satisfaction with the final melody is 3.71 (0.75) in treatment and 3.79 (0.71) in control. Ease of generating ideas is similar across arms as well (3.99 (1.03) in treatment vs. 3.94 (1.00) in control, where higher values indicate greater ease). Self-rated creativity relative to peers shows comparable levels before and after the task in both conditions, with a small average decline that is similar in magnitude across arms. The results are robust to including composer covariates in regression models. We find no evidence that interacting with the AI model during composition changes how much composers like their own melody, how satisfied they are with it, how difficult they found it to come up with ideas for their final melody, or how creative they judge their final melody to be.

4.3 Main Intent-to-Treat Effects on Creativity Ratings

Table 2 Column (1) reports intent-to-treat estimates of the effect of AI access on mean creativity ratings, using data collected in Experiment 2 (N = 500 evaluators). The coefficients are estimated using Equation (1) and OLS, with composer covariates for musical background, CMU School of Music enrollment, prior AI exposure, demographics (age, gender, and employment status), and session timing.

Table 2: Treatment Effects on Creativity Ratings

	(1) Creativity (Melody level)	(2) Creativity (Evaluation level)
Intercept	2.703***	
	(0.208)	
Treatment	−0.154*	−0.182**
	(0.080)	(0.075)
Num.Obs.	140	9000
R2	0.224	0.248
Controls	Yes	Yes
Evaluator FE	No	Yes

* p < 0.1, ** p < 0.05, *** p < 0.01

Giving participants access to the text-to-music model Udio during ideation reduces listener ratings. The estimated treatment effect on mean creativity is −0.154 points on the 1–5 scale (SE = 0.080, $p = 0.055$), which is marginally significant at the 5% level. This corresponds to about 0.3 standard deviations of the cross-melody distribution of mean creativity scores, and the implied 95% confidence interval is approximately $[-0.31, 0.00]$. Table 2 Column (2) uses individual evaluator

ratings as the unit of observation rather than melody-level means, and adds evaluator fixed effects to absorb differences in rating tendencies across listeners. We report two-way clustered standard errors by melody and evaluator. The estimate is similar in magnitude and remains significant. It is surprising that the effect is significant and has a negative impact. Our results indicate that, under the Udio condition, users are evaluated as generating less creative music. However, not all users in the Udio condition report using it for ideation. We now examine how the actual use of Udio affects creativity.

4.4 Local Average Treatment Effect (LATE)

We first provide summary statistics on Udio usage. Table 3 reports engagement with Udio among treated participants. Most participants in the AI-ideation condition (82%) generated at least one additional prompt beyond the required baseline prompt. On average, treated participants generated 3.55 prompts (median = 3), leading to an average of 7.22 Udio tracks (median = 6). The average total number of plays per session was 8.41, and most tracks were listened to at least once.

Variable	All mean (SD)	Median	Min	Max	N
Self-reported AI usage outcomes					
AI usage for composition overall (higher = more usage)	2.19 (1.08)	2.00	1.00	5.00	73
AI helpfulness in ideation (higher = more helpful)	2.58 (1.34)	2.00	1.00	5.00	73
AI fit with intended style (higher = better fit)	2.36 (1.21)	2.00	1.00	5.00	73
AI influence in creativity (higher = more enhancing)	2.49 (1.20)	2.00	1.00	5.00	68
AI vs human control (higher = more human control)	4.40 (1.06)	5.00	1.00	5.00	73
AI suggestions quality (higher = better quality)	2.78 (1.06)	3.00	1.00	5.00	73
Use AI again for music composition? (Yes)	0.25 (0.43)	0.00	0.00	1.00	73
Use AI again for music composition? (No)	0.30 (0.46)	0.00	0.00	1.00	73
Use AI again for music composition? (Maybe)	0.45 (0.50)	0.00	0.00	1.00	73
Observed Udio activity					
Number of Udio prompts created	3.55 (1.98)	3.00	1.00	9.00	73
Number of Udio prompts created (beyond baseline)	2.55 (1.98)	2.00	0.00	8.00	73
Number of Udio tracks generated	7.22 (3.97)	6.00	2.00	18.00	73
Number of Udio plays	8.41 (5.23)	7.00	0.00	24.00	73
Created at least one more prompt (beyond baseline)	0.82 (0.39)	1.00	0.00	1.00	73

Table 3: Summary Statistics of AI usage

Next, we compare simple averages across three groups: control (no access to AI), treated participants who generated at least one additional prompt beyond the required baseline, and treated participants who did not generate any additional prompts. Table 4 reports means and standard deviations for each group. The mean creativity rating is 3.19 in the control condition, 3.04 among prompters, and 3.24 among non-prompters. Creativity ratings are lowest among prompters, while non-prompters in the treatment condition receive ratings similar to those of control users. This pattern suggests that when participants actively use Udio for ideation, their creativity is evaluated as inferior. We formally test this below.

We estimate the causal effect of engaging with Udio beyond the baseline among treated composers. To avoid the selection problem, we implement the instrumental variables strategy described in

Variable	Control mean (SD)	Generated at least one prompt (N=60) mean (SD)	No prompts (N=13) mean (SD)
Creativity	3.19 (0.46)	3.04 (0.50)	3.24 (0.40)

Table 4: Outcomes by Observed Udio Usage

Section 3.4, using randomized assignment to the AI-ideation condition as an instrument for AI usage. We consider two measures of AI usage. The first is a binary indicator equal to one if the composer generated at least one prompt in addition to the baseline prompt. The second is a binary indicator for high self-reported reliance on AI during the task, defined as reporting a score of at least 3 (on a 1–5 scale) on the question "To what extent did you utilize the AI tool during your composition process?".

Table 5 reports the IV estimates for creativity. First-stage regressions show that assignment to the AI-ideation condition significantly increases both prompt-based and self-reported measures of AI usage, indicating a strong first-stage effect. Using the "generated at least one additional prompt" indicator as the endogenous regressor, the 2SLS estimate implies that, for compliers, engaging with Udio beyond the baseline reduces mean creativity ratings by about 0.18 points (SE = 0.094, $p = 0.055$), marginally significant at the 5% level and slightly larger in magnitude than the ITT effect. When we instead use the high self-reported usage indicator, the IV estimate is larger in magnitude (about -0.423, SE = 0.233, $p = 0.072$), also marginally significant at the 10% level. The LATE results are consistent with the ITT findings and suggest that heavier reliance on the tool causally lowers the perceived creativity of the resulting melody for compliers.

Table 5: IV estimates for creativity

	IV: Udio	IV: Self-Reported AI usage
Intercept	2.704***	2.747***
	(0.207)	(0.226)
Udio: Generated at least 1 prompt	−0.182*	
	(0.094)	
Self-Reported: High AI usage		−0.423*
		(0.233)
Num.Obs.	140	140
R2	0.227	0.121
R2 Adj.	0.104	−0.018
Controls	Yes	Yes

* p < 0.10, ** p < 0.05, *** p < 0.01

4.5 Composer Heterogeneity

We explore whether the effect of AI access varies with composer characteristics by extending Equation (1) to include Treatment×Moderator interactions. In a single specification, we interact the treatment indicator with four composer-level moderators: self-reported composition skill, prior experience with AI music tools, having piano as main instrument, and current enrollment as a CMU

School of Music major. The resulting interaction coefficients test whether the intent-to-treat effect differs systematically across levels of these moderators. We report the estimated coefficients in Appendix Table A2.

Because these moderators were not pre-specified, we treat this heterogeneity analysis as exploratory and adjust for multiple testing across the four interaction terms using the Benjamini–Hochberg (BH) procedure. None of the interactions remains statistically significant after BH adjustment. The interaction between treatment and being a School of Music major is statistically significant at the 5% level before correction ($p = 0.03$) and suggests that access to Udio is associated with a more negative effect on creativity ratings. Since School of Music majors tend to have more extensive musical training and experience than the rest of the participants in the sample, this pattern suggests that more experienced composers may see a larger drop in perceived creativity when they use the AI tool. However, the BH-adjusted p-value for this interaction is 0.124, so we consider this suggestive rather than conclusive. The remaining interaction terms are not statistically different from zero even before adjustment, and overall we do not find strong evidence of systematic heterogeneity in the effect of AI access across other composer characteristics.

4.6 Evaluator Heterogeneity

Evaluators differ in observed characteristics, including musical experience, formal training, and composition experience. We examine whether evaluators' musical and compositional backgrounds moderate the treatment effect on creativity ratings. For this, we estimate:

$$Y_{ij} = \alpha_0 + \alpha_1 T_i + \alpha_2 \left(T_i \times Experience_j\right) + \mathbf{X}_i^\top \boldsymbol{\beta} + \mu_j + \epsilon_{ij}, \tag{6}$$

where Y_{ij} is evaluator j's creativity rating of melody i, T_i is treatment assignment, $\mathbf{X}_i$ is a vector of controls, and μ_j denotes evaluator fixed effects. The interaction coefficient α_2 captures how the treatment effect differs for evaluators with $Experience_j = 1$ relative to evaluators with $Experience_j = 0$. We estimate this specification in three separate regressions, defining $Experience_j$ as: (i) any musical experience, (ii) formal musical training (self-reported formal vs. informal/none), and (iii) any composition experience.

We report the estimates from these three models in Table 6. Across all three models, the interaction term is positive but close to zero and statistically insignificant. This indicates that differences in evaluators' musical and compositional experience do not significantly moderate the perceived creativity of treatment melodies.

4.7 Effect on Musical Features

We now assess how AI access affects the musical features of the melodies described in Section 3.3. Table 7 reports, for each feature, the mean in the full sample and by treatment status, as well as the treatment–control difference (T–C) and the corresponding Benjamini–Hochberg q-values. The T–C coefficients are obtained from separate OLS regressions of each feature on the treatment indicator,

Table 6: Treatment heterogeneity by evaluator characteristics

	(1) Musical Exp.	(2) Formal Training	(3) Composition Exp.
Treatment	−0.211**	−0.185**	−0.192**
	(0.081)	(0.076)	(0.079)
Treatment × Experience	0.054	0.012	0.046
	(0.047)	(0.052)	(0.060)
Num.Obs.	9000	9000	9000
R2	0.248	0.248	0.248
Evaluators in group 0	230	363	391
Evaluators in group 1	270	137	109
Controls	Yes	Yes	Yes
Evaluator FE	Yes	Yes	Yes

* p < 0.1, ** p < 0.05, *** p < 0.01

as specified in Equation (4).

Variable	All mean (SD)	Control mean (SD)	Treat. mean (SD)	T-C	q (BH)
Notes: count	44.45 (13.60)	47.37 (13.18)	41.77 (13.51)	-5.61	0.036
Melody duration: total length (sec)	15.57 (1.86)	15.53 (1.74)	15.61 (1.97)	0.08	0.952
Pitch: mean MIDI	67.96 (6.33)	67.93 (5.98)	67.98 (6.67)	0.06	0.958
Pitch: range (semitones)	16.08 (4.33)	17.13 (4.38)	15.11 (4.08)	-2.02	0.016
Pitch entropy: scale-relative (0–1)	0.75 (0.17)	0.75 (0.18)	0.74 (0.17)	-0.00	0.952
Melodic interval: mean (semitones)	3.16 (1.30)	3.25 (1.44)	3.08 (1.15)	-0.17	0.757
Melodic interval: share ≤2 st	0.37 (0.15)	0.36 (0.17)	0.37 (0.13)	0.01	0.905
Melodic contour: arc size mean (semitones)	6.55 (2.49)	6.76 (2.53)	6.35 (2.45)	-0.41	0.691
Melodic interval: chromatic motion share (0–1)	0.12 (0.09)	0.12 (0.10)	0.13 (0.09)	0.01	0.905
Note duration: mean (sec)	0.40 (0.19)	0.33 (0.15)	0.46 (0.20)	0.14	0.000
Note duration: SD (sec)	0.26 (0.15)	0.22 (0.11)	0.29 (0.17)	0.07	0.009
Inter-onset interval: mean (sec)	0.41 (0.15)	0.37 (0.12)	0.45 (0.16)	0.08	0.005
Inter-onset interval: SD (sec)	0.24 (0.13)	0.20 (0.10)	0.27 (0.15)	0.07	0.007
Tonal strength: best-key correlation	0.84 (0.08)	0.85 (0.08)	0.83 (0.09)	-0.01	0.691

Table 7: Treatment effects on musical characteristics (N = 140).

We observe several differences in musical features between AI and control melodies. Note density is lower in the treatment condition, with treated melodies containing on average 41.77 notes compared to 47.37 in control, an estimated difference of −5.61 notes ($q = 0.036$). Pitch range is also narrower: the average range is 15.11 semitones in treatment versus 17.13 in control, a reduction of about 2 semitones ($q = 0.016$). By contrast, mean pitch level is unchanged across conditions.

Most notably, we observe clear shifts in timing and density features. Average note durations are longer in the treatment condition (0.46 seconds vs. 0.33 seconds, an estimated difference of 0.14 seconds, $q < 0.001$). There is also greater variability in note duration, with the standard deviation larger by 0.07 seconds ($q = 0.009$). Inter-onset intervals (the time between note onsets) are also longer, with a mean difference of 0.08 seconds ($q = 0.005$) and are more variable, with a standard deviation difference of 0.07 seconds ($q = 0.007$). These results indicate that melodies composed under the treatment condition are, on average, slower and sparser in time.

In contrast, we do not detect statistically significant effects on tonal strength and measures of

contour and interval. As a robustness check, we compute the effects under Holm and Bonferroni corrections. The main timing shifts remain significant under these stricter corrections, while the note count and pitch range differences are attenuated. In general, AI access steers composers toward slower, sparser, and narrower-range melodies while leaving tonal strength, mean register, and contour characteristics unchanged.

4.8 Musical Features and Listener Ratings

The previous section showed that access to Udio shifts aspects of melodic features, particularly timing, density, and pitch range. We now examine how these musical features relate to listener ratings, in order to understand whether the structural changes induced by access to AI move melodies toward or away from musical aspects that listeners reward.

We first look at simple bivariate relationships between musical features and creativity ratings. In Appendix Figure A1 we plot each musical feature against the mean creativity rating and report Pearson (r) and Spearman (ρ) correlations. Temporal density and timing features show the strongest associations with creativity. Melodies that contain more notes tend to receive higher creativity scores (note count: $r = 0.7$, $\rho = 0.70$), while sparser melodies are associated with lower scores. In particular, longer inter-onset intervals are strongly negatively associated with creativity ($r = -0.70$, $\rho = -0.72$), and longer average note durations are also negatively correlated ($r = -0.61$, $\rho = -0.57$). Pitch and register measures are moderately associated with creativity: melodies with a wider pitch range are rated as more creative (range: $r = 0.33$) and, to a lesser degree, so is higher tonal strength ($r = 0.24$). Melodic contour and interval measures (mean interval size and average arc size) have weak or near-zero correlations with creativity.

We then estimate a model relating creativity ratings to all musical features jointly, as described in Section 3.4. Table 8 reports estimates from an OLS regression of mean creativity on standardized musical features. Appendix Figure A2 shows the corresponding coefficient plot. This model explains more than half of the variation in creativity (adjusted $R^2 \approx 0.6$).

Several features remain strongly associated with creativity after controlling for the others. Note count has a highly significant positive association (standardized coefficient ≈ 0.21, $p < 0.01$), and mean inter-onset interval remains strongly negative (coefficient ≈ -0.21, $p < 0.05$). Mean pitch, pitch range, and tonal strength are positively associated with creativity. Average note duration, which is negatively correlated with creativity in the bivariate analysis, has a coefficient near zero once other features are included. This pattern suggests that the simple negative correlation largely reflects the fact that sparser, slower melodies tend to use longer notes and are rated as less creative. Melodic interval and contour measures contribute little once the other features are accounted for.

To understand where the predictive power of the model comes from, we add blocks of features sequentially and track changes in R^2. With timing features alone (note durations and inter-onset intervals), the model explains roughly 0.52 of the variance in creativity scores. Adding density, pitch/register, contour, and tonality (the full model) increases R^2 to about 0.63. Decomposing the explained variance relative to the full model, timing accounts for about 82% of the explained

variance, density adds about 7%, pitch and register about 5%, contour about 3%, and tonality about 2%. Because these blocks are correlated, the precise attribution depends on the order in which they are added. Nonetheless, the results indicate that timing and temporal features account for most of the model's predictive signal for perceived creativity in this setting. The overall level of explained variance is sizable for subjective ratings of short melodies, suggesting that these musical features capture dimensions that evaluators use when forming creativity judgments.

These results, combined with the treatment effects on musical features in Section 3.3, show that AI access pushes participants toward melodies characterized by fewer notes, longer inter-onset intervals, and narrower pitch ranges, which are associated with lower creativity ratings. The same density and timing dimensions that Udio shifts most strongly are also those that listeners appear to use most when judging creativity, which is consistent with AI-assisted melodies receiving lower creativity scores on average.

Table 8: OLS creativityRating on Standardized Musical Features

	Creativity (std. coeff.)
Intercept	3.131***
	(0.025)
Notes: count	0.212***
	(0.058)
Melody duration: total length (sec)	−0.030
	(0.033)
Pitch: mean MIDI	0.057**
	(0.027)
Pitch: range (semitones)	0.070**
	(0.034)
Melodic interval: mean (semitones)	0.019
	(0.034)
Melodic contour: arc size mean (semitones)	−0.068*
	(0.038)
Note duration: mean (sec)	0.006
	(0.057)
Inter-onset interval: mean (sec)	−0.211**
	(0.084)
Inter-onset interval: SD (sec)	0.120***
	(0.040)
Tonal strength: best-key correlation	0.060**
	(0.029)
Num.Obs.	140
R2	0.631
R2 Adj.	0.602

Standard errors in parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

4.9 Within-Group Variation and Similarity

We now assess whether AI access makes participants' melodies more homogeneous or more diverse within the treatment group, relative to the control group. Prior work on generative AI has

documented both convergence and diversification effects in creative outputs. In our setting, Udio could lead composers toward similar stylistic choices, for example if its outputs reflect central tendencies of the training data or if users gravitate toward high-probability patterns, which would increase within-group similarity. At the same time, access to many quickly generated examples could widen the set of ideas that participants consider and increase dispersion, thereby decreasing within-group similarity. Therefore, the effect of exposure to AI on within-group similarity is unclear.

To study this, we compare within-group similarity between treatment and control using two approaches and report the results in Appendix Table A3. First, we use the symbolic, theory-aligned musical features extracted from the MIDI files and described in Section 3.3. We standardize each feature (z-scores) and compute pairwise distances between melodies using cosine distance and, as a robustness check, Euclidean distance. For this approach, larger values correspond to greater dissimilarity between melodies. For each feature block and distance metric, we compute the mean distance among all treated pairs and among all control pairs. We then test the null of equal within-group means using a permutation test that randomly shuffles treatment labels and recalculates the mean difference. We compute two-sided p-values based on 5,000 permutations.

For the second approach, we use audio-based similarity measures implemented in MiRA (Batlle-Roca et al., 2024), a model-independent tool designed to detect how closely a generated audio sample resembles a reference sample. We compute three pairwise metrics for each pair of audio files: (i) CoverID distances from Essentia for cover-song identification, which aligns pitch-content while allowing for tempo and instrumentation changes (lower values indicate higher similarity); (ii) a symmetric Kullback–Leibler divergence between PaSST classifier posterior distributions (lower values again indicate higher similarity); and (iii) the CLAP score, which compares melodies in a joint audio–text embedding space using cosine similarity (higher values indicate higher similarity). Thus, for CoverID and KL, larger numbers mean more dissimilar (less similar) pairs, whereas for CLAP, smaller numbers mean more dissimilar pairs. As with the feature-based distances, we compare mean within-group (treatment vs. control) values for each metric and obtain permutation-based p-values under random reassignments of the treatment labels.

Using the cosine distance on the full set of standardized features, we find that overall within-group dispersion is very similar across arms. The mean within-group cosine distance is 0.993 for the treatment group and 0.972 for the control group, a difference of 0.021 with a permutation p-value of 0.036. This suggests a modest increase in dispersion under AI access that is significant at the 5% level. The Euclidean analogue is also small and not significant (5.271 vs. 4.983, difference 0.288, $p = 0.446$). Overall, the distribution of melodies in feature space is not dramatically more or less spread out in the treatment group.

However, the pattern differs when we compute distances within musical feature blocks. For rhythm-related features (note durations and inter-onset intervals), the mean within-group cosine distance is 0.968 for treated melodies and 0.861 for control, a difference of 0.107 ($p = 0.002$). This corresponds to a 12% increase in within-group distance relative to the control mean, indicating that AI access increases dispersion along rhythm and timing dimensions. In contrast, several other blocks

show small convergence in the treatment group. For density features, the mean cosine distance is 0.986 in treatment and 1.006 in control, a difference of -0.021 with $p = 0.067$, which corresponds to about a 2% reduction in within-group distance. For interval and contour features, the mean is 0.968 in treatment and 1.008 in control, a difference of -0.040 with $p = 0.072$, or roughly a 4% reduction relative to the control mean. Pitch/register, tonal strength, and pitch-class diversity show very small and statistically insignificant differences in cosine distances. Euclidean distances by block generally exhibit similar signs but none are statistically significant at conventional levels.

The audio-based MiRA measures in the bottom panel of Table A3 show a similar pattern. For CoverID (lower is more similar), the mean within-group value is 0.573 in the treatment group and 0.581 in control (difference $= -0.009$, $p = 0.804$). For CLAP similarity (higher is more similar), the means are 0.904 and 0.899 (difference $= 0.006$, $p = 0.462$). For the symmetric KL divergence (lower is more similar), the means are 0.113 and 0.110 (difference $= 0.003$, $p = 0.881$). Across these metrics, differences in within-group similarity are small, mixed in sign, and not statistically significant.

5 Discussion and Conclusion

This paper studies how access to a generative AI text-to-music system during ideation affects both the musical features and perceived creativity of short musical compositions. Using a randomized controlled trial, we find that giving musically trained participants access to Udio does not increase perceived creativity and instead leads to lower creativity as judged by average listeners. In particular, assignment to the AI condition lowers mean creativity scores by about 0.15 points on the 1–5 scale. This reduction is modest in size and marginally significant at conventional levels. This is surprising, as one would expect that users could ignore ideation if they felt it was not helping them. However, we also find that self-reported enjoyment and satisfaction with one's own melody are similar across treatment arms. Clearly, users do not perceive their output as negatively affected.

We also find differences in compositional choices under AI access that are closely linked to these rating differences. Relative to the control group, treated melodies contain fewer notes, have a narrower pitch range, and are rhythmically sparser, with longer note durations and inter-onset intervals. Timing and density measures explain a substantial share of the variation in creativity scores. In our data, melodies that are more rhythmically active and temporally dense tend to be judged as more creative, whereas slower, sparser melodies receive lower creativity ratings. Udio access therefore nudges composers toward musical patterns that listeners in our sample associate with lower creativity.

These results also highlight a gap between self-reported and external evaluations of creativity. As noted above, treated and control participants report similar enjoyment and satisfaction with their melodies, even though external listeners judge AI-influenced melodies as less creative on average. Generative AI may therefore make the creative process feel equally enjoyable and productive without improving, and in our setting slightly worsening, how creative the output appears to others. For individual creators, this points to a trade-off: AI can support ideation and provide a wide range of

options quickly, while shifting stylistic choices in ways that reduce perceived creativity for listeners.

We view these patterns as consistent with at least two plausible mechanisms, though our design does not directly identify them. The first is creative fixation. Psychological research shows that exposure to examples can anchor people's mental representations and narrow the range of ideas they explore, and recent work on human-AI collaboration suggests that generative models can act as particularly strong anchors. In our experiment, Udio provides polished musical examples in response to text prompts. Even though participants must still compose and record their own 15-second melodies, the AI-generated tracks likely serve as salient references for tempo, density, and overall pacing. The musical structure shifts we observe affect precisely the dimensions most strongly correlated with creativity ratings, consistent with AI-guided ideation steering composers toward regions of the musical space that listeners regard as less creative.

A second, listener-side mechanism involves musical arousal. Prior work links faster, denser musical stimuli to higher arousal states, and arousal has been associated with heightened perceptions of novelty, engagement, and problem solving (Williams et al., 2015). In our setting, the musical features that predict higher creativity ratings, such as more notes and faster pacing, are also those likely to produce higher arousal. If Udio's outputs encourage participants to adopt less dense textures or more relaxed pacing, exposure to the model may lower the arousal profile of the melodies that composers ultimately create. Listeners may then interpret these pieces as less creative even if they are competently structured.

Several limitations of our study suggest directions for future work. First, our experimental design abstracts away from an important feature of real-world creativity, namely throughput. In the experiment, each participant produces a single 15-second melody under time constraints. In practice, text-to-music models may allow musicians to generate many ideas quickly and discard most of them. A reduction in average creativity across all AI-influenced ideas might be less concerning if higher throughput substantially increases the likelihood of producing a few exceptionally creative outputs. Our results speak to how AI access affects the creativity of a given melody under controlled conditions, but not to the overall distribution of work that a composer might select after exploring many AI-assisted ideas. Studying this trade-off between average creativity and the probability of rare, highly creative outcomes is an important direction for future work.

Second, the musical task in our experiment is intentionally constrained. Participants compose fixed-length, short, monophonic piano melodies in a laboratory setting. This allows for precise control, but limits generalization to longer, fully produced pieces and to more natural workflows. We also study a single commercial text-to-music system and a particular way of interacting with the tool. Other models, especially symbolic AI tools that operate on scores or piano-roll representations, may have different effects on stylistic choices and perceived creativity. Our composer sample consists of musically trained individuals mostly from a single institution (CMU), and our main evaluator pool comprises average listeners recruited online. A broader and more diverse sample of both composers and listeners would help test our findings across genres, populations, and levels of expertise.

Third, our evaluator sample is aligned with our main research question but also suggests

directions for future work. We focus on perceived creativity among typical listeners, so our primary evaluators are online participants with varied but mostly non-professional musical backgrounds, which is appropriate for our construct of interest, creativity, as experienced by average listeners. It would be interesting to conduct this study with a large number of evaluators with significant composition experience.

Future research can build on our results in several ways. One question is whether there are modes of using text-to-music models that increase, rather than reduce, perceived creativity. For example, designs that explicitly promote diverse generations or highlight atypical outputs (perhaps informed by music theory and genre) may help users avoid fixation on high-probability patterns. Another direction is to study other classes of AI music models, especially symbolic composition tools designed to be more deeply integrated into compositional workflows and to give users greater control over the structure of the output. Models that operate on symbolic representations and allow users to iteratively accept, reject, and modify suggestions in context may yield different outcomes than one-shot text-to-music models. Comparing how text-to-music and symbolic systems affect both the mean and the tails of creative outcomes would help to understand whether and how generative AI can expand creative possibilities in music.

Finally, our findings relate to ongoing policy debates about training generative models on copyright-protected music. Many current text-to-music systems are trained on copyrighted recordings without permission, and fair use arguments often emphasize the potential for such training to generate more innovative and creative works. In our setting, we do not find evidence that access to a text-to-music model raises perceived creativity of the resulting compositions; instead, we find a small but significant decline. Although our evidence comes from one model and one task, it suggests that large creativity gains in music from such training are not guaranteed.

References

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5):997–1013.

Amabile, T. M. (1983). The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology*, 45(2):357–376.

Amabile, T. M. (1996). *Creativity in Context: Update to the Social Psychology of Creativity*. Westview Press, Boulder, CO.

Batlle-Roca, R., Liao, W.-H., Serra, X., Mitsufuji, Y., and Gómez, E. (2024). Towards assessing data replication in music generation with music similarity metrics on raw audio. In *Proceedings of the 25th International Society for Music Information Retrieval Conference*, pages 1004–1011. ISMIR.

Bick, A., Blandin, A., and Deming, D. J. (2024). The rapid adoption of generative ai. Working Paper 32966, National Bureau of Economic Research.

Brynjolfsson, E., Li, D., and Raymond, L. (2025). Generative ai at work*. *The Quarterly Journal of Economics*, 140(2):889–942.

Chen, C., KASOF, J., HIMSEL, A., DMITRIEVA, J., DONG, Q., and XUE, G. (2005). Effects of explicit instruction to "be creative" across domains and cultures. *The Journal of Creative Behavior*, 39(2):89–110.

Chen, Z. and Chan, J. (2024). Large language model in creative work: The role of collaboration modality and user expertise. *Management Science*, 70(12):9101–9117.

Cuthbert, M. and Ariza, C. (2010). Music21: A toolkit for computer-aided musicology and symbolic music data. pages 637–642.

Diehl, M. and Stroebe, W. (1987). Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of Personality and Social Psychology*, 53(3):497–509.

Donahue, C., Wu, S.-L., Kim, Y., Carlton, D., Miyakawa, R., and Thickstun, J. (2025). Hookpad aria: A copilot for songwriters.

Doshi, A. R. and Hauser, O. P. (2024). Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290.

Epstein, Z., Hertzmann, A., of Human Creativity, I., et al. (2023). Art and the science of generative ai. *Science*, 380(6650):1110–1111.

Geringer, J. and Johnson, C. (2007). Effects of excerpt duration, tempo, and performance level on musicians' ratings of wind band performances. *Journal of Research in Music Education*, 55:289–301.

Guilford, J. P. (1967). *The Nature of Human Intelligence*. McGraw-Hill, New York, NY.

Kohn, N. W. and Smith, S. M. (2011). Collaborative fixation: Effects of others' ideas on brainstorming. *Applied Cognitive Psychology*, 25(3):359–371.

Kumar, H., Jordan, E., Jonathan, V., and Anderson, A. (2025). Human creativity in the age of llms randomized experiments on divergent and convergent thinking. *CHI, Yokohama, Japam*.

Loui, P., Parrish, C., Wu, X., Lee, J., and Margulis, E. (2024). Sequence generation and evaluation: A novel assessment of musical creativity. *Creativity Research Journal*. Publisher Copyright: © 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.

Marrone, R., Cropley, D., and Medeiros, K. (2024). How does narrow ai impact human creativity? *Creativity Research Journal*, pages 1–11. Advance online publication; published online 2024-07-15.

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review*, 69(3):220–232.

Mednick, S. A. (1968). Remote associates test. *Journal of Creative Behavior*, 2(3):213–214.

Nijstad, B. A., Stroebe, W., and Lodewijkx, H. F. (2003). Production blocking and idea generation: Does blocking interfere with cognitive processes? *Journal of Experimental Social Psychology*, 39(6):531–548.

Oppenheimer, D. M. and Patterson, M. T. (2025). Thinking outside the box means thinking outside the search engine. *Memory & Cognition*.

Porter, B. and Machery, E. (2024). Ai-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports 14*, 26133.

Priest, T. (2001). Using creativity assessment experience to nurture and predict compositional creativity. *Journal of Research in Music Education*, 49(3):245–257.

Said-Metwaly, S., Noortgate, W. V. d., and Kyndt, E. (2017). Approaches to measuring creativity: A systematic literature review. *Creativity. Theories – Research - Applications*, 4(2):238–275.

Smith, S. M. (1995). Fixation, incubation, and insight in memory and creative thinking. In Smith, S. M., Ward, T. B., and Finke, R. A., editors, *The Creative Cognition Approach*, pages 135–156. MIT Press, Cambridge, MA.

Smith, S. M., Ward, T. B., and Schumacher, J. S. (1993). Constraining effects of examples in a creative generation task. *Memory & Cognition*, 21(6):837–845.

Sternberg, R. J. and Lubart, T. I. (1995). *Defying the Crowd: Cultivating Creativity in a Culture of Conformity*. Free Press, New York, NY.

Taylor, D. W., Berry, P. C., and Block, C. H. (1958). Does group participation when using brainstorming facilitate or inhibit creative thinking? *Administrative Science Quarterly*, 3(1):23–47.

Torrance, E. P. (1966). *The Torrance Tests of Creative Thinking: Norms-technical manual*. Personal Press, Lexington, MA.

Torrance, E. P. (2008). *The Torrance Tests of Creative Thinking: Norms-technical manual-figural (streamlined) forms A and B*. Scholastic Testing Service, Bensenville, IL.

Wapnick, J., Campbell, L., Siddell-Strebel, J., and Darrow, A.-A. (2005). Effects of excerpt tempo and duration on musicians' ratings of high-level piano performances. *Journal of Research in Music Education*, 53(2):162–176.

Williams, D., Kirke, A., Miranda, E., Daly, I., Hallowell, J., Weaver, J., Malik, A., Roesch, E., Hwang, F., and Nasuto, S. (2015). Investigating perceived emotional correlates of rhythmic density in algorithmic music composition. *ACM Trans. Appl. Percept.*, 12(3).

Xusen, C. and Zhang, L. (2025). Inspiration booster or creative fixation? the dual mechanisms of llms in shaping individual creativity in tasks of different complexity. *Humanities and Social Sciences Communicationscientific Reports*, 12-1563.

Zhou, E. and Lee, D. (2024). Generative artificial intelligence, human creativity, and art. *PNAS Nexus*, 3(3):pgae052.

Appendix

Pilot Survey: Effect on Creativity, Musicality and Enjoyment Ratings

We first designed and executed a pilot survey to test our survey design. In this phase, we recruited 176 evaluators using Prolific and collected 3,143 ratings. Each melody received at least 21 evaluations, with a median of 22 and a maximum of 24. After listening, participants answered three questions on 5-point Likert scales: (1) "How creative do you find this melody?", (2) "How musical do you find this melody?", and (3) "How much do you like this melody?". Responses ranged from 1 ("Not at all") to 5 ("Extremely"). Creativity is our primary outcome, and musicality and enjoyment are secondary outcomes that capture broader listener judgments of the piece.

We find that treatment decreased the ratings of participant melodies by 0.139 points on the 1–5 scale (SE = 0.077, $p = 0.075$), which is statistically significant at the 10% level. This corresponds to about 0.3 standard deviations of the cross-melody distribution of mean creativity scores, and the implied 95% confidence interval is approximately $[-0.29, 0.01]$.

The effects on musicality and enjoyment are similar in magnitude. For musicality, the treatment effect is -0.138 (SE = 0.079, $p = 0.082$), and for enjoyment it is -0.164 (SE = 0.086, $p = 0.057$). Both are statistically significant at the 10% level. The three measures are also highly correlated (creativity–enjoyment: Pearson r = 0.91; creativity–musicality: r = 0.91; enjoyment–musicality: r = 0.94; all $p < 0.001$). This pattern suggests that evaluators may not sharply distinguish among the constructs and/or that responses on one dimension may influence the others when elicited jointly. Given this redundancy, we asked only a single creativity question in the main survey (N = 500).

Variable	Control (N=4316)	Treatment (N=4684)	All (N=9000)	$\lvert SMD \rvert$	q (BH)
Survey duration (min)	9.53 (5.33)	9.44 (5.05)	9.48 (5.19)	0.017	0.795
Total Prolific approvals	4299.71 (3823.87)	4284.33 (3726.16)	4291.71 (3773.14)	0.004	0.887
Age	42.49 (12.52)	42.42 (12.73)	42.45 (12.63)	0.006	0.860
Has musical training	0.54 (0.50)	0.54 (0.50)	0.54 (0.50)	0.011	0.798
Main instrument: Piano	0.20 (0.40)	0.20 (0.40)	0.20 (0.40)	0.005	0.871
Main instrument: Woodwind	0.07 (0.26)	0.07 (0.26)	0.07 (0.26)	0.009	0.798
Main instrument: Guitar	0.11 (0.32)	0.11 (0.31)	0.11 (0.31)	0.022	0.795
Main instrument: Other	0.16 (0.36)	0.16 (0.36)	0.16 (0.36)	0.004	0.891
Formal musical training: 1-3 years	0.19 (0.39)	0.20 (0.40)	0.19 (0.40)	0.035	0.245
Formal musical training: 4-7 years	0.16 (0.37)	0.15 (0.36)	0.15 (0.36)	0.024	0.684
Formal musical training: 8+ years	0.20 (0.40)	0.18 (0.39)	0.19 (0.39)	0.028	0.684
Piano experience: 1-3 years	0.15 (0.36)	0.17 (0.38)	0.16 (0.37)	0.049	0.127
Piano experience: 4-7 years	0.09 (0.28)	0.07 (0.26)	0.08 (0.27)	0.041	0.055
Piano experience: 8+ years	0.10 (0.30)	0.09 (0.29)	0.10 (0.29)	0.024	0.684
Composes music	0.22 (0.41)	0.22 (0.41)	0.22 (0.41)	0.001	0.945
Composes $\leq$ once/month	0.14 (0.35)	0.15 (0.36)	0.15 (0.35)	0.017	0.795
Composes $\geq$ few times/month	0.06 (0.24)	0.06 (0.23)	0.06 (0.23)	0.024	0.571
Sex: Female	0.49 (0.50)	0.50 (0.50)	0.49 (0.50)	0.007	0.850
Ethnicity: White	0.71 (0.45)	0.70 (0.46)	0.71 (0.46)	0.012	0.795
Country of birth: USA	0.92 (0.26)	0.92 (0.27)	0.92 (0.27)	0.019	0.684
Age 18-24	0.05 (0.21)	0.05 (0.22)	0.05 (0.21)	0.011	0.795
Age 25-34	0.25 (0.44)	0.26 (0.44)	0.26 (0.44)	0.010	0.795
Age 35-44	0.31 (0.46)	0.30 (0.46)	0.31 (0.46)	0.016	0.795
Age 45-54	0.21 (0.41)	0.20 (0.40)	0.20 (0.40)	0.014	0.795
Age 55+	0.18 (0.39)	0.19 (0.39)	0.18 (0.39)	0.013	0.795
College degree	0.61 (0.49)	0.62 (0.49)	0.61 (0.49)	0.006	0.871
Employment: Full-time	0.51 (0.50)	0.48 (0.50)	0.50 (0.50)	0.048	0.055
Employment: Part-time	0.16 (0.37)	0.17 (0.38)	0.17 (0.37)	0.017	0.795
Employment: Not paid	0.10 (0.30)	0.09 (0.29)	0.10 (0.29)	0.018	0.795
Employment: Unemployed	0.09 (0.28)	0.10 (0.30)	0.09 (0.29)	0.048	0.024

Table A1: Evaluator characteristics across ratings of control and treatment melodies

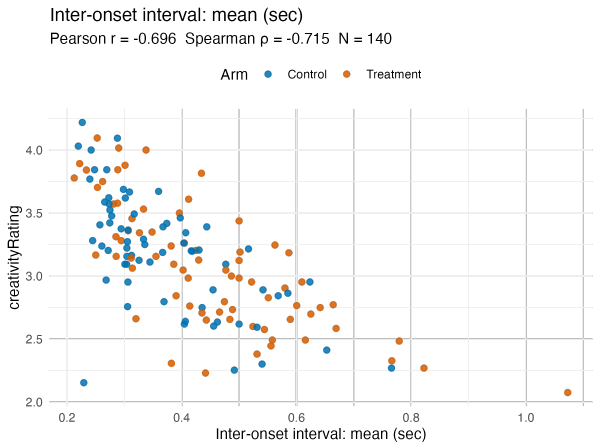
Table A2: ITT with Treatment×Moderator Interactions

	ITT with interactions
Intercept	2.698***
	(0.218)
Treatment	−0.159
	(0.131)
CMU School of Music major	0.342**
	(0.146)
Piano is main instrument	−0.032
	(0.143)
Piano ≥8 yrs	0.149
	(0.107)
Main Instrument ≥8 yrs	0.017
	(0.095)
High non-Western music exposure	0.065
	(0.085)
Used any AI for music (past)	−0.089
	(0.146)
Self-reported Composition skill ≥4	0.093
	(0.135)
Age 18–24	0.346*
	(0.193)
Age 25–34	0.354*
	(0.189)
Female	−0.080
	(0.085)
Full-time worker	0.449**
	(0.205)
Student	0.085
	(0.167)
Task start period (Morning=1/Afternoon=0)	0.022
	(0.096)
Treatment × Self-reported Composition skill ≥4	0.127
	(0.185)
Treatment × Used any AI for music (past)	0.124
	(0.204)
Treatment × Piano is main instrument	0.047
	(0.169)
Treatment × CMU School of Music major	−0.415**
	(0.190)
Num.Obs.	140
R2	0.257
R2 Adj.	0.118

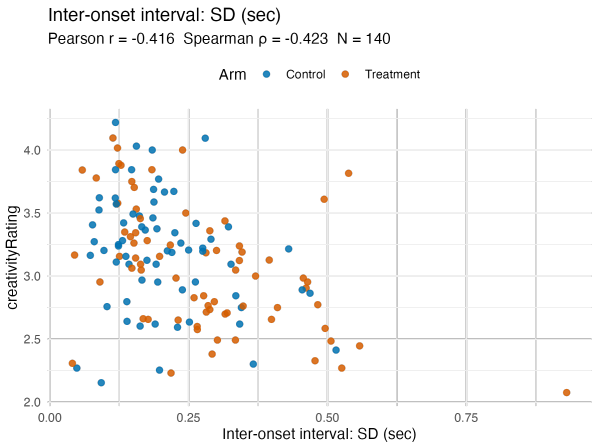
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A3: Within-group similarity scores and differences. For CoverID and KL divergence, lower values indicate higher similarity. For CLAP, higher values indicate higher similarity.

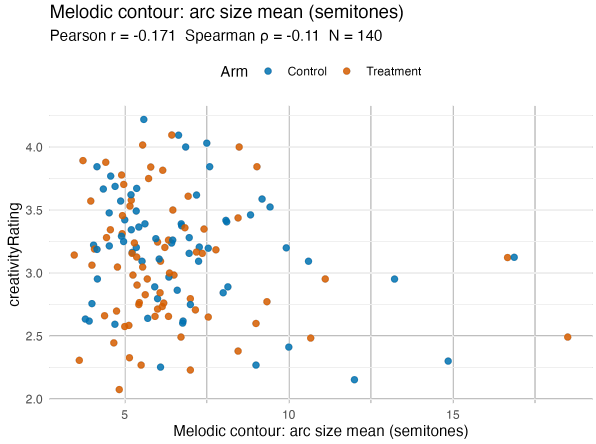
Method	Feature Group	Within Treatment	Within Control	Delta	p-value
Cosine distance	All	0.993	0.972	0.021	0.036
Cosine distance	Rhythm	0.968	0.861	0.107	0.002
Cosine distance	Density	0.986	1.006	-0.021	0.067
Cosine distance	Pitch/Register	0.991	1.001	-0.010	0.209
Cosine distance	Intervals/Contour	0.968	1.008	-0.040	0.072
Cosine distance	Tonal Strength/Pitch Class diversity	1.003	1.005	-0.001	0.956
Euclidean distance	All	5.271	4.983	0.288	0.446
Euclidean distance	Rhythm	2.119	1.859	0.260	0.278
Euclidean distance	Density	1.736	1.653	0.083	0.689
Euclidean distance	Pitch/Register	1.799	1.781	0.018	0.873
Euclidean distance	Intervals/Contour	2.667	2.934	-0.267	0.421
Euclidean distance	Tonal Strength/Pitch Class diversity	1.612	1.692	-0.081	0.769
CoverID	Audio File	0.573	0.581	-0.009	0.804
CLAP	Audio File	0.904	0.899	0.006	0.462
KL divergence	Audio File	0.113	0.110	0.003	0.881



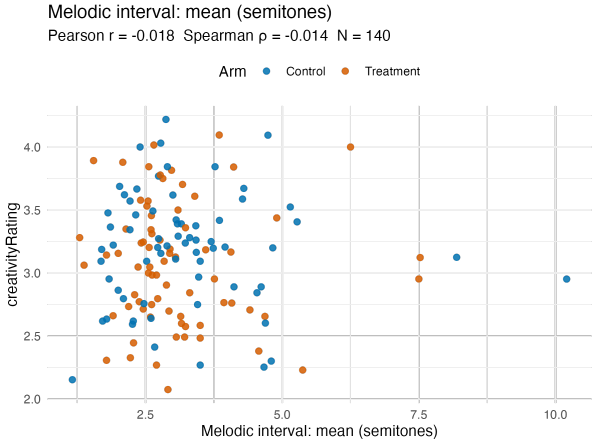
(a) Inter-onset interval: mean (s)



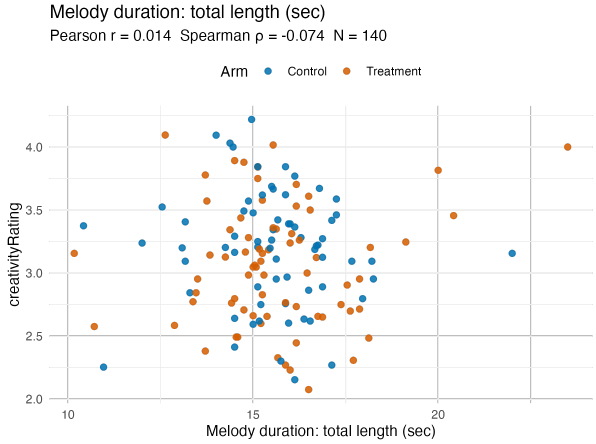
(b) Inter-onset interval: SD (s)



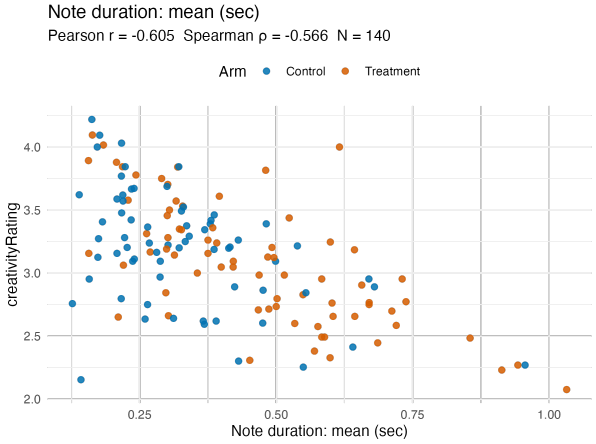
(c) Melodic contour: arc size mean (semitones)



(d) Melodic interval: mean (semitones)

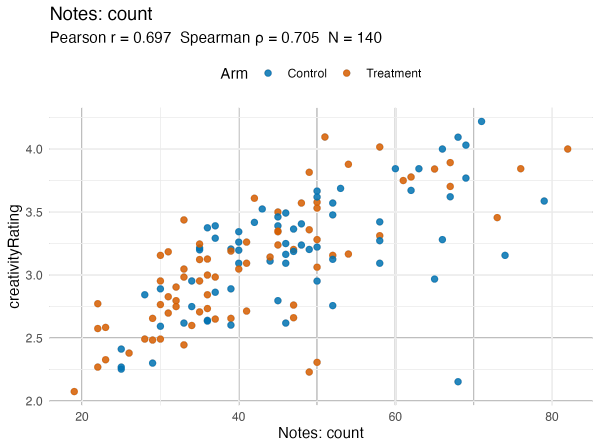


(e) Melody duration: total length (s)

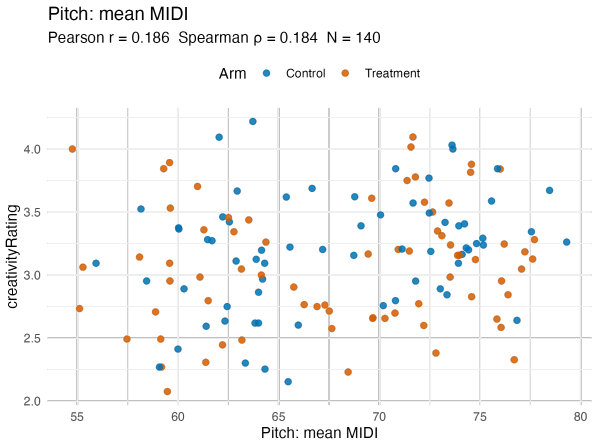


(f) Note duration: mean (s)

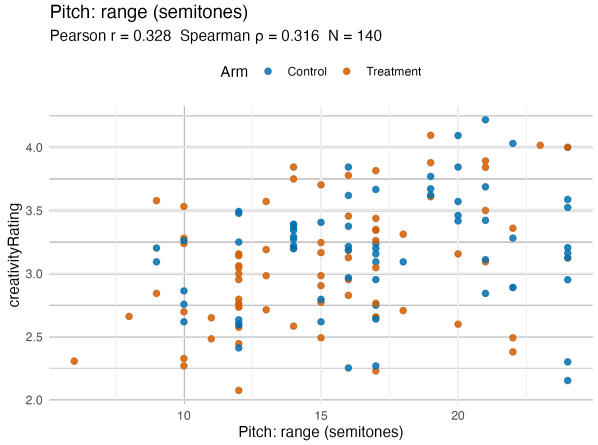
Figure A1: Creativity vs. feature blocks.



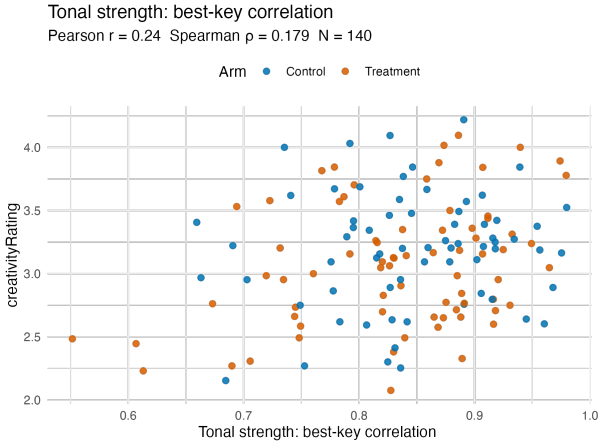
(g) Notes: count



(h) Pitch: mean (MIDI)



(i) Pitch range (semitones)



(j) Tonal strength: best key correlation

Figure A1: Creativity vs. feature blocks (continued).

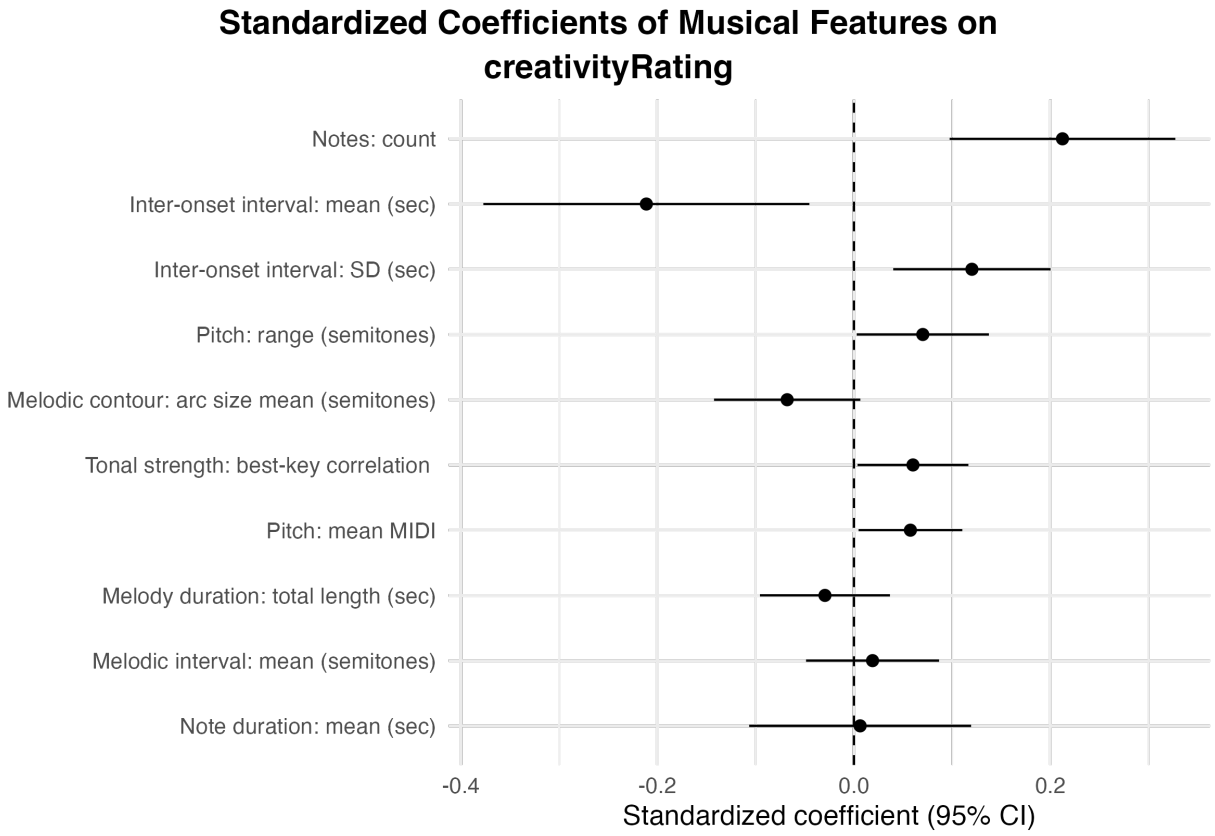


Figure A2: Standardized coefficients (HC3, 95% CI) of musical features predicting creativity ratings.