Metaphors for Communicating Musical Expression: Comparing LLMs to Humans Interpreting Metaphors in Piano Performance

Jonathan C. Hosea (s3712826)

Supervisors:
Dr. Tessa Verhoef & Dr. Rebecca Schaefer

Bachelor Thesis– Informatica

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

August 12, 2026

Abstract

In music education and performance, metaphors are frequently used to communicate complex emotional and expressive concepts that basic technical instructions cannot easily capture. As music generation AI is in full development, effectively communicating these nuanced, concepts to generative models remains a challenge. This thesis investigates the capability of LLMs (Gemini 3.1 Pro, GPT-5.5 and Claude Sonnet 4.6) to interpret metaphorical instructions in musical performance and compares their outputs to human pianists. Using a replicated experimental framework, the models were prompted to generate piano performances of two neutral melodies based on eight metaphors varying in valence, arousal, and type. Analysis of the simulated motor performance data (keypress velocity, tempo, note length, and their respective variabilities) reveals that while LLMs successfully capture the directional expressivity of humans, primarily driven by arousal, they significantly exaggerate contrasts. Furthermore, the models frequently defaulting to robotic, zero variance timing. These findings indicate that while current LLMs lack the organic nuance required for autonomous humanlike piano performance, they hold strong potential as interactive, metaphor-driven toolboxes for music education and generative experimentation, bridging abstract emotional intent with concrete performance control without replacing human artistry.

Contents

1 Introduction

In the last century, music has gotten more and more accessible to most people, imbuing itself into everyday life [HM01]. People are nowadays listening to music during all kinds of activities such as travel, physical exercise and studying [LGS01]. Martin Clayton describes music as 'mediation between self and other', since music is often used for interaction where regular verbal communication is insufficient [Cla01]. Besides, Clayton explains that musical performance also helps regulating an individual's cognitive physiological, or emotional state . This relation between music and emotion is a topic that has been a blooming field of research in the last few decades [JS01].

Moreover, the role of emotions within listening to music has been widely explored [Slo10, Jus01]. Similarly, emotion in music performance has been a prevalent research direction. For example, Gabrielsson and Juslin conducted an experiment in 1996, showing the role of music performers' expressive intentions in contributing to the listener's experience [GJ96]. Juslin, together with others, also extensively dove into how different emotions are usually conveyed by music artists and how accurate they are able to communicate the emotions to listeners [JL01, JT10].

The artists' ability to convey emotion towards listeners is a skill that is taught from the beginning of music education, and many music teachers believe that teaching musical expression to new students starts with physical freedom, connection to the instrument, and repertoire choice [BS13]. Additionally, other fundamental teaching methods have been tested, for instance in a study by Woody [Woo06], where three different approaches were used: aural modeling, verbal instruction addressing concrete musical properties, and verbal instruction using imagery and metaphor.

Metaphors are a common but not canonical way to verbally communicate desired musical expression between musicians [Kra87]. Despite the fact that metaphors can be interpreted differently, they still offer an intuitive way for artists to embody the emotions in their musical performances and is therefore widely used in music education [MT20, Bro25].

Within the last few years, chatbots built on large language models (LLMs) have surged in popularity and changed the human relation to computers drastically [KP24]. LLMs have become a greatly researched topic in all research areas, including linguistics, where for instance metaphor understanding by LLMs is constantly being evaluated and improved [TCLS24, KND25].

A recent study by Schaefer and Verhoef explored metaphors in music as a means of expressive communication. They conducted an experiment, collecting motor performance data from trained pianists playing predefined melodies influenced by different kinds of metaphors [SV24]. As a possible direction for future work, the authors propose connecting the explored metaphors to music making AI.

Recently, there have been many advancements in modeling emotions for music generation [SHYX25], but an unexplored field of generative music AI is communicating emotions to LLMs using metaphors to generate expressive music. For instance, specialized text-based music generation models exist such as the popular Suno and Udio [CVD+25], and the open source YuE [YLG+25] and NotaGen [WWH+25]. However, we found that these music generation models had trouble

using any text input other than lyrics and genre specifications, which limits the user's ability to input desired musical expression into prompts.

This thesis aims to evaluate and compare current state-of-the-art LLMs with humans in understanding metaphors in piano performance by using the human motor performance data from the previously mentioned study by Schaefer and Verhoef [SV24], answering the following research question:

Compared to human pianists, how do current state-of-the-art LLMs interpret metaphors in piano performance?

Answering this question allows us to assess the use of metaphors as a means to communicate musical expression towards generative music AI, potentially opening a new and effective way to model emotion into music generation algorithms.

This thesis is structured as follows: Section 2 provides necessary background and discusses related work; Section 3 describes the methodology used to answer the research question; Section 4 presents the results of the experiment; Section 5 discusses the results, places them in context and presents possible future research. Finally, Section 6 concludes.

2 Background

2.1 Metaphors in performance art

As introduced in Section 1, metaphors are an intuitive and widely used tool in music education. Metaphors act as cross-domain mappings that allow musicians to understand an abstract domain in terms of another, more tangible domain [SGRG19]. Rather than providing a strict instruction, a metaphor actively modifies the performer's internal musical representation. This allows musicians to translate complex cognitive concepts and emotional states into expressive physical motor performance, connecting the discrepancy between internal feeling and mechanical execution.

2.2 LLMs' understanding of metaphors

Because metaphors reflect fundamental cognitive processes such as analogical reasoning and categorization, metaphor comprehension has become an essential evaluation metric for LLMs [TCLS24]. While LLMs are specialized for literal text processing, interpreting metaphorical language remains a significant challenge due to its linguistic complexity, variability, and cultural embeddedness [KND25]. Datasets show that LLMs sometimes struggle with full metaphor interpretation, occasionally resorting to lexical similarities rather than demonstrating true understanding [TCLS24]. However, because state-of-the-art models have been trained on data including music education literature where metaphors are widely used, they theoretically should be able to interpret musical metaphors. This study shifts the focus from general linguistic comprehension to how these models interpret metaphors in a specialized, musical context.

2.3 Bridging LLMs and Music Generation

The extent to which AI models can comprehend and reason about music is a growing field of research. While LLMs exhibit understanding of basic musical rules and can perform successfully on simple symbolic tasks like defining note intervals, they often struggle with more complex logical reasoning in music, suggesting their capabilities rely more on memorizing general patterns than on a robust understanding of musical structures [ZWW+24].

Simultaneously, specialized generative music models (such as Suno, Udio, and YuE) have made massive progress in music generation, but are limited in their semantic control. As noted earlier, these systems rely on explicit genre tags or lyrical inputs rather than expressive textual instructions. We intend to build upon this research by exploring whether generalized LLMs can serve as a semantic bridge. By evaluating how LLMs translate metaphors into concrete motor performance data, we can assess whether they are adequately capable to convert abstract human intent into the strict numerical parameters that generative music systems require.

2.4 Pianists experiment

To measure this translation, this thesis builds directly upon the experimental framework established by Schaefer and Verhoef [SV24]. In their original study, six human pianists performed two neutral melodies while emulating musical expression through metaphorical prompts. These metaphors were systematically categorized across three distinct semantic dimensions:

- **Valence:** The emotional charge of the metaphor (*positive* or *negative*).
- **Arousal:** The level of physical or emotional activation (*high* or *low*) implied by the prompt.
- **Type:** Whether the metaphor was worded as an active, physical *action* or an internal *emotion*.

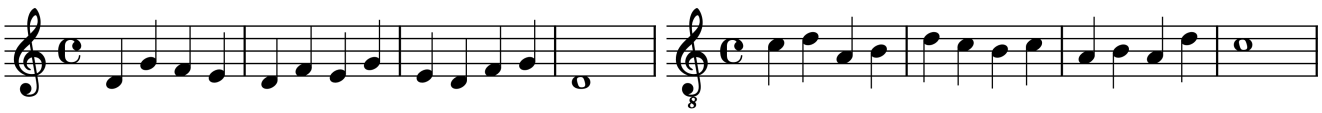
By applying this exact experimental framework to state-of-the-art LLMs and analyzing the resulting simulated motor data (keypress force, tempo, and tempo variability), we can analyze exactly how AI models interpret expressive prompts compared to human musicians.

3 Methods

3.1 Melodies and metaphors

To ensure a rigorous and fair comparison between the performance of LLMs and human musicians, we replicated the experimental design established by Schaefer and Verhoef [SV24] as closely as possible. The LLMs were instructed to apply the exact same set of metaphors to the identical melodies used in the original human pianist trials.

In their study, Schaefer and Verhoef designed two neutral melodies, shown in Figure 1. These note sequences were intentionally composed to be emotionally ambiguous, ensuring that any expressive variations in motor performance such as dynamic changes or tempo shifts would be driven entirely by the metaphorical instruction rather than the musical structure of the melody.



(a) Melody 1 (b) Melody 2

Figure 1: The two melodies used in the experiment.

The metaphors were systematically selected based on three semantic dimensions: valence (positive or negative), arousal (high or low), and type (action or emotion). This resulted in eight distinct metaphorical instructions, shown in Table 1. This structured selection ensures that the metaphors cover a wide range of physical and psychological states, allowing us to evaluate how LLMs parameterize specific semantic dimensions into motor performance output.

Arousal	Valence	Type: Emotion	Type: Action
High	Positive	In a jubilant manner	Like you are leaping
High	Negative	In a violent manner	Like you are having an argument
Low	Positive	In a sweet manner	Like you are walking at sunset
Low	Negative	In a bored manner	Like you are passively disengaging

Table 1: The eight metaphors used in the experiment, which are either high or low in arousal, positive or negative in valence, and use emotion or action words.

3.2 Model selection and prompt engineering

We selected three current state-of-the-art Large Language Models for this experiment: Google's Gemini 3.1 Pro, OpenAI's GPT-5.5, and Anthropic's Claude Sonnet 4.6. These models were chosen due to their advanced reasoning capabilities, widespread accessibility and popularity.
Initially, we investigated specialized text-based music generation models, including the popular closed-source platforms Suno and Udio [CVD+25], as well as open-source alternatives like YuE [YLG+25] and NotaGen [WWH+25]. However, we found that these specialized models struggle to process text inputs other than lyrics and explicit genre specifications. Because they are not designed to process metaphorical instructions, a direct and fair comparison to the human motor data is impossible. Therefore, we utilized general-purpose LLMs, testing their ability to bridge semantic comprehension and music performance generation.

To simulate the experiment, we relied on structured prompt engineering. Our initial prompt attempt (see Listing 1) was not robust enough, as the models often eliminated any variance in tempo or note length (GPT-5.5 had zero variance for all trials), or modified variable names.
To resolve these issues, we designed a more constrained, updated prompt (see Listing 2). This prompt explicitly defined exact parameter constraints (e.g. standardizing MIDI integers, defining velocity between 0 and 127, and enforcing clear timestamps) while actively giving the LLMs the freedom to systematically and dynamically vary the timing and attack parameters within trials based on the metaphors.

```
You are a trained pianist.
participant={PP1}
melody1={D4, G4, F4, E4, D4, F4, E4, G4, E4, D4, F4, G4, D4};
melody2={C4, D4, A3, B3, D4, C4, B3, C4, A3, B3, A3, D4, C4};
metaphors={"like you are leaping", "like you are walking at sunset",
           "like you are having an argument",
           "like you are passively disengaging",
           "in a jubilant manner", "in a sweet manner",
           "in a violent manner", "in a bored manner"}
Trials are the combinations of melodies and metaphors.
On the piano, play every trial. Based on the metaphor,
vary with note length, time between notes, and attack of the notes.
Do not add additional notes to the melodies.
List every note of the trials in a csv file
as if the piano had MIDI output.
Use column_names={"participant", "trial", "midi_pitch", "midi_velocity",
                  "timestamp", "notelength", "melody", "instruction"}
Give the timestamps and note lengths in seconds, rounded to 3 decimals.
```

Listing 1: Initial prompt.

```
You are an AI simulating a highly trained pianist.

participant="PP1"
melody1=["D4", "G4", "F4", "E4", "D4", "F4", "E4", "G4", "E4", "D4",
         "F4", "G4", "D4"]
melody2=["C4", "D4", "A3", "B3", "D4", "C4", "B3", "C4", "A3", "B3",
         "A3", "D4", "C4"]
metaphors=["like you are leaping", "like you are walking at sunset",
           "like you are having an argument",
           "like you are passively disengaging",
           "in a jubilant manner", "in a sweet manner",
           "in a violent manner", "in a bored manner"]
parameters=["note length", "time between notes", "attack velocity"]

Trials are all possible combinations of melodies and metaphors.

Your task is to simulate playing every trial. Based on the specific
metaphor, you must systematically vary the ‘parameters‘ to reflect the
emotion/instruction. You may also vary parameters dynamically within
a single trial to be expressive.

Constraints:
1. Do not add or remove any notes from the melodies.
2. Convert all pitches to their standard MIDI integer values (C4 = 60).
3. ‘midi_velocity‘ must be an integer between 0 and 127.
4. Timestamps and note lengths must be in seconds, rounded to 3 decimals.
5. The ‘timestamp‘ should represent the start time of the note.
   Reset the timestamp to 0.000 at the beginning of each trial.
6. The ‘instruction‘ column should contain the metaphor used.
7. The ‘trial‘ column should be numbered 1 through 16.
8. The ‘melody‘ column should either contain "melody1" or "melody2".

Output the results as a raw CSV file. Use exactly these column names:
participant, trial, melody, instruction, midi_pitch, midi_velocity,
timestamp, notelength

CRITICAL: You must generate the complete CSV with all rows for all trials.
Do not truncate the data, summarize, or use ellipses (...).
Output only the raw CSV text inside a single code block.
```

Listing 2: Updated prompt.

To generate a dataset robust enough for statistical modeling, we prompted each of the three LLMs to generate eight unique 'participants'. For every new participant, the LLM was launched in a completely fresh, zero-context session. This ensured that previous outputs did not influence subsequent participants, maintaining the independence of the trials.

Once the raw CSV files were generated, the data was compiled and processed. We calculated the Inter-Onset Interval (IOI) for every note by measuring the difference in timestamps between consecutive notes. Besides, the data from all three LLMs and the original human dataset were appended into a single master CSV file. An 'Agent' column was added to categorize the source of each performance (Human, Gemini, GPT, or Sonnet), allowing the statistical model to calculate interaction terms that directly compare each LLM's behavioral deviations against the human baseline.

3.3 Analysis

Statistical analysis was conducted using the R programming environment [R C21]. We used Bayesian multilevel regression models utilizing the brms package [Bü17], serving as an interface for cmdstanr [GB25]. Weakly informative (non-informative) priors were used across all models to ensure the data strictly drove the posterior distributions. Posterior distributions were estimated using Markov Chain Monte Carlo (MCMC) sampling, running 4 chains of 4000 iterations each, including 2000 warmup iterations per chain.

Five distinct motor performance metrics were evaluated: Velocity (keypress force), IOI (tempo), IOI CV (tempo variability), Note length, and Note length CV (articulation variability). Due to the distribution shapes of the data, the Velocity model utilized a skew_normal family, while the timing and variance metrics (IOI, CV, Note length, and Note length CV) were modeled using a lognormal family to appropriately handle strictly positive, right-skewed data.
To analyze the models' internal logic and baseline parameterization, we formulated the regression equations as $x \sim \text{valence} * \text{arousal} * (\text{type} + \text{melody})$.
However, to benchmark the LLMs against human performance, the formula was expanded to $x \sim \text{Agent} * \text{valence} * \text{arousal} * (\text{type} + \text{melody})$. This allowed the models to calculate precise statistical deviations between human expression and AI simulation across all semantic dimensions.

4 Results

This section is structured into five subsections, each representing one of the measures. Every subsection presents the results for each of the tested LLMs. This includes how every model reacts to valence, arousal, type and melody. This is described by (compounding) effects relative to the baseline, which is the negative high arousal action metaphor. The first three subsections also describe how much the model's behavior differs from the human pianists in Schaefer and Verhoef's experiment.

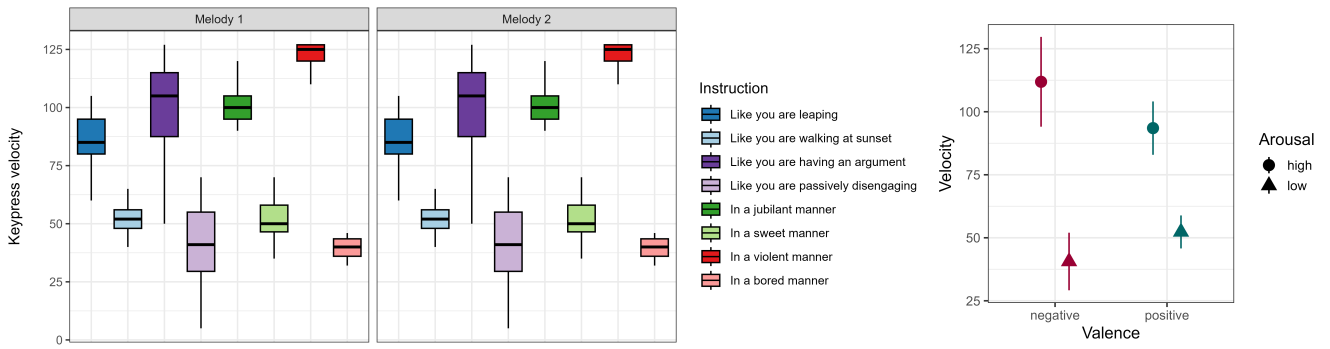
Besides, to show that the results are replicable with similar prompts and are not overfitted to the updated prompt from Section 3.2, the figures belonging to the initial prompt from Section 3.2 are attached in Appendix A for comparison to the corresponding figures in this section.

4.1 Velocity (Force)

4.1.1 Gemini 3.1 Pro

When operating under baseline parameters (negative action metaphors), Gemini played low arousal metaphors with much less force compared to high arousal metaphors ($b = -60.53$, Bayesian 95 % Credible Interval $[-64.53, -56.52]$). Positive metaphors also caused Gemini to play with lower velocity on average, although the effect was weaker ($b = -19.90$, 95 % CrI $[-22.35, -17.44]$). For emotion-type metaphors, Gemini played slightly louder ($b = 13.31$, 95 % CrI $[9.54, 17.00]$), whereas melody had negligible impact ($b = 0.00$, 95 % CrI $[-1.86, 1.87]$). We also found a significant relation between valence and arousal, where the difference in velocity between high and low arousal metaphors was much smaller when the metaphor was also positive ($b = 23.57$, 95 % CrI $[20.34, 26.84]$). This behavior was also stronger for emotion metaphors compared to action metaphors, reducing the velocity difference between high and low arousal even more for positive emotion metaphors ($b = 12.09$, 95 % CrI $[8.21, 15.99]$). Lastly, when playing low arousal metaphors, Gemini played significantly softer for emotion metaphors compared to action metaphors ($b = -22.91$, 95 % CrI $[-25.53, -20.38]$). These results are visualized in Figure 2.

Generally, Gemini showed similar behavior compared to the pianists, when prompted with the trials. Both Gemini and the pianists systematically used less force with low arousal metaphors than with high arousal metaphors. However, Gemini tended to exaggerate this difference, reducing velocity far more than the pianists did ($b = -38.86$, 95 % CrI $[-45.63, -31.97]$). Furthermore, valence and metaphor type had no significant impact on velocity for the pianists, whereas Gemini played softer for positive metaphors compared to negative metaphors, and slightly louder for emotion metaphors compared to action metaphors, as described earlier. These two effects can be attributed to Gemini's interpretation of the negative emotion metaphor "in a violent manner", which Gemini played significantly more forceful than the other metaphors, unlike the pianists. Besides this, the effect of a smaller difference in velocity between high and low arousal for positive metaphors found in the human data is amplified by Gemini, as it played positive, low arousal metaphors louder than the pianists ($b = 17.05$, 95 % CrI $[12.70, 21.51]$). In addition to this, Gemini had lower velocity compared to humans when playing low arousal emotion metaphors ($b = -31.74$, 95 % CrI $[-35.36, -28.12]$), but a decreased effect for low arousal emotion metaphors that were also positive ($b = 23.95$, 95 % CrI $[18.41, 29.53]$).



(a) Velocity for each metaphor

(b) Velocity by valence/arousal

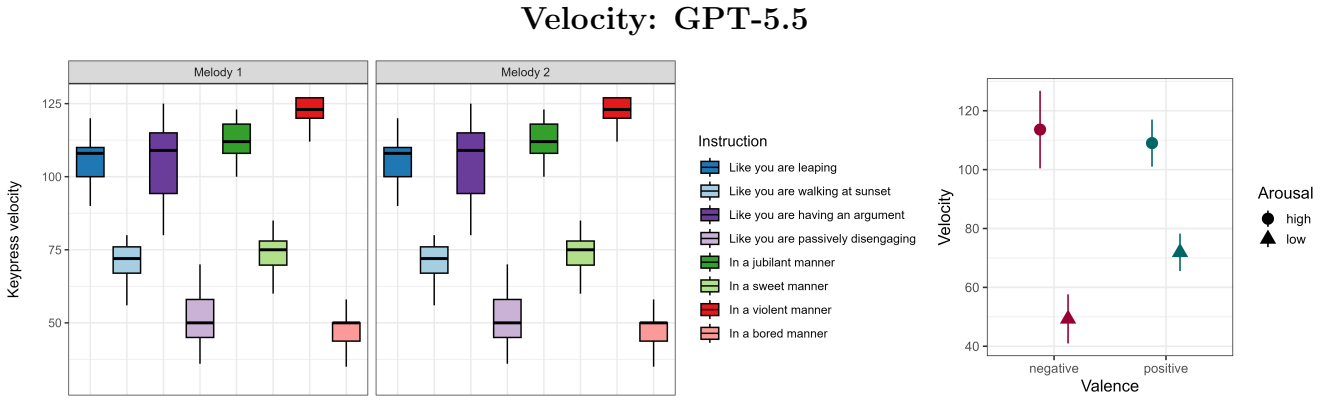
Figure 2: The keypress velocity across eight Gemini 3.1 Pro simulations for the two melodies and eight metaphors. In (a) action metaphors are shown in blue and purple, whereas emotion metaphors are shown in green and red. Positive metaphors are displayed in blue and green, negative metaphors in purple and red. High arousal metaphors are visualized with stronger color intensity, while low arousal metaphors are more transparent. Regardless of metaphor type and melody, Gemini played low arousal metaphors with less force, with a greater velocity drop in the negative valence category, as shown in (a) and (b).

4.1.2 GPT-5.5

Similar to Gemini, GPT-5.5 created a high contrast between high and low arousal metaphors, playing low arousal metaphors approximately twice as soft as high arousal metaphors ($b = -56.13$, Bayesian 95 % Credible Interval $[-60.14, -52.03]$). Valence ($b = -0.78$, 95 % CrI $[-5.71, 4.12]$) and melody ($b = 0.35$, 95 % CrI $[-2.09, 2.84]$) had no significant impact on velocity. Additionally, we found several significant interactions between valence, arousal and type. When looking at positive metaphors in comparison with negative metaphors, both the difference in velocity between high and low arousal metaphors ($b = 19.55$, 95 % CrI $[17.58, 21.52]$) and emotion versus action metaphors ($b = -5.26$, 95 % CrI $[-7.18, -3.37]$) was smaller. Furthermore, when looking at high arousal negative metaphors, emotion metaphors were played louder than action metaphors ($b = 10.77$, 95 % CrI $[8.37, 13.26]$). When arousal was low, a significant interaction ($b = -20.36$, 95 % CrI $[-22.18, -18.57]$) inverted this effect, resulting in emotion metaphors being played softer than action metaphors. However, this interaction shift was significantly mitigated when GPT-5.5 played positive metaphors ($b = 16.20$, 95 % CrI $[13.57, 18.79]$). These results are visualized in Figure 3.

For the most part, GPT-5.5 managed to interpret the metaphors similarly to the pianists. While GPT-5.5 correctly emulated the main tendencies, it often exaggerated the effect. The drop in velocity between high and low arousal metaphors was more than doubled for GPT-5.5 compared to the pianists ($b = -34.72$, 95 % CrI $[-41.44, -27.98]$). Moreover, for positive metaphors, GPT-5.5 exaggerated the effect of smaller differences between high and low arousal ($b = 13.85$, 95 % CrI $[9.66, 18.18]$) and emotion versus action metaphors ($b = -3.61$, 95 % CrI $[-7.10, -0.14]$). On the other hand, GPT-5.5's complex three-way interaction where emotion metaphors were played louder for both high arousal negative and low arousal positive metaphors, but not for low arousal negative metaphors, was completely opposite of the pianists' behavior. GPT-5.5's estimate for the

interaction between arousal and type, deviated severely from the human baseline ($b = -28.57$, 95 % CrI $[-32.13, -25.01]$). Similarly, its three-way interaction between arousal, type and valence significantly overshot the estimate found in the human data ($b = 27.58$, 95 % CrI $[22.31, 32.82]$).



(a) Velocity for each metaphor

(b) Velocity by valence/arousal

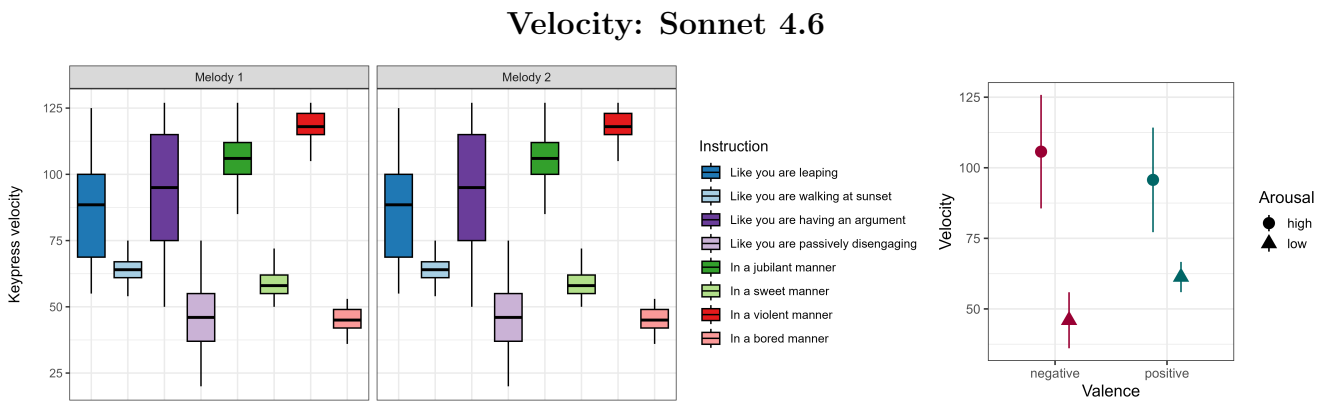
Figure 3: The keypress velocity across eight GPT-5.5 simulations for the two melodies and eight metaphors. In (a) action metaphors are shown in blue and purple, whereas emotion metaphors are shown in green and red. Positive metaphors are displayed in blue and green, negative metaphors in purple and red. High arousal metaphors are visualized with stronger color intensity, while low arousal metaphors are more transparent. Regardless of metaphor type and melody, GPT-5.5 played low arousal metaphors with less force, with a greater velocity drop in the negative valence category, as shown in (a) and (b).

4.1.3 Sonnet 4.6

Just like Gemini and GPT-5.5, Sonnet 4.6 played low arousal metaphors roughly twice as soft as high arousal metaphors ($b = -52.85$, Bayesian 95 % Credible Interval $[-58.13, -47.81]$). Besides, Sonnet 4.6 played emotion metaphors slightly louder than action metaphors ($b = 9.34$, 95 % CrI $[5.28, 13.22]$). Unlike GPT-5.5, Sonnet 4.6 also demonstrated a significant main effect for valence, playing positive metaphors softer than negative metaphors ($b = -8.71$, 95 % CrI $[-11.64, -5.76]$). However, melody had no significant impact on velocity ($b = 0.35$, 95 % CrI $[-2.06, 2.80]$). We also found several significant interactions between valence, arousal and type. When looking at positive action metaphors, the difference in velocity between high and low arousal was decreased ($b = 16.46$, 95 % CrI $[13.38, 19.59]$). This effect was augmented for positive emotion metaphors ($b = 8.91$, 95 % CrI $[4.83, 12.96]$). Finally, Sonnet 4.6 played low arousal emotion metaphors significantly softer than low arousal action metaphors ($b = -21.82$, 95 % CrI $[-24.52, -19.04]$). These results are visualized in Figure 4.

Overall, Sonnet 4.6 demonstrated similar behavior to the pianists. Although Sonnet 4.6 correctly captured the main tendencies, it consistently exaggerated the effects. Compared to the pianists, Sonnet 4.6 exaggerated the decrease in velocity for low arousal compared to high arousal metaphors ($b = -31.80$, 95 % CrI $[-38.19, -25.09]$). Additionally, it reduced the difference between high and low arousal for positive metaphors far more than the pianists ($b = 12.32$, 95 % CrI $[8.20, 16.47]$). While the pianists slightly augmented this effect for action metaphors, Sonnet

4.6 deviated significantly ($b = 19.28$, 95 % CrI $[14.30, 24.38]$), resulting in an opposite effect. Comparable to Gemini, valence and type had a significant effect on Sonnet's velocity, which was not the case with the pianists. Lastly, Sonnet 4.6's estimate for the interaction between arousal and type deviated severely from the human baseline ($b = -29.29$, 95 % CrI $[-32.78, -25.82]$).



(a) Velocity for each metaphor

(b) Velocity by valence/arousal

Figure 4: The keypress velocity across eight Sonnet 4.6 simulations for the two melodies and eight metaphors. In (a) action metaphors are shown in blue and purple, whereas emotion metaphors are shown in green and red. Positive metaphors are displayed in blue and green, negative metaphors in purple and red. High arousal metaphors are visualized with stronger color intensity, while low arousal metaphors are more transparent. Regardless of metaphor type and melody, Sonnet 4.6 played low arousal metaphors with less force, with a greater velocity drop in the negative valence category, as shown in (a) and (b).

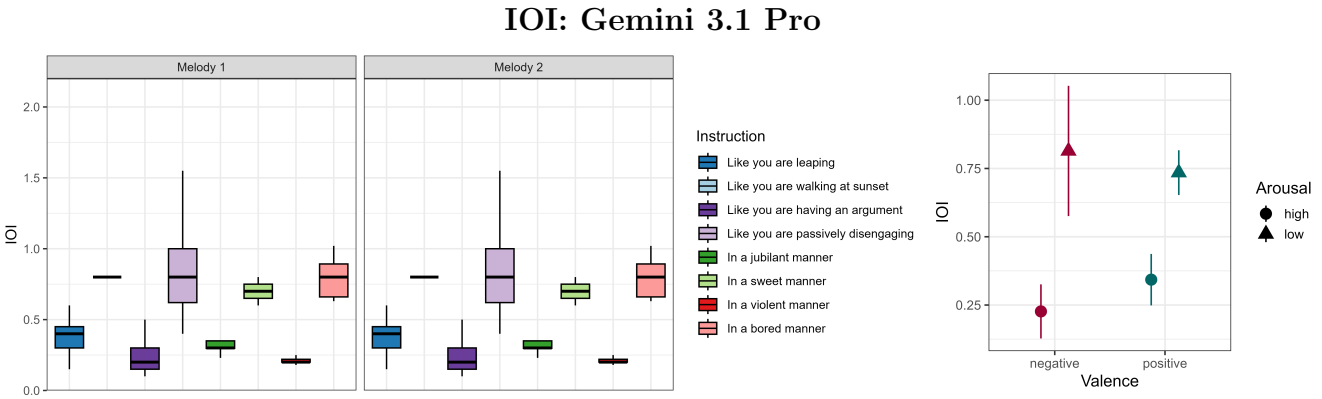
4.2 Inter-onset interval (Tempo)

4.2.1 Gemini 3.1 Pro

Gemini 3.1 Pro created drastic tempo modifications based on the metaphor's arousal. When interpreting low arousal metaphors, Gemini increased its Inter-Onset Intervals (IOI) immensely, playing approximately 3.6 times slower than it did for high arousal metaphors ($b = 1.289$, Bayesian 95 % Credible Interval $[1.133, 1.436]$, $e^{1.289} \approx 3.6$). Furthermore, when looking strictly at high arousal metaphors, positive valence elicited a slower tempo, causing a roughly 60% increase in IOI duration compared to negative metaphors ($b = 0.473$, 95 % CrI $[0.370, 0.577]$). However, the model showed a significant interaction between valence and arousal. When metaphors were both positive and low arousal, the massive slowdown was heavily mitigated ($b = -0.505$, 95 % CrI $[-0.591, -0.420]$). In other words, positive valence only caused a slower tempo in high arousal contexts. Melody ($b = -0.000$, 95 % CrI $[-0.052, 0.053]$) and type ($b = -0.068$, 95 % CrI $[-0.223, 0.093]$) had no significant influence on IOI. These results are visualized in Figure 5.

When evaluated against the human baseline, Gemini generally captured the correct tendencies of the pianists but severely exaggerated the proportions. Gemini multiplied the human tendency to slow down for low arousal metaphors by an additional factor of 2.26 ($b = 0.815$, 95 % CrI $[0.507, 1.115]$). Similarly, it significantly overshot the human tempo reduction for positive high

arousal metaphors ($b = 0.262$, 95 % CrI $[0.065, 0.458]$). While human musicians also mitigated the low arousal slowdown when playing positive metaphors, Gemini overcompensated during this interaction as well ($b = -0.245$, 95 % CrI $[-0.451, -0.037]$). Finally, Gemini slightly deviated from the human behavior for the interaction between arousal and type. While human pianists barely slowed down when playing low arousal emotion metaphors as opposed to low arousal action metaphors, Gemini notably accelerated its relative tempo for this exact combination ($b = -0.184$, 95 % CrI $[-0.355, -0.013]$).



(a) IOI for each metaphor

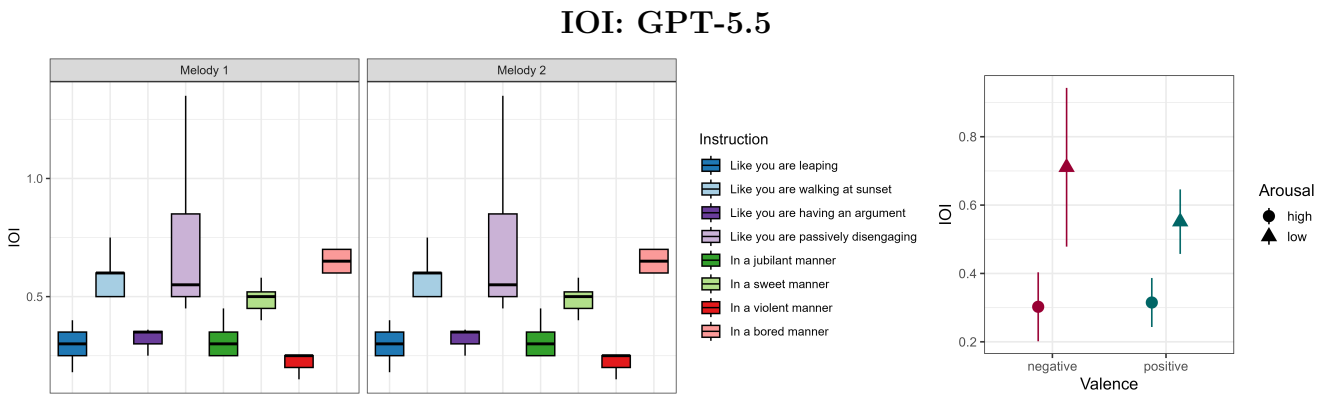
(b) IOI by valence/arousal

Figure 5: Inter-onset interval across eight Gemini 3.1 Pro simulations for the two melodies and eight metaphors. In (a) action metaphors are shown in blue and purple, whereas emotion metaphors are shown in green and red. Positive metaphors are displayed in blue and green, negative metaphors in purple and red. High arousal metaphors are visualized with stronger color intensity, while low arousal metaphors are more transparent. Regardless of metaphor type and melody, Gemini contrasted high and low arousal by significantly slowing down for low arousal metaphors, as shown in (a) and (b).

4.2.2 GPT-5.5

GPT-5.5's tempo was significantly influenced by all parameters except melody ($b = 0.000$, Bayesian 95 % Credible Interval $[-0.029, 0.029]$). It played low arousal metaphors roughly 1.8 times slower than it played high arousal metaphors ($b = 0.587$, 95 % CrI $[0.465, 0.700]$). Conversely, we found several complex interactions influencing the tempo. Positive valence reduced the IOI by approximately 16% for high arousal action metaphors ($b = -0.178$, 95 % CrI $[-0.283, -0.071]$, but for low arousal action metaphors the effect was the opposite ($b = 0.124$, 95 % CrI $[0.075, 0.174]$). While positive valence caused a slightly faster tempo on average, this effect was significantly countered for low arousal action metaphors ($b = 0.124$, 95 % CrI $[0.075, 0.174]$) and for high arousal emotion metaphors ($b = 0.484$, 95 % CrI $[0.443, 0.524]$). The model also revealed that emotion type reduced IOI by 34% for high arousal negative metaphors ($b = -0.419$, 95 % CrI $[-0.520, -0.314]$), but when interpreting low arousal negative metaphors this effect inverted ($b = 0.533$, 95 % CrI $[0.493, 0.573]$). However, this severe interaction shift was drastically mitigated when combined with positive valence ($b = -0.771$, 95 % CrI $[-0.828, -0.714]$). Ultimately, this massive three-way interaction drives the overall behavioral average, resulting in the visually smaller arousal gap for positive metaphors seen in Figure 6b. The complete results are visualized in Figure 6.

When compared directly to the human pianists, GPT-5.5 correctly slowed down when prompted with low arousal metaphors, without significantly overreacting ($b = 0.175$, 95 % CrI $[-0.122, 0.464]$). However, GPT-5.5 deviated from the human behavior for all the interactions. The model played positive metaphors ($b = -0.376$, 95 % CrI $[-0.565, -0.183]$) and emotion metaphors ($b = -0.183$, 95 % CrI $[-0.329, -0.033]$) significantly faster than the human baseline. Moreover, GPT-5.5 consistently overcompensated in its two-way interactions. It significantly amplified the human tempo adjustments for positive low arousal metaphors ($b = 0.339$, 95 % CrI $[0.139, 0.538]$), positive emotion metaphors ($b = 0.371$, 95 % CrI $[0.210, 0.526]$), and low arousal emotion metaphors ($b = 0.297$, 95 % CrI $[0.135, 0.456]$). Finally, GPT-5.5's calculation for the three-way interaction between valence, arousal, and type fell severely below the estimates found in the human data ($b = -0.594$, 95 % CrI $[-0.817, -0.368]$).



(a) IOI for each metaphor

(b) IOI by valence/arousal

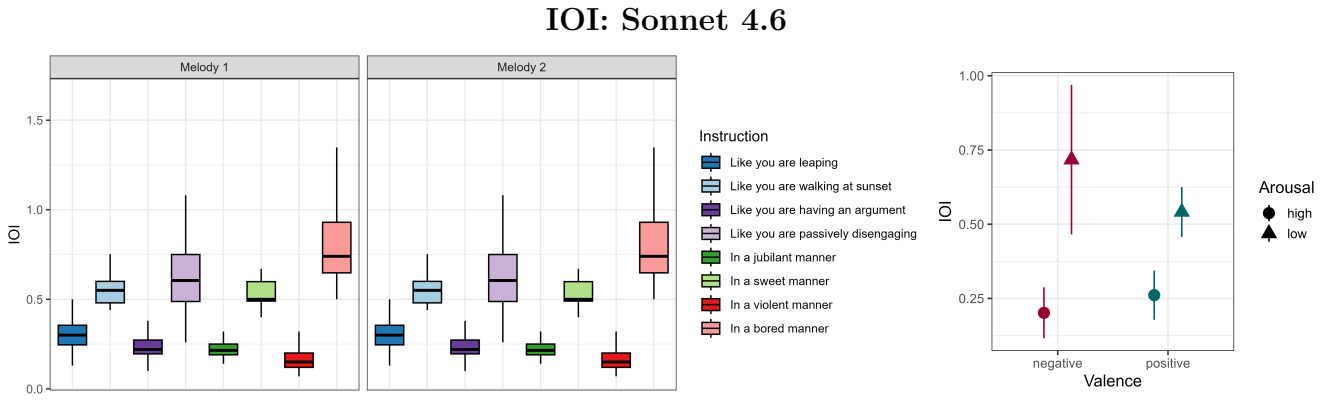
Figure 6: Inter-onset interval across eight GPT-5.5 simulations for the two melodies and eight metaphors. In (a) action metaphors are shown in blue and purple, whereas emotion metaphors are shown in green and red. Positive metaphors are displayed in blue and green, negative metaphors in purple and red. High arousal metaphors are visualized with stronger color intensity, while low arousal metaphors are more transparent. Regardless of metaphor type and emotion, GPT-5.5 played low arousal metaphors slower, as shown in (a) and (b).

4.2.3 Sonnet 4.6

Sonnet 4.6 also showed rhythmic sensitivity to the parameters of the metaphors, modifying its tempo across multiple conditions. Only melody had no significant effect ($b = -0.001$, Bayesian 95 % Credible Interval $[-0.048, 0.045]$). Under baseline parameters, Sonnet played low arousal metaphors approximately 2.6 times slower than high arousal metaphors ($b = 0.946$, 95 % CrI $[0.750, 1.131]$). Furthermore, positive valence contributed to a slower tempo for high arousal metaphors, increasing the IOI by roughly 28% ($b = 0.243$, 95 % CrI $[0.092, 0.402]$). Besides, the massive tempo gap between high and low arousal was notably reduced when combined with positive valence ($b = -0.321$, 95 % CrI $[-0.399, -0.246]$). Emotion metaphors were interpreted to be played significantly faster than action metaphors in high arousal contexts, reducing the IOI by approximately 34% ($b = -0.409$, 95 % CrI $[-0.476, -0.344]$). Additionally, Sonnet 4.6 completely inverted this rule for emotion types depending on the arousal level. While it normally played emotion metaphors

faster, it decreased the tempo when processing low arousal emotion metaphors ($b = 0.685$, 95 % CrI $[0.622, 0.747]$). However, similar to the behavior observed in GPT-5.5, this interaction shift was heavily mitigated when Sonnet processed metaphors that were simultaneously positive, low arousal, and emotion based ($b = -0.419$, 95 % CrI $[-0.508, -0.332]$). These results are visualized in Figure 7.

Generally, Sonnet correctly captured the global trends but consistently overstated their proportions. The model exaggerated the human tendency to slow down for low arousal concepts by an additional factor of 1.6 ($b = 0.484$, 95 % CrI $[0.180, 0.773]$). It also deviated from the human standard by typically playing emotion metaphors faster than action metaphors ($b = -0.188$, 95 % CrI $[-0.331, -0.039]$). Lastly, Sonnet's estimates for several interactions significantly overshot the human data. The agent significantly exaggerated the human tempo adjustments for low arousal emotion metaphors ($b = 0.465$, 95 % CrI $[0.303, 0.626]$) while similarly deviating from human rhythmic behavior during the three-way interaction between valence, arousal, and type ($b = -0.257$, 95 % CrI $[-0.483, -0.033]$).



(a) IOI for each metaphor

(b) IOI by valence/arousal

Figure 7: Inter-onset interval across eight Sonnet 4.6 simulations for the two melodies and eight metaphors. In (a) action metaphors are shown in blue and purple, whereas emotion metaphors are shown in green and red. Positive metaphors are displayed in blue and green, negative metaphors in purple and red. High arousal metaphors are visualized with stronger color intensity, while low arousal metaphors are more transparent. Regardless of metaphor type and emotion, Sonnet 4.6 played low arousal metaphors slower, as shown in (a) and (b).

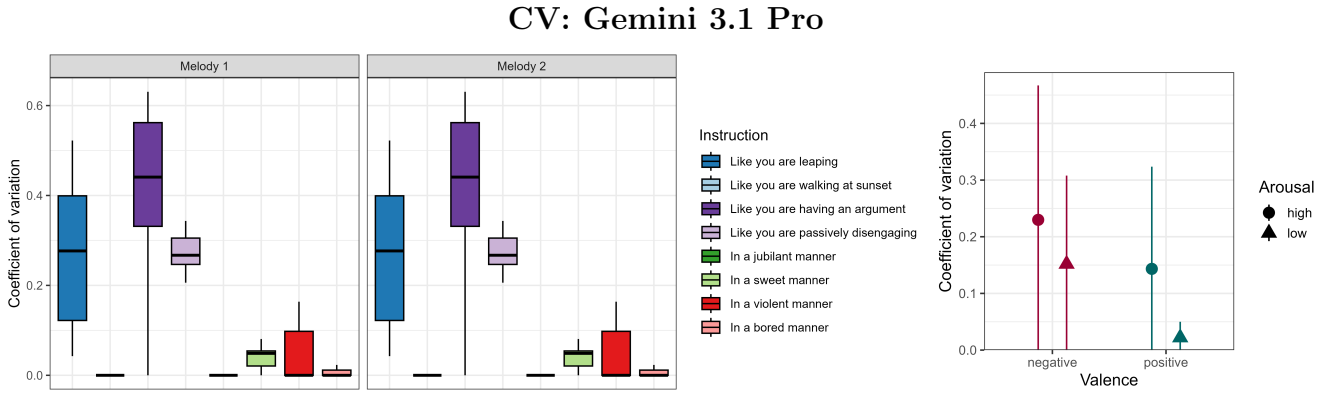
4.3 Coefficient of variation (Rhythmic variability)

4.3.1 Gemini 3.1 Pro

Gemini 3.1 Pro showed little expressive rhythmic variability, often struggling to generate organic rhythm and instead defaulting to rigid, robotic timing. The model often produced completely uniform IOIs throughout entire trials, resulting in an effective CV of zero. For the low arousal positive action metaphor and high arousal positive emotion metaphor trials, Gemini did not manage to generate a CV higher than zero at all. Consequently, the statistical analysis struggled measuring Gemini's behavior, yielding only two statistically significant interactions.
Specifically, when processing action metaphors, combining positive valence with low arousal re-

sulted in a complete collapse of rhythmic variation ($b = -2.341$, Bayesian 95 % Credible Interval $[-3.453, -1.202]$), suggesting that the arousal gap for positive valence metaphors is huge. However, this was heavily counteracted in a significant three-way interaction when the low arousal positive metaphor was emotion based rather than action based ($b = 3.922$, 95 % CrI $[2.495, 5.259]$). This represents a massive 50-fold increase in CV ($e^{3.922} \approx 50.5$), indicating that Gemini abruptly shifted from robotic constant rhythm to more expressive rhythmic variability. Noteworthy is that low arousal, action type and negative valence all independently elicited more rhythmic variability than their opposites on average. This is shown clearly in Figure 8.

When evaluated against human performance, Gemini 3.1 Pro's rhythmic variability deviated severely from the organic rhythmic variability of the pianists. While human musicians organically vary with rhythm based on the metaphors, Gemini evidently struggled simulating natural rhythmic variability. For only two metaphors, namely "Like you are leaping" and "Like you are passively engaging", Gemini managed to emulate rhythmic variability for all corresponding trials. Despite this, some general patterns seen in Gemini's data are similar to human tendencies, although Gemini tends to overadjust again, as we have seen earlier. For example, increased CV for negative versus positive valence and increased CV for high versus low arousal can be observed for both the pianists and Gemini, although the LLM exaggerated the proportional differences. Lastly, Gemini's tendency to increase CV for action metaphors compared to emotion metaphors, is not reflected in the human data.



(a) CV for each metaphor

(b) CV by valence/arousal

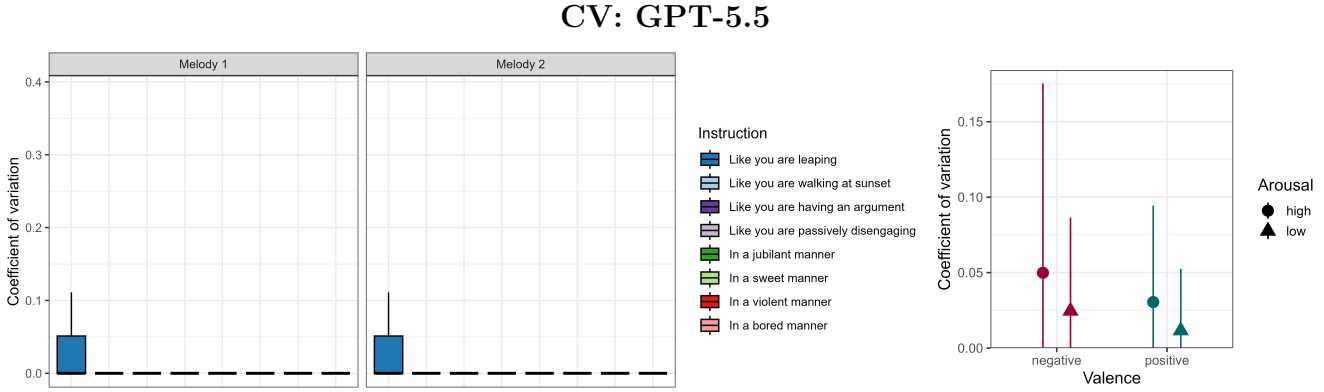
Figure 8: Coefficient of variation (CV) across eight Gemini 3.1 Pro simulations for the two melodies and eight metaphors. In (a) action metaphors are shown in blue and purple, whereas emotion metaphors are shown in green and red. Positive metaphors are displayed in blue and green, negative metaphors in purple and red. High arousal metaphors are visualized with stronger color intensity, while low arousal metaphors are more transparent. Although Gemini often struggled to emulate rhythmic variability, it generally decreased CV for low arousal metaphors and for positive metaphors, as shown in (a) and (b).

4.3.2 GPT-5.5

GPT-5.5 displayed extremely little rhythmic variability rather than organic rhythm. A close examination of the trial-level data reveals that GPT-5.5 operated almost exclusively in one of two

extremes: metronomic consistency or big systematic fluctuations. For instance, the vast majority of trials yielded a CV of exactly zero. This included 87.5% of trials for negative high-arousal action metaphors ("in a violent manner") and 75% of trials for positive high-arousal action metaphors ("like you are leaping"). However, these trials with perfectly constant rhythm were occasionally interrupted by trials where GPT-5.5 varied with rhythm, with CVs reaching up to 0.389. Because GPT-5.5's CV was polarized between zero and 'extreme' outliers, the statistical model identified no significant results, and the mathematical median ended up being zero for all metaphors, as visually evident in Figure 9a). As "like you are leaping" was the only metaphor that had four trials with a non-zero CV for both melodies, it was the only metaphor with an upper quartile higher than zero, resulting in the interquartile range being drawn. Looking at the outlier data where GPT-5.5 attempted to implement rhythmic variability, reveals that GPT-5.5 generated a slightly higher CV for negative and for high arousal metaphors, as visualized in Figure 9b.

Evidently, GPT-5.5 exhibited a massive lack of rhythmic variability when compared to the pianists. In fact, the comparative statistical model reveals that GPT-5.5 played with only about 6.5% of the rhythmic variance seen in the human data ($b = -2.729$, Bayesian 95 % Credible Interval $[-3.957, -1.428]$). Furthermore, none of the interactions comparing GPT-5.5 to human performance were statistically significant. However, looking at the trials in which GPT-5.5 did vary with tempo, it lowered the CV for low arousal metaphors, which the pianists also structurally did.



(a) CV for each metaphor

(b) CV by valence/arousal

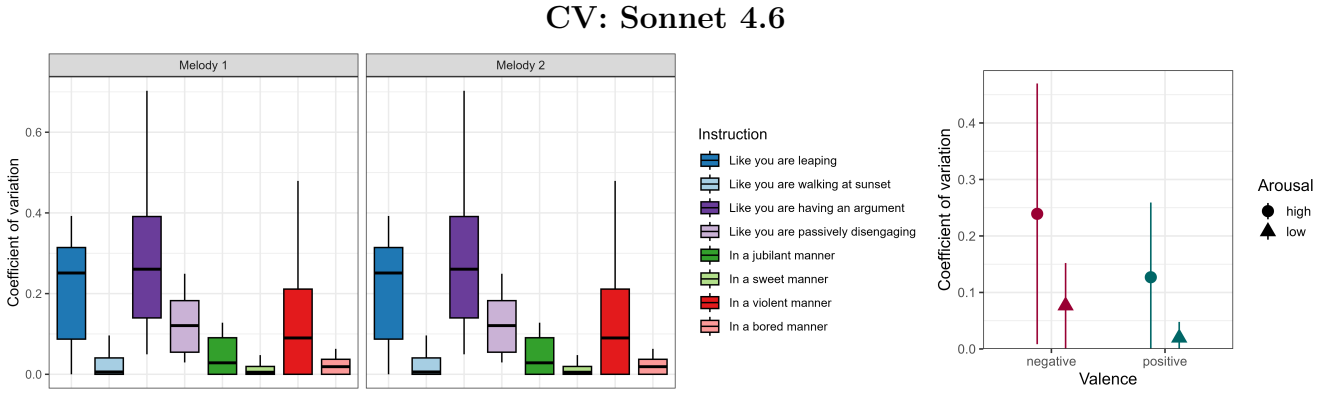
Figure 9: Coefficient of variation (CV) across eight GPT-5.5 simulations for the two melodies and eight metaphors. In (a) action metaphors are shown in blue and purple, whereas emotion metaphors are shown in green and red. Positive metaphors are displayed in blue and green, negative metaphors in purple and red. High arousal metaphors are visualized with stronger color intensity, while low arousal metaphors are more transparent. Robotic, constant rhythm plagued GPT-5.5's CV as shown in (a), with only a few outliers showing rhythmic variability, which is visualized in (b).

4.3.3 Sonnet 4.6

Out of the three LLMs, Sonnet 4.6 demonstrated the most rhythmic variance within trials. It played low arousal metaphors with significantly less variation than high arousal metaphors ($b = -0.927$, Bayesian 95 % Credible Interval $[-1.706, -0.159]$), representing an approximate 60.4% reduction in

CV. Similarly, the LLM played emotion metaphors with roughly 81.3% less rhythmic variability than action metaphors ($b = -1.679$, 95 % CrI $[-2.631, -0.516]$). Furthermore, when valence was positive, the arousal gap grew immensely for action metaphors ($b = -1.332$, 95 % CrI $[-2.281, -0.363]$). However, this effect was inverted for emotion metaphors ($b = 1.418$, 95 % CrI $[0.315, 2.521]$). These results are visualized in Figure 10.

When evaluated against human performance, Sonnet showed two large deviations. While human pianists organically adjust tempo within the trials, Sonnet significantly exaggerated the CV reduction for emotion metaphors compared to the human baseline, reducing its relative variability to about 25.6% of what the pianists performed ($b = -1.364$, 95 % CrI $[-2.339, -0.317]$). Additionally, Sonnet's tendency to drop its rhythmic variation for low arousal metaphors in the positive category drastically overshot human behavior by roughly 69.2% ($b = -1.178$, 95 % CrI $[-2.131, -0.173]$). These deviations show that while Sonnet shows an understanding of rhythmic variability, it utilizes large adjustments as opposed to the nuanced expressiveness of human musicians.



(a) CV for each metaphor

(b) CV by valence/arousal

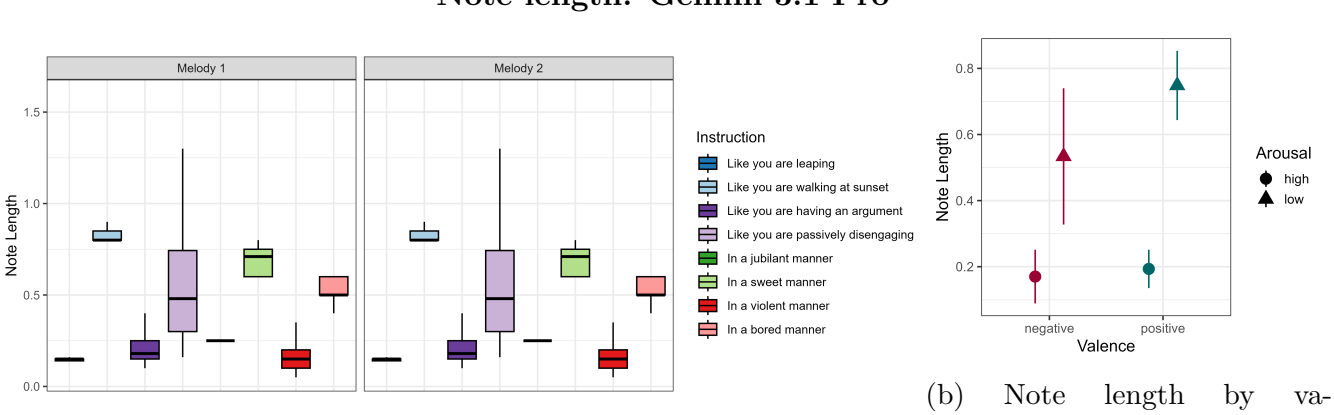
Figure 10: Coefficient of variation (CV) across eight Sonnet 4.6 simulations for the two melodies and eight metaphors. In (a) action metaphors are shown in blue and purple, whereas emotion metaphors are shown in green and red. Positive metaphors are displayed in blue and green, negative metaphors in purple and red. High arousal metaphors are visualized with stronger color intensity, while low arousal metaphors are more transparent. Sonnet 4.6 generally plays high arousal metaphors with higher CV and emotion metaphors with lower CV, as shown in (a) and (b).

4.4 Note length

4.4.1 Gemini 3.1 Pro

Gemini 3.1 Pro exhibited stark contrasts in its note lengths depending on the metaphors. It generated significantly longer notes for low arousal metaphors, increasing note lengths by a factor of 2.7 compared to high arousal metaphors ($b = 1.002$, Bayesian 95 % Credible Interval $[0.888, 1.110]$). Conversely, the model interpreted emotion metaphors with slightly shorter notes at the baseline, reducing the note length by approximately 27% compared to action metaphors ($b = -0.318$, 95 % CrI $[-0.510, -0.115]$). Melody had no significant effect ($b = 0.000$, 95 % CrI $[-0.048, 0.049]$). Gemini's logic also relied on significant interactions that shifted these baseline rules. For instance,

positive valence massively amplified the lengthening effect of low arousal for action metaphors, more than doubling the arousal gap ($b = 0.714$, 95 % CrI $[0.634, 0.791]$). Furthermore, the model's tendency to play emotion metaphors with shorter notes was completely overridden by both positive valence and low arousal. Combining emotion types with positive valence ($b = 0.856$, 95 % CrI $[0.792, 0.922]$) or low arousal ($b = 0.369$, 95 % CrI $[0.306, 0.434]$) independently increased note length significantly.
However, a large three-way interaction heavily suppressed these combined effects. When Gemini interpreted metaphors that were simultaneously positive, low arousal, and emotion based, a strong negative interaction drastically shortened the note lengths by roughly 64% compared to what the previous compounding effects would predict ($b = -1.026$, 95 % CrI $[-1.118, -0.935]$). These results are visualized in Figure 11.



(a) Note length for each metaphor

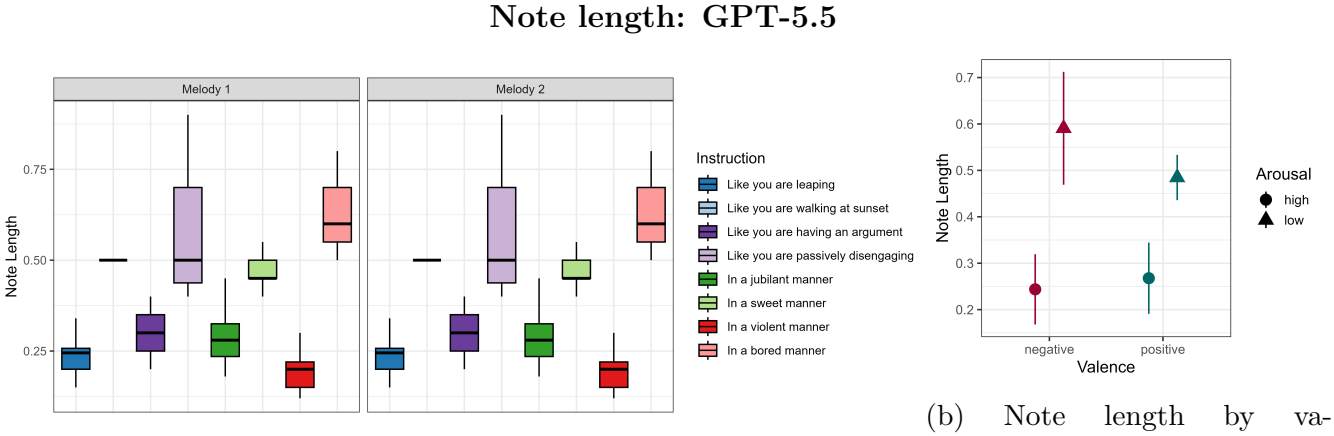
(b) Note length by valence/arousal

Figure 11: Note length across eight Gemini 3.1 Pro simulations for the two melodies and eight metaphors. In (a) action metaphors are shown in blue and purple, whereas emotion metaphors are shown in green and red. Positive metaphors are displayed in blue and green, negative metaphors in purple and red. High arousal metaphors are visualized with stronger color intensity, while low arousal metaphors are more transparent. Gemini interpreted high and low arousal metaphors very differently, utilizing short notes for high arousal and longer notes for low arousal metaphors. In the action type scope, Gemini increased this contrast for positive metaphors, as shown in (a) and (b).

4.4.2 GPT-5.5

GPT-5.5's note lengths were also mostly dependent on arousal level. In the baseline condition, the model nearly doubled note duration when playing low arousal metaphors compared to high arousal ones, resulting in a roughly 92% increase ($b = 0.651$, Bayesian 95 % Credible Interval $[0.449, 0.840]$). While melody had no significant impact ($b = 0.000$, 95 % CrI $[-0.024, 0.025]$), both positive valence and emotion metaphors independently elicited shorter notes. Positive valence caused a 16% reduction in note length ($b = -0.178$, 95 % CrI $[-0.285, -0.065]$), while emotion metaphors yielded notes that were approximately 33% shorter than those generated for action metaphors ($b = -0.396$, 95 % CrI $[-0.477, -0.318]$).
GPT-5.5's behavior was further complicated by several two-way interactions that changed these baseline rules. While positive valence generally shortened notes for high arousal action metaphors,

it slightly expanded the note length gap between low and high arousal for action metaphors by 11% ($b = 0.107$, 95 % CrI $[0.065, 0.150]$). Moreover, the model completely reversed its tendency to play emotion metaphors with shorter notes for metaphors with different parameters. Pairing an emotion metaphor with positive valence counteracted the compounding shortening effect with an offset of 77% ($b = 0.569$, 95 % CrI $[0.536, 0.602]$), meaning that positive high arousal emotion metaphors ended up with virtually the same note lengths as the baseline negative high arousal action metaphors ($e^{-0.396-0.178+0.569} = e^{-0.005} \approx 0.995$). Combining a negative emotion metaphor with low arousal similarly stretched the note lengths by about 67% ($b = 0.511$, 95 % CrI $[0.478, 0.543]$), making the arousal gap the biggest of all four valence/type combinations, as is visible in Figure 12a.
Ultimately, a strong three-way interaction balanced these compounding effects. For metaphors that were positive, low arousal, and emotion based all at once, GPT-5.5 applied a strong negative correction, cutting the expected note length by roughly 54% ($b = -0.775$, 95 % CrI $[-0.822, -0.727]$). These results are visualized in Figure 12.



(a) Note length for each metaphor

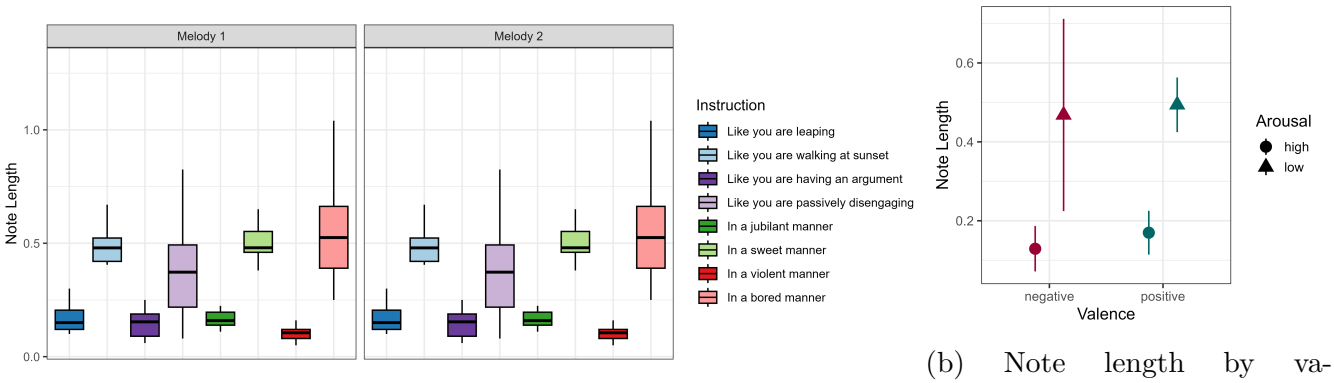
(b) Note length by valence/arousal

Figure 12: Note length across eight GPT-5.5 simulations for the two melodies and eight metaphors. In (a) action metaphors are shown in blue and purple, whereas emotion metaphors are shown in green and red. Positive metaphors are displayed in blue and green, negative metaphors in purple and red. High arousal metaphors are visualized with stronger color intensity, while low arousal metaphors are more transparent. GPT-5.5 typically played high arousal metaphors with notes roughly twice as short compared to low arousal metaphors, as shown in (a) and (b).

4.4.3 Sonnet 4.6

Under baseline conditions, Sonnet 4.6 held notes significantly longer for low arousal metaphors, increasing the note length by a factor of 2.3 compared to high arousal metaphors ($b = 0.830$, Bayesian 95 % Credible Interval $[0.639, 1.012]$). In contrast, the model interpreted emotion metaphors with notably shorter notes by approximately 28% ($b = -0.332$, 95 % CrI $[-0.414, -0.253]$). Valence ($b = 0.159$, 95 % CrI $[-0.153, 0.438]$) and melody ($b = -0.000$, 95 % CrI $[-0.045, 0.043]$) had no significant impact.
Similar to the other LLMs, Sonnet relied on several significant interactions that heavily modified these baseline tendencies. Positive valence further increased the lengthening effect of low arousal for action metaphors, adding a 28% increase to the existing gap ($b = 0.247$, 95 % CrI $[0.174, 0.321]$).

Besides, combining emotion metaphors with positive valence increased note lengths by roughly 36% ($b = 0.308$, 95 % CrI $[0.249, 0.367]$), and pairing them with low arousal increased note lengths by a massive 120% ($b = 0.788$, 95 % CrI $[0.729, 0.848]$).
Finally, these compounding effects on note lengths was regulated by a massive three-way interaction. When Sonnet interpreted positive low arousal emotion metaphors, a negative interaction strongly suppressed the effects. This specific combination reduced the expected note duration by approximately 53% ($b = -0.746$, 95 % CrI $[-0.829, -0.662]$). These results are visualized in Figure 13.



(a) Note length for each metaphor

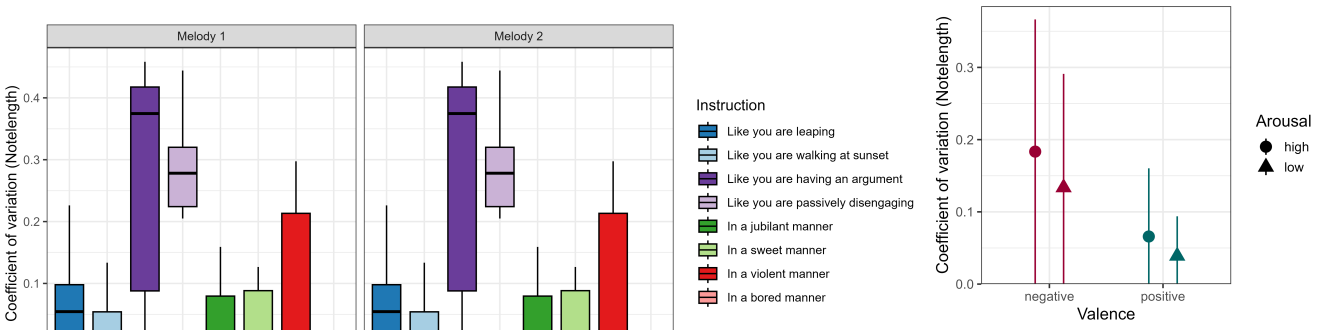
(b) Note length by valence/arousal

Figure 13: Note length across eight Sonnet 4.6 simulations for the two melodies and eight metaphors. In (a) action metaphors are shown in blue and purple, whereas emotion metaphors are shown in green and red. Positive metaphors are displayed in blue and green, negative metaphors in purple and red. High arousal metaphors are visualized with stronger color intensity, while low arousal metaphors are more transparent. Sonnet 4.6 played low arousal metaphors with much longer notes, making a heavy contrast with high arousal metaphors, as shown in (a) and (b).

4.5 Coefficient of variation of note length (Note length variability)

4.5.1 Gemini 3.1 Pro

Gemini 3.1 Pro typically showed no variance in note length (note length CV) within trials. For five out of eight metaphors, the note length CV median was zero. Remarkably, negative action metaphors stood out with medians above 0.25, including the baseline metaphor. Because the baseline note length CV was an extreme outlier among all metaphors, the statistical model applied Bayesian partial pooling, shrinking the baseline toward Gemini's global tendency to play with zero note length variance. This explains the discrepancy between the median of the baseline "Like you are having an argument" of roughly 0.365 seen in Figure 14a and the baseline intercept of approximately 0.044 ($b = -3.131$, Bayesian 95 % Credible Interval $[-5.244, -1.209]$, $e^{-3.131} \approx 0.044$).
With this in mind, Gemini played emotion metaphors without any note length variability most of the times, reducing the baseline variance by approximately 76%, to virtually zero ($b = -1.446$, 95 % CrI $[-2.593, -0.102]$). Additionally, it is notable that Gemini applied note length variability more often to action metaphor trials (36/64) than trials with emotion metaphors (18/64). The full results are visualized in Figure 14.



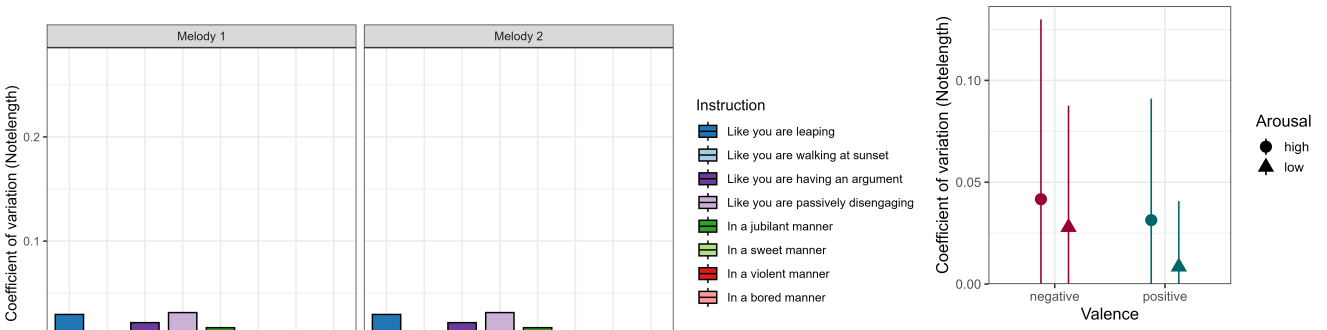
(a) Note length CV for each metaphor

(b) Note length CV by valence/arousal

Figure 14: Coefficient of variation of note length across eight Gemini 3.1 Pro simulations for the two melodies and eight metaphors. In (a) action metaphors are shown in blue and purple, whereas emotion metaphors are shown in green and red. Positive metaphors are displayed in blue and green, negative metaphors in purple and red. High arousal metaphors are visualized with stronger color intensity, while low arousal metaphors are more transparent. Gemini 3.1 Pro struggled implementing note length variance, succeeding more with action metaphors than emotion metaphors, as seen in (a).

4.5.2 GPT-5.5

GPT-5.5 applied note length variance to just 26 out of 128 trials, yielding no statistically significant main effects, indicating static playing across the metaphors. However, there were two significant interactions. When processing action metaphors, the combination of positive valence and low arousal triggered a significant collapse in variation, pulling note length CV to zero ($b = -0.574$, 95 % CrI $[-1.093, -0.060]$). A similar reduction occurred when GPT-5.5 evaluated low arousal emotion metaphors ($b = -0.514$, 95 % CrI $[-0.943, -0.075]$). Ultimately, GPT-5.5's behavior for note durations lacked organic expressiveness entirely. Its only significant interactions only suppressed note length variance further from an already rigid, robotic baseline. The full results are visualized in Figure 15.



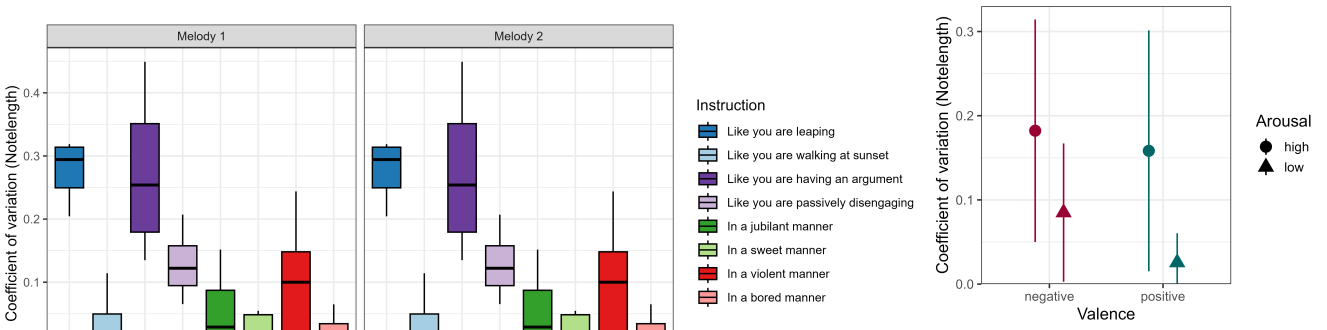
(a) Note length CV for each metaphor

(b) Note length CV by valence/arousal

Figure 15: Coefficient of variation of note length across eight GPT-5.5 simulations for the two melodies and eight metaphors. In (a) action metaphors are shown in blue and purple, whereas emotion metaphors are shown in green and red. Positive metaphors are displayed in blue and green, negative metaphors in purple and red. High arousal metaphors are visualized with stronger color intensity, while low arousal metaphors are more transparent. GPT-5.5 exhibited virtually zero note length variability, as shown in (a) and (b).

4.5.3 Sonnet 4.6

Sonnet 4.6 demonstrated the most note length variability of the tested LLMs (88/128 trials), as every metaphor had a note length CV median above zero. Under baseline parameters, Sonnet exhibited significantly less note length variation for low arousal metaphors compared to high arousal metaphors, by around 58% ($b = -0.878$, 95 % CrI $[-1.692, -0.024]$). This variance penalty was even more severe for emotion metaphors, which yielded an 83% reduction compared to action metaphors ($b = -1.774$, 95 % CrI $[-2.879, -0.387]$). Furthermore, when positive valence was paired with low arousal for baseline action metaphors, the note length CV collapsed by an additional 79% ($b = -1.575$, 95 % CrI $[-2.519, -0.634]$). However, Sonnet applied a drastic countermeasure during the three-way interaction. When interpreting positive low arousal emotion metaphors, the model generated a massive 550% increase in variance, though the note length CV remains close to zero ($b = 1.872$, 95 % CrI $[0.772, 2.964]$). The full results are visualized in Figure 16.



(a) Note length CV for each metaphor

(b) Note length CV by valence/arousal

Figure 16: Coefficient of variation of note length across eight Sonnet 4.6 simulations for the two melodies and eight metaphors. In (a) action metaphors are shown in blue and purple, whereas emotion metaphors are shown in green and red. Positive metaphors are displayed in blue and green, negative metaphors in purple and red. High arousal metaphors are visualized with stronger color intensity, while low arousal metaphors are more transparent. Sonnet showed more note length variance for high arousal metaphors than low arousal metaphors, with a larger arousal gap for positive metaphors, as shown in (a) and (b).

5 Discussion

5.1 Evaluation of the results

In this study based on the experiment with six pianists by Schaefer and Verhoef [SV24], we prompted three state-of-the-art LLMs (Gemini 3.1 Pro, GPT-5.5 and Claude Sonnet 4.6) to simulate pianists playing two melodies while emulating musical expression in the form of metaphors varying in several semantic dimensions, being valence (positive/negative), arousal (high/low) and type (action/emotion). We collected the motor performance data to evaluate how the LLMs reacted to the metaphors. This data included measures such as keypress force, tempo, tempo variability, note length and note length variability. By comparing this data to the human data, we aim to examine how similar LLMs translate musical metaphors to motor performance compared to humans.

We found strong behavioral patterns from all three LLMs across the five motor performance metrics. Among the three semantic dimensions tested, arousal ended up being the most dominant driver of musical expression across the LLMs. Low arousal metaphors consistently elicited softer keypresses, slower tempos, and longer note lengths, whereas high arousal metaphors triggered the opposite. This global response to arousal was present across Gemini 3.1 Pro, GPT-5.5, and Claude Sonnet 4.6, demonstrating that LLMs reliably associate emotional intensity with motor performance parameters.

However, there was a stark contrast between these global parametric adjustments and the models' timing flexibility, evaluated through tempo variability and note length variability. The LLMs frequently struggled to generate organic, continuous expressive fluctuations, and defaulted

to rigid, robotic timings instead. This was especially noticeable in GPT-5.5, which only applied timing variability for a small portion of the experiment, resulting in a lot of zero-variance trial data with 'high' variance outliers. Similarly, Gemini 3.1 Pro had a collapse in note length variability when processing emotion metaphors, yielding medians of zero.

When evaluated against the pianists, the LLMs successfully approximated human musical expressivity, translating arousal, valence, and metaphor types to similar directional shifts in motor performance as the pianists. However, the models consistently exaggerated these adjustments, while simultaneously stripping away expressive variance, especially evident as they applied the exact same motor performance rules regardless of the melody played. This tendency aligns with established findings in LLM evaluation, where they often reduce variance and exaggerate contrasts when translating nuanced concepts [ANE+24].

Additionally, the strong similarity in directional tendencies across all three independently developed LLMs points to comparable and shared training material. These models have been trained on data including musical pedagogy literature, performance science research, and music theory textbooks. When prompted with metaphorical instructions, the LLMs likely retrieve associations between expressive principles and cues in motor performance, that are described in these texts, for example on pp. 86-87 of Gabrielsson and Juslin's research [GJ96].

5.2 Limitations and future research

Several limitations must be acknowledged when interpreting these results. The first limitation lies in the statistical behavior of the LLMs' tempo variability and note length variability. Because the models (most notably GPT-5.5 and Gemini 3.1 Pro) frequently generated robotic, zero-variance trials, the dataset suffered from zero-inflation. This impeded statistical modeling, resulting in flattened results and high standard errors in the Bayesian models. This binary behavior between robotic timing and extreme variation shows a significant limitation in current LLM capabilities: without the capacity for controlling smooth, organic variance in timing, current LLMs cannot yet serve as reliable intermediaries for translating nuanced human musical expression into generative music systems. Though, Sonnet showed potential, and perhaps LLMs might perform more human-like when explicitly prompted to implement small variance in timing.

Besides, the evaluation of note length and its variability was constrained by the available human baseline data. While keypress velocity, tempo, and timing variability were directly compared against human pianist performances, note length and note length CV lacked a corresponding human dataset. Consequently, while the general note length behavior of the models is somewhat defined, the extent to which it mirrors actual human behavior remains unverified.

Additionally, while this study explicitly investigates how LLMs simulate motor tasks, the generation mechanism itself is heavily formulaic, regardless of melody. In music, the direction and utilized note length values of melodies also heavily influence expressive motor performance. In the future, a similar experiment could be executed, using different types of melodies that are not neutral, unlike the melodies given to the LLMs and the pianists. Possible categories are simple melodies with more pronounced direction in pitch, or famous melodies in classical music repertoire.

In general, the results show potential for innovations in music generation algorithms. However, music AI has lately been a controversial topic. Some people are concerned about models stealing copyrighted music for training. Moreover, a big reason why humans listen to music is because it is a "dynamic aspect of human interconnection" according to Finnegan [Fin02]. Music generated by AI is by some seen as a threat to culture, removing the human aspect of music. Consequently, a lot of people prefer music made by humans rather than generated by AI [EP26]. Rather than framing these models as replacements for human creation, a more constructive application lies in developing future metaphor-driven LLMs as interactive tools for musical experimentation and education. By embedding LLMs into existing music generation models, users could apply expressive metaphors to existing melodies to explore how different instructions alter performance parameters in real time. In an educational setting, this could assist students in visualizing and understanding the relationship between abstract concepts and concrete motor controls, turning music AI into expressive toolboxes rather than substitutes for human artistry.

6 Conclusion

This study investigated how state-of-the-art LLMs (Gemini 3.1 Pro, GPT-5.5, and Claude Sonnet 4.6) simulate human musical expression, benchmarking their motor performance against human pianists prompted with metaphorical instructions across valence, arousal, and type.
The results show that while LLMs successfully approximate human directional expressivity, with arousal as the main driver, they heavily exaggerate contrasts, and lack adaptation across different melodies, characterized by robotic timing.
While current models lack the timing precision required for humanlike performance, they show potential to be developed into interactive, metaphor-driven toolboxes. In educational and creative settings, these music generation models can help musicians and students experiment with expressive parameters, bridging abstract emotional intent with concrete performance control without replacing human artistry.

References

[ANE+24] Guilherme F.C.F. Almeida, José Luiz Nunes, Neele Engelmann, Alex Wiegmann, and Marcelo de Araújo. Exploring the psychology of LLMs' moral and legal reasoning. *Artificial Intelligence*, 333:104145, 2024.

[Bro25] Jacob Brown. Bridging the gap: Fostering emotional expression through bridging the gap: Fostering emotional expression through teaching strategies in the middle school choral classroom, 2025.

[BS13] Brenda Brenner and Katherine Strand. A case study of teaching musical expression to young performers. *Journal of research in music education*, 61(1):80–96, 2013.

[Bü17] Paul-Christian Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017.

[Cla01] Martin Clayton. The social and personal functions of music in cross-cultural perspective. In *The Oxford Handbook of Music Psychology*. Oxford University Press, 2016-01.

[CVD+25] Luca Casini, Laura Cros Vila, David Dalmazzo, Anna-Kaisa Kaila, and Bob L. T. Sturm. Data-driven analysis of text-conditioned ai-generated music: A case study with suno and udio, 2025.

[EP26] Andrew B. Edelblum and Joshua Poe. Liking without endorsing: Consumer dilemmas in responses to ai-generated music. *Psychology & Marketing*, 43(9):2229–2244, 2026.

[Fin02] Ruth H Finnegan. *Communicating: The multiple modes of human interconnection*. Psychology Press, 2002.

[GJ96] Alf Gabrielsson and Patrik N. Juslin. Emotional expression in music performance: Between the performer's intention and the listener's experience. *Psychology of Music*, 24(1):68–91, 1996.

[GB25] Jonah Gabry, Rok Češnovar, Andrew Johnson, and Steve Bronder. *cmdstanr: R Interface to 'CmdStan'*, 2025. R package version 0.9.0, https://discourse.mc-stan.org.

[HM01] Susan Hallam and Raymond MacDonald. The effects of music in community and educational settings. In *The Oxford Handbook of Music Psychology*. Oxford University Press, 2016-01.

[JL01] Patrik N. Juslin and Erik Lindström. Emotion in music performance. In *The Oxford Handbook of Music Psychology*. Oxford University Press, 2016-01.

[JS01] Patrik N. Juslin and John A. Sloboda. The past, present, and future of music and emotion research. In *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford University Press, 2010-01.

[JT10] Patrik N Juslin and Renee Timmers. Expression and communication of emotion in music performance. In *Handbook of Music and Emotion: Theory, Research, Applications*, pages 453–489. Oxford University Press, 2010.

[Jus01] Patrik N. Juslin. Emotional reactions to music. In *The Oxford Handbook of Music Psychology*. Oxford University Press, 2016-01.

[KND25] A. Karakanta, M.O. Nas, and A.G. Dorst. Metaphors in literary machine translation: close but no cigar?, 2025.

[KP24] Jan Kietzmann and Andrew Park. Written by chatgpt: Ai, large language models, conversational chatbots, and their place in society and business. *Business horizons*, 67(5):453–459, 2024.

[Kra87] Steven C. Krantz. Metaphor in music. *The Journal of aesthetics and art criticism*, 45(4):351–360, 1987.

[LGS01] Alexandra Lamont, Alinka Greasley, and John Sloboda. Choosing to hear music: Motivation, process, and effect. In *The Oxford Handbook of Music Psychology*. Oxford University Press, 2016-01.

[MT20] Henrique Meissner and Renee Timmers. Young musicians' learning of expressive performance: The importance of dialogic teaching and modeling. *Frontiers in Education*, Volume 5, 2020.

[R C21] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[SGRG19] Simon Schaerlaeken, Donald Glowinski, Marc-André Rappaz, and Didier Grandjean. "hearing music as . . .": Metaphors evoked by the sound of classical music. *Psychomusicology*, 29(2-3):101, 2019.

[SHYX25] Xueer Sun, Xiao Han, Fei Yao, and Jiawei Xu. Emotion-aware cross-modal music generation based on multimodal emotion recognition. *Alexandria Engineering Journal*, 133:254–270, 2025.

[Slo10] John A Sloboda. Music in everyday life. In *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford University Press, 2010.

[SV24] Rebecca Schaefer and Tessa Verhoef. Metaphors in music performance: from semantics and motor performance to expressive communication. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46:1061–1067, 2024.

[TCLS24] Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. Metaphor understanding challenge dataset for LLMs, 2024.

[Woo06] Robert H. Woody. The effect of various instructional conditions on expressive music performance. *Journal of research in music education*, 54(1):21–36, 2006.

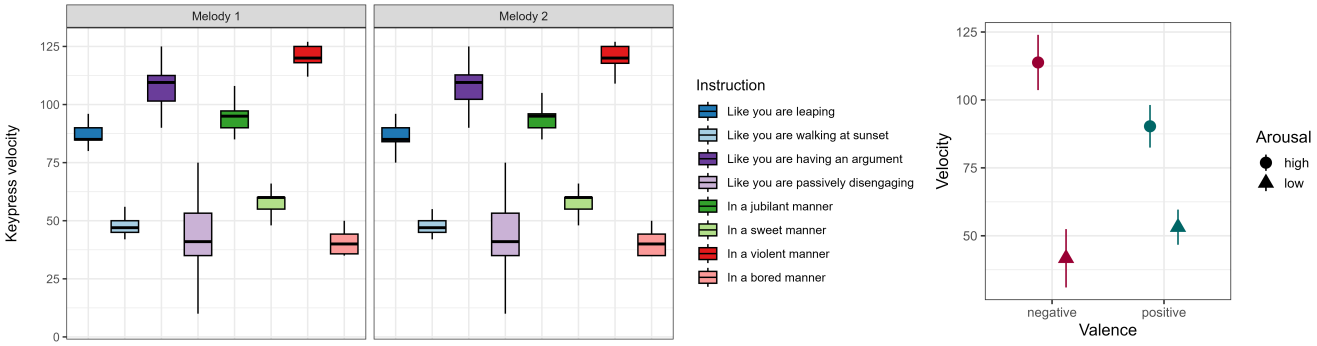
[WWH+25] Yashan Wang, Shangda Wu, Jianhuai Hu, Xingjian Du, Yueqi Peng, Yongxin Huang, Shuai Fan, Xiaobing Li, Feng Yu, and Maosong Sun. NotaGen: Advancing musicality in symbolic music generation with large language model training paradigms, 2025.

[YLG+25] Ruibin Yuan, Hanfeng Lin, Shuyue Guo, Ge Zhang, Jiahao Pan, Yongyi Zang, Haohe Liu, Yiming Liang, Wenye Ma, Xingjian Du, Xinrun Du, Zhen Ye, Tianyu Zheng, Zhengxuan Jiang, Yinghao Ma, Minghao Liu, Zeyue Tian, Ziya Zhou, Liumeng Xue, Xingwei Qu, Yizhi Li, Shangda Wu, Tianhao Shen, Ziyang Ma, Jun Zhan, Chunhui Wang, Yatian Wang, Xiaowei Chi, Xinyue Zhang, Zhenzhu Yang, Xiangzhou Wang, Shansong Liu, Lingrui Mei, Peng Li, Junjie Wang, Jianwei Yu, Guojian Pang, Xu Li, Zihao Wang, Xiaohuan Zhou, Lijun Yu, Emmanouil Benetos, Yong Chen, Chenghua Lin, Xie Chen, Gus Xia, Zhaoxiang Zhang, Chao Zhang, Wenhu Chen, Xinyu Zhou, Xipeng Qiu, Roger Dannenberg, Jiaheng Liu, Jian Yang, Wenhao Huang, Wei Xue, Xu Tan, and Yike Guo. YuE: Scaling open foundation models for long-form music generation, 2025.

[ZWW+24] Ziya Zhou, Yuhang Wu, Zhiyue Wu, Xinyue Zhang, Ruibin Yuan, Yinghao Ma, Lu Wang, Emmanouil Benetos, Wei Xue, and Yike Guo. Can LLMs "reason" in music? an evaluation of LLMs' capability of music understanding and generation. 2024.

A Experiment results with initial prompt

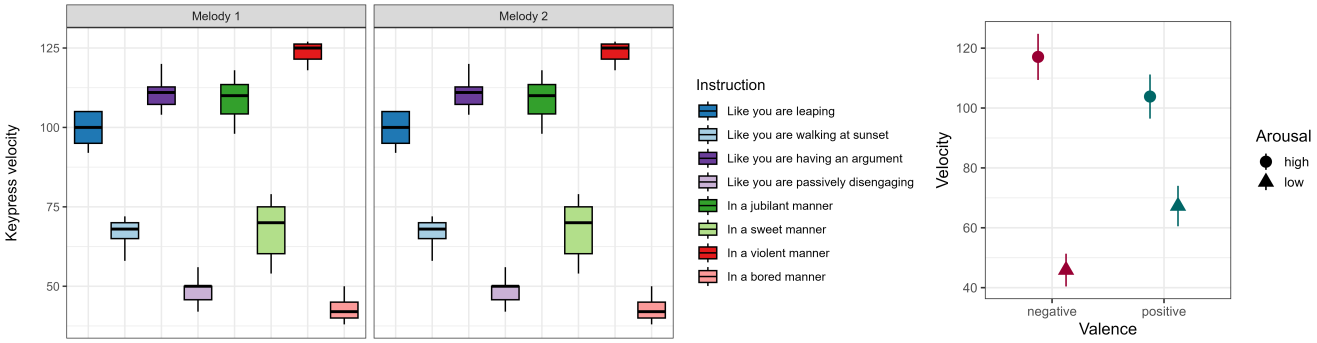
Velocity: Gemini 3.1 Pro



(a) Velocity for each metaphor (b) Velocity by valence/arousal

Figure 17: The keypress velocity across eight Gemini 3.1 Pro simulations.

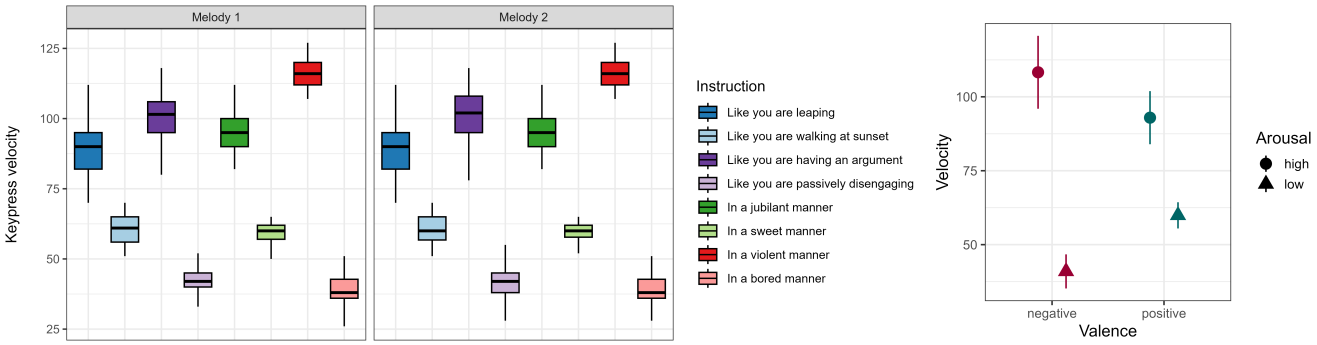
Velocity: GPT-5.5



(a) Velocity for each metaphor (b) Velocity by valence/arousal

Figure 18: The keypress velocity across eight GPT-5.5 simulations.

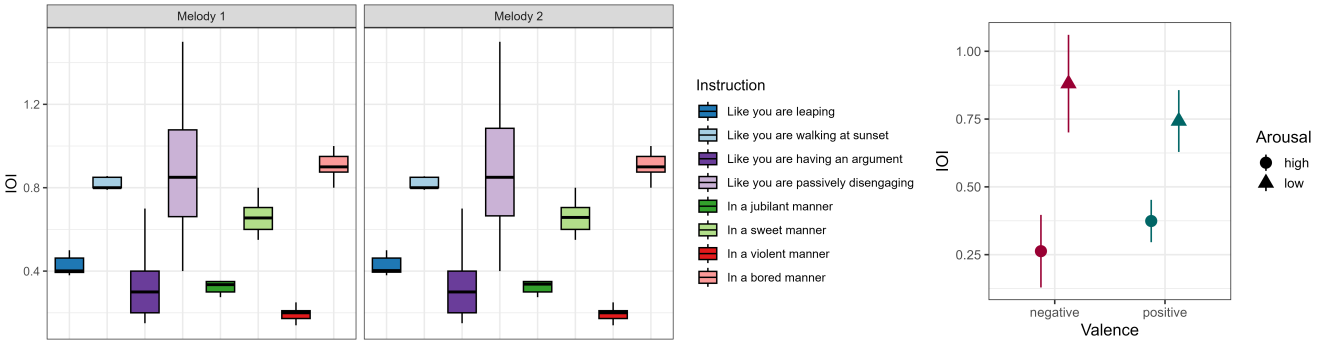
Velocity: Sonnet 4.6



(a) Velocity for each metaphor (b) Velocity by valence/arousal

Figure 19: The keypress velocity across eight Sonnet 4.6 simulations.

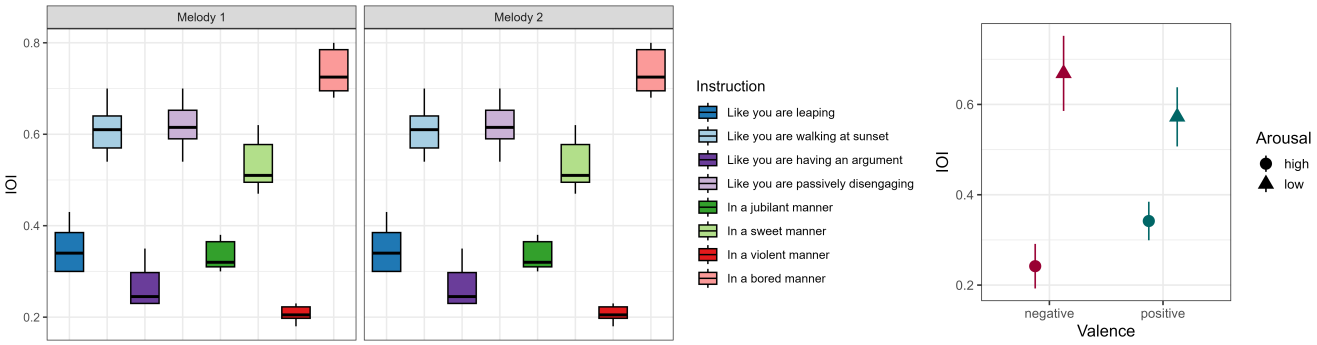
IOI: Gemini 3.1 Pro



(a) IOI for each metaphor (b) IOI by valence/arousal

Figure 20: Inter-onset interval across eight Gemini 3.1 Pro simulations.

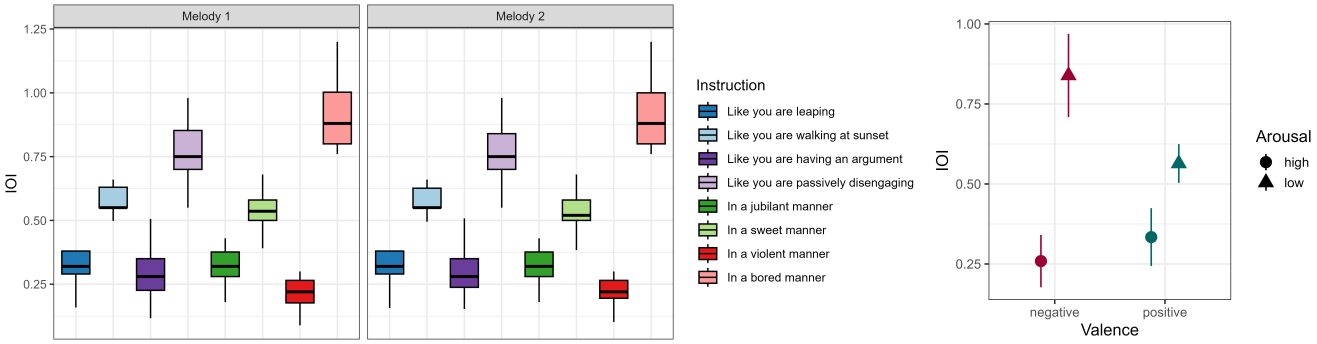
IOI: GPT-5.5



(a) IOI for each metaphor (b) IOI by valence/arousal

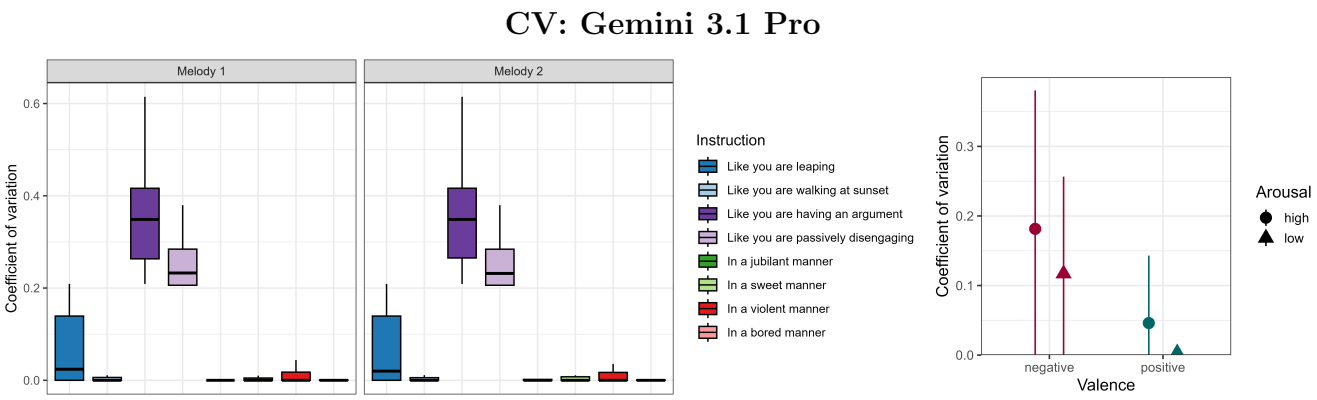
Figure 21: Inter-onset interval across eight GPT-5.5 simulations.

IOI: Sonnet 4.6



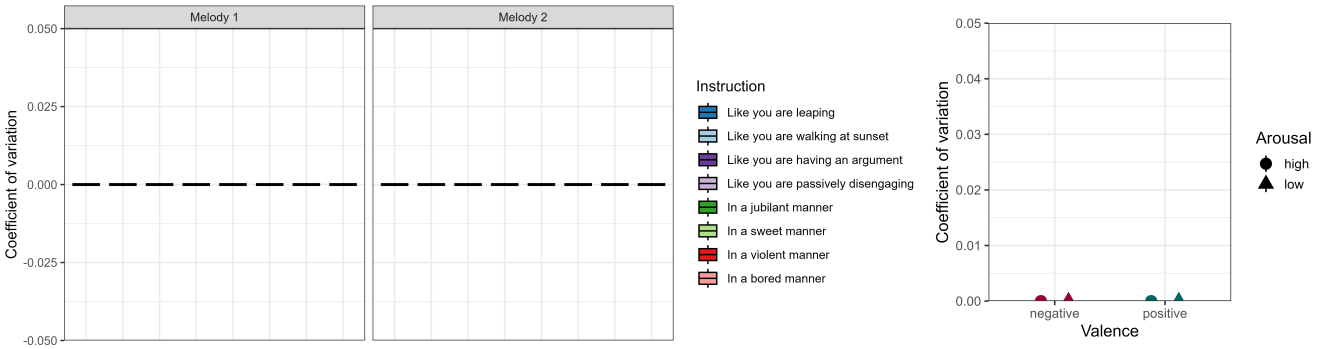
(a) IOI for each metaphor (b) IOI by valence/arousal

Figure 22: Inter-onset interval across eight Sonnet 4.6 simulations.



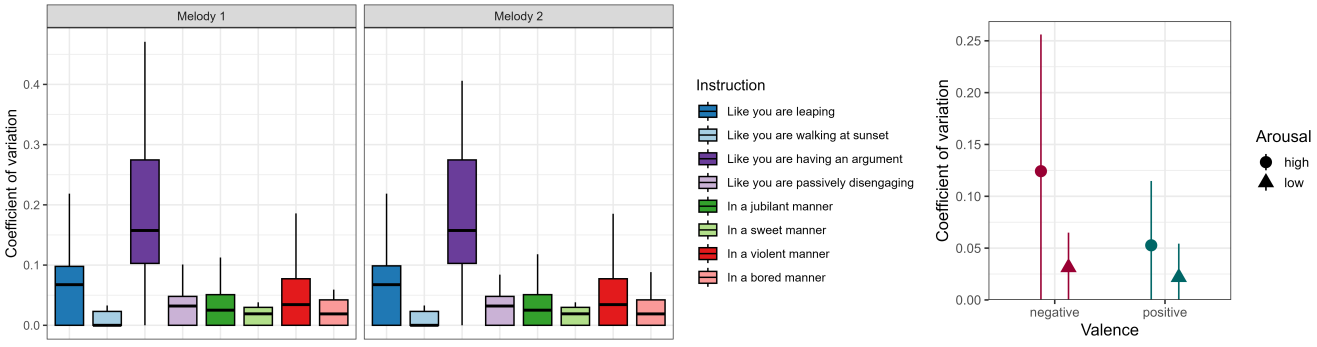
(a) CV for each metaphor (b) CV by valence/arousal

Figure 23: Coefficient of variation (CV) across eight Gemini 3.1 Pro simulations.



(a) CV for each metaphor (b) CV by valence/arousal

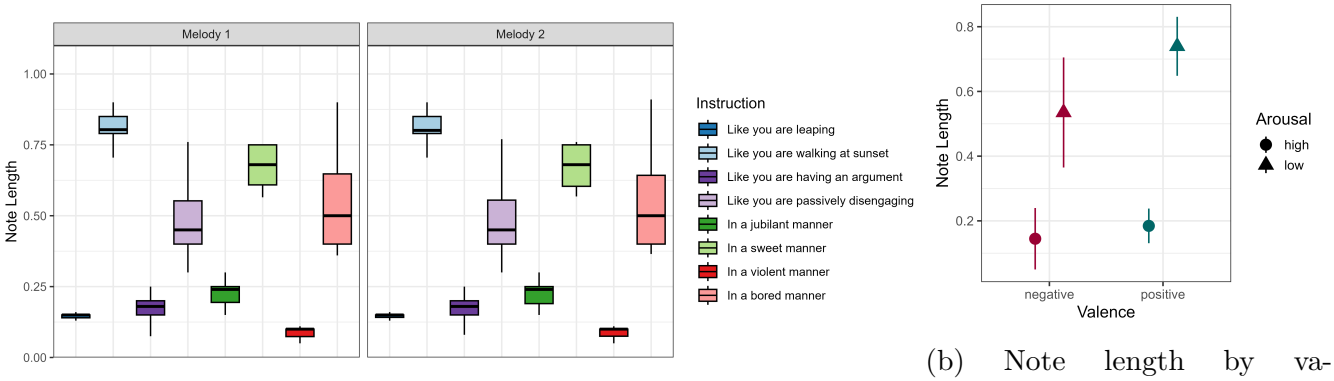
Figure 24: Coefficient of variation (CV) across eight GPT-5.5 simulations.



(a) CV for each metaphor (b) CV by valence/arousal

Figure 25: Coefficient of variation (CV) across eight Sonnet 4.6 simulations.

Note length: Gemini 3.1 Pro

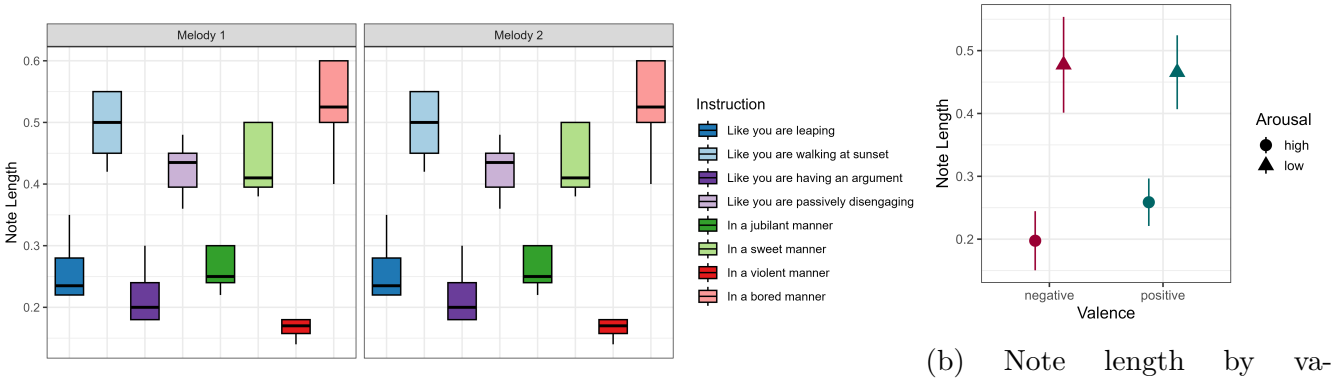


(a) Note length for each metaphor

(b) Note length by valence/arousal

Figure 26: Note length across eight Gemini 3.1 Pro simulations.

Note length: GPT-5.5

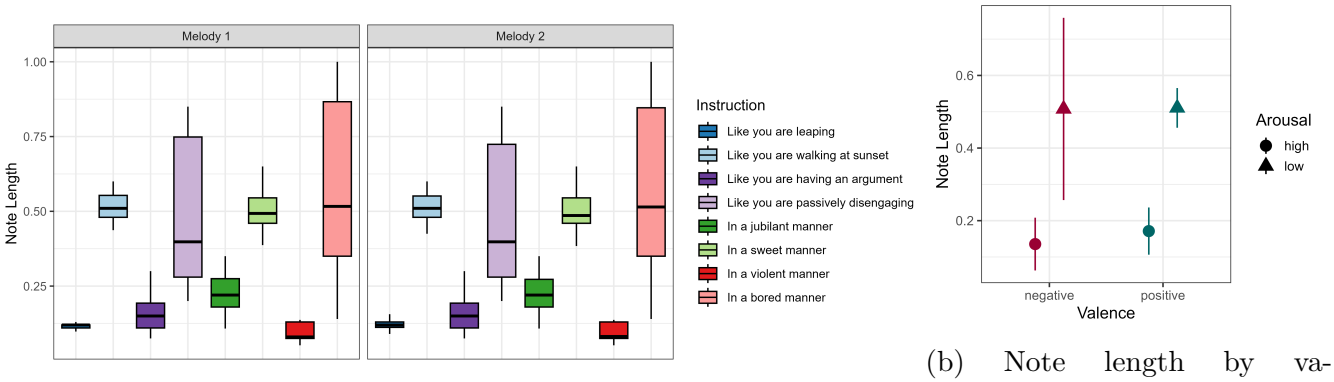


(a) Note length for each metaphor

(b) Note length by valence/arousal

Figure 27: Note length across eight GPT-5.5 simulations.

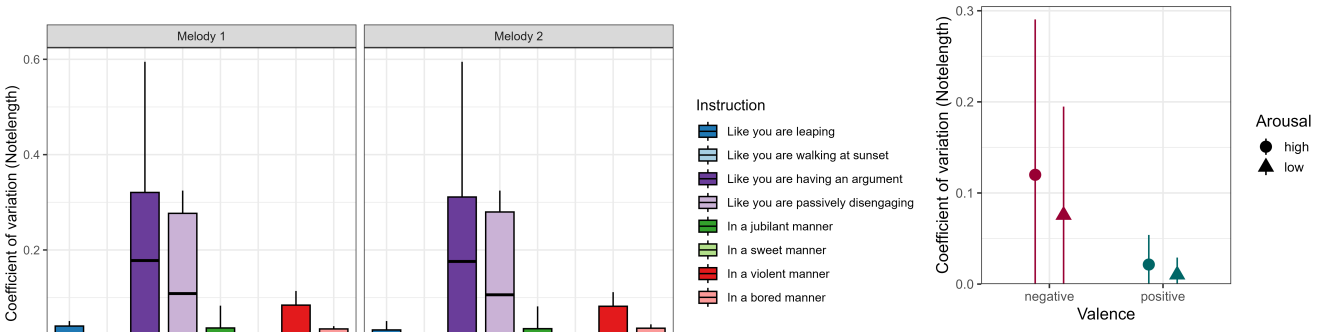
Note length: Sonnet 4.6



(a) Note length for each metaphor

(b) Note length by valence/arousal

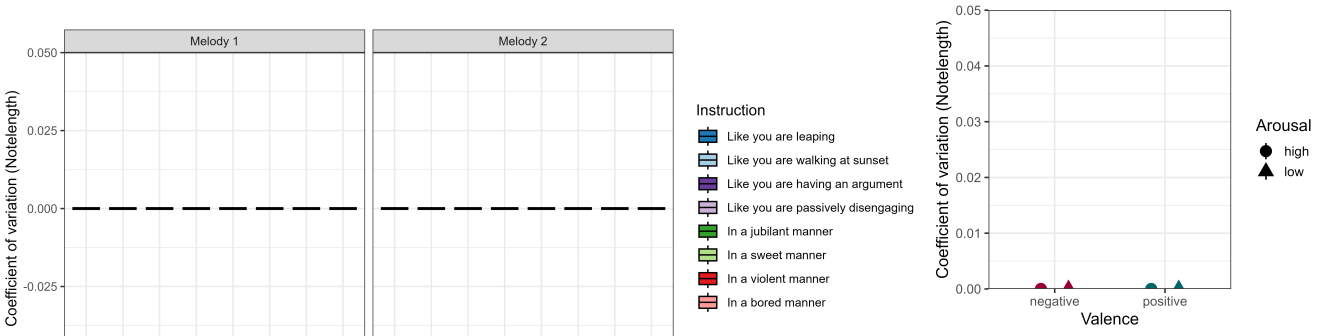
Figure 28: Note length across eight Sonnet 4.6 simulations.



(a) Note length CV for each metaphor

(b) Note length CV by valence/arousal

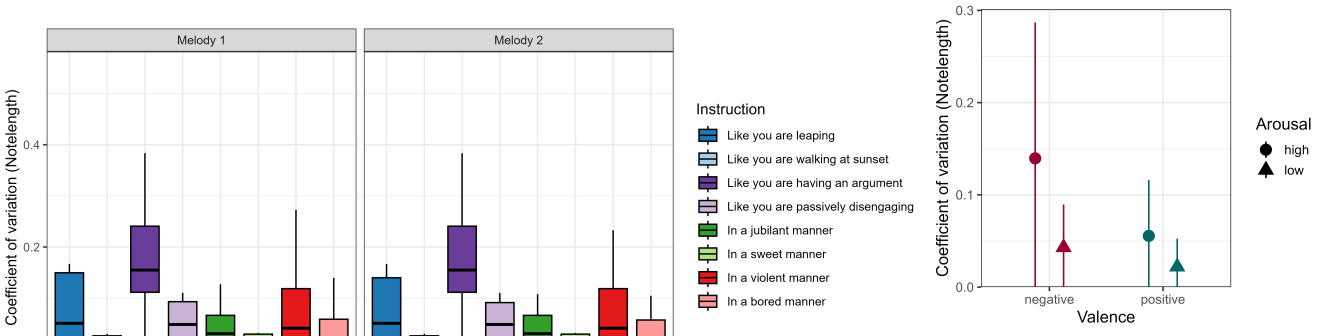
Figure 29: Coefficient of variation of note length across eight Gemini 3.1 Pro simulations.



(a) Note length CV for each metaphor

(b) Note length CV by valence/arousal

Figure 30: Coefficient of variation of note length across eight GPT-5.5 simulations.



(a) Note length CV for each metaphor

(b) Note length CV by valence/arousal

Figure 31: Coefficient of variation of note length across eight Sonnet 4.6 simulations.