Computational Modeling of Artistic Inspiration: A Framework for Predicting Aesthetic Preferences in Poetic Lines Using Linguistic and Stylistic Features

Gaurav Sahu and **Olga Vechtomova**

David R. Cheriton School of Computer Science
University of Waterloo, Canada
{gaurav.sahu, olga.vechtomova}@uwaterloo.ca

Abstract

Artistic inspiration remains one of the least understood aspects of the creative process, yet plays a crucial role in producing works that resonate with audiences. This paper introduces a novel computational framework for modeling individual artistic preferences in poetic lines through key linguistic and stylistic features. Our approach consists of two components: (1) a feature extraction module that quantifies poetic imagery, word energy, abstraction level, emotional valence, and linguistic complexity, and (2) a calibration network that learns to predict what content will inspire specific individuals. To evaluate our framework, we present EvocativeLines, a dataset of poetic lines annotated as either "inspiring" or "not inspiring" across diverse preference profiles. Experiments demonstrate that our framework significantly outperforms state-of-the-art language models, surpassing LLaMA-3-70b by nearly 18 percentage points in accuracy. The framework's design prioritizes interpretability and flexibility, making it adaptable to analyzing various types of artistic preferences across different creative domains and skill levels. By formalizing the measurement of subjective aesthetic responses, our work provides a foundation for computational systems that can support the early stages of the creative process. [1]

Introduction

The creative process, along with the inspiration that drives the creation of art—whether visual, poetic, or musical—remains one of the least understood aspects of human experience. Rubin (2023) describes three distinct stages to a creative process: the Seed phase, where one is entirely open to any and all forms of inspiration; the Experimentation phase, where one tries multiple approaches to allow the nascent ideas or "seeds" to grow and take shape; and the Crafting phase, where the artist refines these ideas into a cohesive work. Our research focuses specifically on the Seed phase, investigating how AI-generated poetic lines can inspire new creative pathways for artists. At this stage, AI-generated lines can serve not as direct inputs into the final artwork but as catalysts that guide artists into a creative mindset.

Assessing creativity in AI-generated content presents significant challenges due to its subjective and individual nature. Traditional evaluation methods such as the Consensual Assessment Technique (CAT) (Amabile 1982) and the Torrance Test of Creative Thinking (TTCT) (Torrance 1966) provide structured frameworks but have notable limitations. CAT relies on expert consensus, yet fails to assess the inspirational potential of outputs or their ability to spark creative development in individual artists. Similarly, TTCT measures divergent thinking but has been criticized for not reflecting the diverse forms that creativity can take (Baer 2011).

In this work, we propose a novel framework that addresses these limitations by: **1)** Identifying and formalizing key linguistic and poetic properties of AI-generated lines that best explain artists' subjective preferences during the Seed phase; **2)** Developing a system that can adapt to the artistic preferences of a wide spectrum of creative individuals; **3)** Focusing on the inspirational quality of poetic lines rather than evaluating them as final artifacts.

A key application of our framework is the ability to personalize creative suggestions for individual artists. By modeling the specific stylistic and linguistic features that resonate with each artist, our system can filter and prioritize AI-generated poetic lines that align with that person's unique artistic sensibilities. This personalization enables more effective creative assistance by providing artists with inspiration that matches their aesthetic preferences, rather than generic suggestions that may not resonate with their creative vision. Furthermore, our framework does not assume a pre-requisite skill level of an artist and can be useful to a wide spectrum of creative individuals for any task that is inherently subjective.

To evaluate our framework, we create *EvocativeLines*, a collection of 3,025 AI-generated poetic lines. Our experiments demonstrate that our approach accurately predicts artistic preferences across diverse preference profiles, significantly outperforming recent large language models like LLaMA-3-70b.

Our contributions include: **1)** A novel framework for measuring various characteristics of AI-generated poetic lines that identifies key linguistic and stylistic features driving subjective preferences during the early creative process; **2)** An interpretable system adaptable to diverse creative individuals regardless of their artistic skill level; **3)** A personalization mechanism that filters and recommends AI-generated poetic lines based on individual artistic preferences; **4)** Experimental validation demonstrating our framework's ability to predict

[1] Project Website:
https://sites.google.com/view/modeling-artistic-preferences

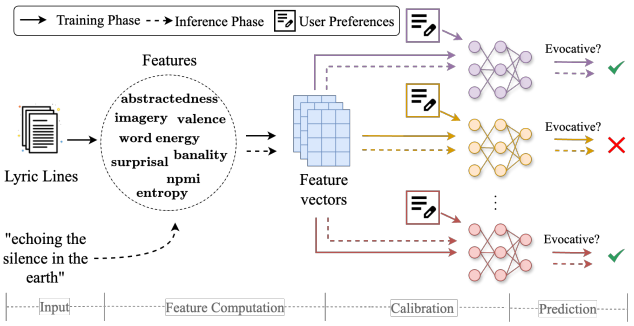

Figure 1: Framework for predicting artistic preferences in poetic lines, consisting of (1) a Feature Computation Module that quantifies linguistic and stylistic characteristics and (2) a Preference Calibration Network that learns individual artistic preferences to predict which lines will be perceived as 'inspiring'.

artistic preferences with significantly higher accuracy than state-of-the-art language models, outperforming LLaMA-3-70b by nearly 18 percentage points; **5)** Insights into the challenges of modeling diverse artistic preferences in highly subjective domains using large language models.

Related Work

Recent advancements in natural language processing (NLP), particularly in the space of LLMs, have facilitated the development of evaluation metrics, such as LLM-Rubric (Hashemi et al. 2024), GPTScore (Fu et al. 2024), and G-Eval (Liu et al. 2023), that are far better aligned with human judgments than traditional NLP metrics like BLEU (Papineni et al. 2002), ROUGE (Lin 2004), or METEOR (Banerjee and Lavie 2005); however, these LLM-based metrics primarily assess general language quality rather than creative attributes. There are also evaluation approaches that measure aspects like diversity and fluency but do not capture the multifaceted nature of creative expression. For instance, Hashimoto, Zhang, and Liang (2019) proposed an evaluation framework balancing diversity and fluency, while Pillutla et al. (2024) introduced MAUVE, an automatic measure of how close generated text is to human text.

Recently, Chakrabarty et al. (2024) extended the TTCT framework to the Torrance Test of Creative Writing (TTCW) to evaluate creativity in outputs from both humans and large language models (LLMs), but it also suffers from TTCT's limitation of being unable to capture the diverse forms creativity can take. In our work, we develop multifaceted metrics that capture various linguistic and stylistic characteristics of generated text and we also investigate how these metrics can predict personal preferences of individuals with diverse artistic skills and aesthetic sensibilities.

Our framework is informed by an aesthetic theory and creative practices. The relationship between aesthetic pleasure and complexity, known as the inverted U-shape or "Wundt" effect (Berlyne 1975), suggests that stimuli of intermediate complexity are most pleasing, an effect observed across domains including music (Gold et al. 2019) and design (Althuizen 2021). This theoretical foundation has practical implications for poetic creativity. Many artists, including David Bowie (Jones 2017) and Kurt Cobain (Cross 2001), have employed techniques like the "cut-up" method (Burroughs 2012) to create unexpected language patterns that inspire creative thoughts. Similarly, literary techniques such as Spanbauer's "burnt tongue" deliberately introduce complexity to force deeper engagement (Palahniuk 2020), an effect supported by evidence that predictable words receive less attention from readers (Lowder et al. 2018).

While the "Wundt" effect theory suggests general patterns in aesthetic responses, preferences for artistic stimuli remain highly individual (Gold et al. 2019; Althuizen 2021). This observation motivates our two-step framework that first measures multiple linguistic and stylistic characteristics of poetic lines, then uses these measurements to learn the specific artistic preferences of individual creators.

The Proposed Framework

Our framework consists of two main components: (1) a feature extraction module that quantifies key linguistic and stylistic characteristics of poetic lines, and (2) a preference calibration network that learns individual artistic preferences based on these features (see Figure 1).

The feature extraction module processes each poetic line to compute a broad set of measurements across multiple dimensions: poetic imagery, word energy, level of abstraction, emotional valence, linguistic surprisal, contextual entropy, word associations (NPMI), and banality. These features capture both the stylistic aspects that make lines evocative and the linguistic properties that contribute to their cognitive processing. Each feature vector preserves both the overall characteristics of the line and its temporal development through word-by-word analysis.

The preference calibration network then takes these feature vectors as input and learns to predict whether a given poetic line will be perceived as "inspiring" by a specific individual. This network is trained on labeled examples of poetic lines that reflect the artistic preferences of the individual. By separating feature extraction from preference modeling, our framework achieves both interpretability and adaptability to diverse artistic tastes. In the following sections, we describe each component of our framework in detail, beginning with the linguistic and poetic features that form the foundation of our approach.

Step 1: Identifying The Key Poetic and Stylistic Characteristics

Our framework quantifies both poetic/stylistic and statistical-linguistic aspects of poetic lines. We measure five creative dimensions: poetic imagery (sensory evocation), word energy (emotional impact), abstraction level (conceptual ambiguity), emotional valence (affective tone), and banality (originality of expression). Complementing these are statistical measures: linguistic surprisal (word unexpectedness), contextual entropy (prediction uncertainty), and normalized pointwise mutual information (word associations).

Our feature selection was guided by literary theory and empirical observation. While many poetic devices act together to create desired effects, a poem is foremost intended to evoke emotion in the reader and heighten perceptual awareness (Dunnigan 2014). Rather than capturing all poetic devices—which would be computationally intractable—we identified fundamental characteristics based on their theoretical importance in achieving these core poetic functions. We complement stylistic features with statistical ones derived from information theory to quantify linguistic patterns influencing cognitive processing. The following sections detail our approach to measuring each characteristic.

Poetic Imagery

Poetic Imagery is a stylistic device using references to physical objects that appeal to readers' senses, evoking emotions and making abstract ideas tangible (Pound 1913; Brooks and Warren 1976; Kao and Jurafsky 2012).To measure this feature, we prompt a LLaMA-3-70b model to output a rating between 1-5 [2]. Since poetic imagery is a well-established stylistic device in English, we perform prompting in a zero-shot fashion. This further avoids the induction of preference bias in the LLM outputs.

We compute the score for a sentence S using weighted summation:

$$score_{img}(S) = \sum_{i=1}^{5} p(r_i) \times r_i, \tag{1}$$

where $p(r_i)$ is the probability of the LLM assigning rating r_i. This provides more fine-grained scores and takes into account the model's confidence rather than directly using output tokens as scores. The final imagery feature vector combines scores of all subsequences to capture temporal development:

$$\mathbf{S_{imagery}} = \langle score_{img}(\mathcal{S}_t) \rangle_{t=1}^{T}, \tag{2}$$

where $\mathcal{S}_t = (w_1, w_2, \cdots, w_t)$ represents a subsequence of the first t words.

This effectively captures poetic imagery differences: "there's a little red to the sea" scores 4.01, while "this is all" only reaches 1.27.

Word Energy

Word Energy is a composite feature encompassing symbolism and diction, which contribute to a poem's emotional impact. Symbolism uses objects to evoke abstract concepts and emotions (Cassirer 1946) (e.g., "chains" representing entrapment), while diction refers to word choice for achieving specific effects (Brooks and Warren 1976) (e.g., "beseeching" versus "searching").

Similar to imagery, we prompt a LLaMA-3-70b model to rate poetic lines between 1-5 in a zero-shot fashion to obtain the score $\mathbf{S_{energy}}$. Since word energy is a composite feature, we include three example words that exemplify this concept (e.g., 'beseeching', 'chains', 'burning'). A higher rating indicates higher word energy.

[2]Examples of prompts are listed on the project website.

The final feature vector takes the same form as the imagery vector, except ratings reflect word energy rather than imagery. For instance, "all the words will drown" scores 4.11, while "where i know about you" receives only 1.92.

Level of Abstraction

We measure abstraction or ambiguity in poetic lines, as higher abstraction allows multiple interpretations and can be more engaging. The *Linguistic Category Model (LCM)* (Semin and Fiedler 1988; Johnson-Grey et al. 2020) provides a framework for measuring abstraction by assessing four word categories: Descriptive Action Verbs (DAVs), Interpretative Action Verbs (IAVs), State Verbs (SVs), and Adjectives (ADJs).

This measure captures stylistic devices like diction and represents the balance between abstract and concrete language, essential for defining tone, mood, and imagery vividness. It also indirectly indicates figurative language through abstract expressions. We compute the abstraction feature vector as:

$$\mathbf{S_{abs}} = \langle score_{abs}(\mathcal{S}_t) \rangle_{t=1}^{T} \tag{3}$$

where $\mathcal{S}_t = (w_1, w_2, \cdots, w_t)$ denotes a subsequence with first t words, and:

$$score_{abs} = \frac{D + (2 \times I) + (3 \times S) + (4 \times A)}{D + I + S + A} \tag{4}$$

Here, D, I, S and A represent counts of the four word categories. Weights reflect theorized abstraction levels from a study of 40,000 English words (Brysbaert, Warriner, and Kuperman 2014), with concrete verbs (DAVs) weighted lowest (1.0) and abstract adjectives highest (4.0). We use LLaMA-3-70b to obtain these counts.

Our scores distinguish between definitive phrases with lower abstraction ("rearranging this stage") and more unresolved ones ("the greatest began to deny"). Our experiments indicate abstraction level is a prominent feature in identifying potentially evocative lines.

Banality

Banality refers to common or clichéd expressions lacking originality or depth. It reflects cultural and temporal context—artists may deliberately use clichés to resonate with broader audiences or evoke certain eras. This makes banality crucial for studying diverse artistic preferences.

Similar to imagery and word energy, we prompt a LLaMA-3-70b model in zero-shot fashion to rate expressions from 1-5 for banality. Higher ratings indicate more banal expressions, while lower ratings reflect uniqueness or originality.

The banality feature vector is constructed as:

$$\mathbf{S_{banality}} = \langle score_{bnl}(\mathcal{S}_t) \rangle_{t=1}^{T}, \tag{5}$$

where $\mathcal{S}_t = (w_1, w_2, \cdots, w_t)$ denotes a subsequence with the first t words, and $score_{bnl}(\mathcal{S}_t)$ is computed similarly to imagery scores but reflects banality levels.

For example, "take the test" achieves a high banality score of 3.18, while "breathe back the world" scores only 0.98, indicating greater originality.

Valence

Valence reflects emotions expressed in poetic lines and is crucial for assessing an artist's preferred emotional tone (Scherer 1984; Charland 2005). Artists might favor lines that blend positive and negative connotations ("the greatest began to deny") or purely positive sentiments ("tranquility in the lovers").

To measure valence, we build a computational emotion model based on 29 fine-grained emotion categories derived from the GoEmotions dataset (Demszky et al. 2020), plus "nostalgia" to capture lyrical emotional nuances. We employ a LLaMA-3-70b prompting-based classifier to obtain probabilities for the top 5 emotional categories for each line.

Let S represent the input sentence and E_k the set of top-k emotions predicted by the classifier (with $k = 5$). The probability vector is constructed as: $\mathbf{p}(S) = \langle \tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_m \rangle$, where m is the total emotion categories (29) and:

$$\tilde{p}_e = \begin{cases} p_e, & \text{if } e \in E_k \\ p_{rem}, & \text{if } e \notin E_k \end{cases} \tag{6}$$

with $p_{rem} = \frac{1-\sum_{e \in E_k} p_e}{m-k}$ representing remaining probability mass distributed evenly across other categories. The complete valence vector for a poetic line $S = (w_1, w_2, \ldots, w_T)$ is given by:

$$\mathbf{S_{vfull}} = \langle \mathbf{p}(\mathcal{S}_t) \rangle_{t=1}^{T}, \tag{7}$$

where $\mathcal{S}_t = (w_1, w_2, \cdots, w_t)$ denotes a subsequence capturing temporal development.

We also construct a uni-dimensional variant by grouping emotions into positive, neutral, and negative sets (E_+, E_0, E_-), and collapsing the probability vector to a score $\tilde{p} \in [-1, 1]$:

$$\mathbf{p_{bin}}(S) = \tilde{p}_+ = \sum \mathbb{I}_e \cdot \tilde{p}_e, \tag{8}$$

where $\mathbb{I}_e$ is defined as:

$$\mathbb{I}_e = \begin{cases} +1, & \text{if } e \in E_+ \\ 0, & \text{if } e \in E_0 \\ -1, & \text{if } e \in E_- \end{cases} \tag{9}$$

yielding the final feature vector:

$$\mathbf{S_{vbin}} = \langle \mathbf{p_{bin}}(\mathcal{S}_t) \rangle_{t=1}^{T} \tag{10}$$

Our preliminary results showed that features from the LLaMA-based classifier have higher predictive power than a RoBERTa classifier (Liu 2019) fine-tuned on GoEmotions.

Surprisal

Surprisal (Shannon 1948) indicates a word's predictability in a sentence, measured as the negative log probability of the word given its context:

$$s(w_t) = -\log_2 P(w_t | \mathbf{w}_{<t}), \tag{11}$$

where $s(w_t)$ denotes the surprisal value for the t-th word.

To compute sentence-level surprisal, we compose a vector of word-level surprisals, appended with their average:

$$\mathbf{S_{surprisal}} = \langle s(w_t) \rangle_{t=1}^{T} \oplus \langle \overline{s} \rangle, \tag{12}$$

where T is the number of words, $\overline{s} = \sum s(w_t)/T$ is the mean surprisal, and $\oplus$ denotes concatenation.

Since we lack access to the true probability distribution P, we use GPT-2 (Radford et al. 2019)[3], a powerful autoregressive language model, to approximate the log probabilities.

This feature distinguishes between sentences like "a knowledge of the world," with lower surprisal ($s_{avg} = 10.12$) indicating higher predictability, and "a just dream with the fire," with higher surprisal ($s_{avg} = 11.21$) indicating greater novelty potential (note these values are on a log scale, so the difference is exponential).

Contextual Entropy

Contextual Entropy (Shannon 1948) of a word is defined as the expected value of its Surprisal:

$$h(w_t) = \sum_{w_i \in V} p(w_t | \mathbf{w}_{<t}) \log_2 P(w_t | \mathbf{w}_{<t}), \tag{13}$$

where $h(w_t)$ denotes the contextual entropy for the t-th word, and V represents the vocabulary of possible next words.

Similar to surprisal, we use GPT-2 to compute probabilities and construct the sentence-level entropy feature vector by combining word-level entropy values with their mean:

$$\mathbf{S_{entropy}} = \langle h(w_t) \rangle_{t=1}^{T} \oplus \langle \overline{h} \rangle, \tag{14}$$

where T is the number of words and $\overline{h} = \sum h(w_t)/T$ is the mean entropy. Contextual entropy helps gauge unpredictability/novelty in sentences. For example, "i remember you" has lower entropy ($h_{avg} = 6.21$) compared to "pure crime of the lost" ($h_{avg} = 9.53$). Entropy also correlates positively with reading times (Lowder et al. 2018), indicating its ability to capture sentence complexity (higher entropy $\rightarrow$ higher complexity).

Normalized Pointwise Mutual Information (NPMI)

NPMI measures word associations by comparing co-occurrence probabilities with individual occurrence probabilities in a corpus (Bouma 2009). Formally, the NPMI of words x and y is defined as:

$$N(x, y) = \left(\ln \frac{p(x, y)}{p(x)p(y)} \right) / - \ln p(x, y) \tag{15}$$

Here, $p(x) = count(x)/N$, where $count(x)$ is the word's occurrence count and N is the total word count. NPMI values range between $[-1, 1]$, with higher values indicating stronger associations.

For example, ("see", "you") has a high NPMI value (0.47) while ("never", "the") has a low value (-0.11), as "see" and "you" frequently co-occur, whereas "never" typically precedes verbs rather than determiners.

We compute two NPMI variants: unidirectional (N_{uni}), considering co-occurrences only when y follows x, and bidirectional (N_{bi}), also counting when y precedes x. The final feature vectors take the form:

[3] specifically, we use the 774M parameter `gpt2-large` model from Huggingface: https://huggingface.co/openai-community/gpt2-large

$\mathbf{S_{npmi,uni}} = \langle N_{uni}(w_t, w_{t+1}) \rangle_{t=1}^{T-1} \oplus \langle \overline{N}_{uni} \rangle$ (similar for $\mathbf{S_{npmi,ubi}}$), where T is the sentence length, $\overline{N}_{uni} = \sum N_{uni}(w_t, w_{t+1})/T$ is the mean unidirectional NPMI, and $\overline{N}_{bi} = \sum N_{bi}(w_t, w_{t-1})/T$ is the mean bidirectional NPMI.

Step 2: Training the Calibration Network

Artistic preferences vary significantly across individuals—a poetic line with identical stylistic and linguistic properties might inspire one person while leaving another unmoved. This subjective nature of creative inspiration necessitates a personalized approach that adapts to individual tastes.

To address this challenge, we develop a calibration network that learns to map the extracted features to an individual's unique artistic preferences. This personalization layer is critical for accurately predicting whether a specific person will find a given poetic line inspiring.

We formulate this as a supervised classification problem: given the linguistic and stylistic features of a poetic line, the network predicts whether the line will be inspiring to a specific individual. Formally, consider a dataset of artistic preferences $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where $x_i \in \mathcal{X}$ represents the feature vector of the i-th poetic line and $y_i \in \mathcal{Y}$ indicates whether the artist found that line inspiring.

The calibration network learns a function $f_\theta : \mathcal{X} \to \mathcal{Y}$ that maps each feature vector x_i to a probability distribution over classes $\hat{y} = f_\theta(x_i) = P(y|x_i; \theta)$. For our binary classification setup, $\mathcal{Y} = \{\text{inspiring}, \text{not inspiring}\}$.

This approach offers several advantages over alternative methods. Unlike end-to-end models that might directly map raw text to preferences, our feature-based approach provides interpretability by revealing which specific linguistic and stylistic elements resonate with each individual. Additionally, by avoiding direct use of sentence embeddings, we focus explicitly on creative attributes rather than general semantic properties, which often fail to capture the nuanced elements that make lyrics inspirational.

The calibration network can be implemented using various architectures, from traditional machine learning models to deep neural networks, depending on the dataset size and complexity of preference patterns. Once trained, this personalized layer enables our framework to filter and prioritize lyrical content according to each artist's unique creative sensibilities.

Dataset

Data Collection and Annotation. We collected 3,025 poetic lines from the LyricJam platform, a publicly available research system that generates poetic lines with the goal of assisting creative individuals in finding novel lyrical and poetic ideas for use in the Seed phase of the creative process (Vechtomova, Sahu, and Kumar 2021) [4]. All lines were preserved in their original lowercase format without additional preprocessing. We refer to this collection as the *EvocativeLines* dataset.

For the initial annotation phase, the first and last authors independently labeled each line as either "inspiring" or "not inspiring." In this context, "inspiring" indicates that a line created an emotional impact and stimulated creative ideation. Both annotators actively engage in creative writing (poetry and fiction), providing them with relevant expertise for this evaluation task. The complete annotation process spanned approximately 12 hours over multiple weeks.

[4] https://lyricjam.ai

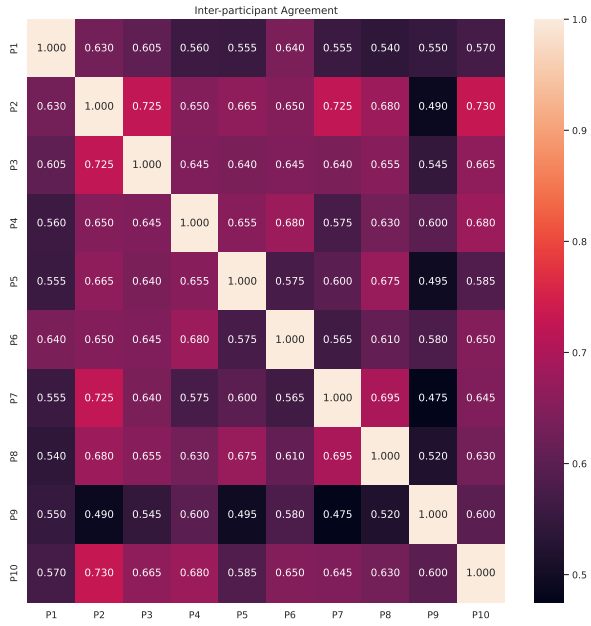

Figure 2: Degree of agreement among participants on the EvocativeLines *(small)* dataset. Lower agreement (darker tones) demonstrates the diversity of subjective preferences captured across our pool of creative individuals.

Diverse Preference Profiles. To evaluate our framework's ability to model diverse artistic preferences, we recruited eight additional participants, each actively engaged in creative activities such as poetry writing or musical performance. These participants rated a subset of 200 poetic lines, the *EvocativeLines (small)* dataset, on a scale of 1-10 based on their inspirational quality.

The subset was constructed by randomly selecting 100 "inspiring" and 100 "not inspiring" lines from the original dataset, balanced across different length categories (short, medium, and long). This sampling approach kept the annotation task manageable (30-45 minutes per participant), minimizing decision fatigue while still capturing a broad spectrum of creative preferences. We applied min-max scaling to normalize the ratings and convert them to binary "inspiring"/"not inspiring" judgments.

We denote our complete set of participants as $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_{10}\}$, where $\mathcal{P}_1$ and $\mathcal{P}_2$ represent the first and last authors, and $\mathcal{P}_{3-10}$ represent the additional participants. Each participant's set of labels constitutes a unique preference profile.

Inter-participant Agreement Analysis. Figure 2 visualizes the inter-participant agreement on the EvocativeLines

(small) dataset. The agreement score represents the percentage of lines that pairs of participants both found either inspiring or not inspiring. We observed substantial variation in agreement, ranging from 47.5% to 72.5%, with a mean agreement of 65.3% (±3.1%). This indicates that participants disagreed on approximately 70 out of 200 lines on average.

This level of disagreement aligns with previous research on the perception of aesthetic stimuli (Gold et al. 2019) and confirms that our dataset effectively captures the subjective and individualized nature of artistic preferences.

Alternative Generation Sources. We also explored alternative AI generation platforms, including ChatGPT and Claude Sonnet. However, less than 1% of lines from these sources were judged as evocative by our participants, providing insufficient data for training our calibration network. Figure S.2 on our project website showcases examples with comparative scores for poetic and stylistic characteristics across different sources. Notably, we observe that these systems tend to add superficial flourishes to lines to make them "poetic." Such lines score high in imagery and energy; however, they also score high on banality (e.g. "stars watch over empty fields," and "dreams scatter like fallen petals" that sound poetic but use well-worn expressions).

Experiments

Step 1: Computing Features. To test our framework on the EvocativeLines dataset, we first compute all the features described in Step 1 of the proposed framework. Specifically, we combine all the features *along the time dimension* to obtain the following overall feature vector for a given poetic line: $\mathbf{S_{all}} = [\mathbf{S_{imagery}}|\cdots|\mathbf{S_{banality}}]$ where $|$ denotes the concatenation operation along the time dimension, T is the number of time steps (= number of words in the sentence), and $\mathbf{S_{all}} \in \mathbb{R}^{T\times n}$ with $n = (8T + m)$ when using $\mathbf{S_{vfull}}$ and $n = (8T + 1)$ when using $\mathbf{S_{vbin}}$.

Step 2: Training the Calibration Network. We create a 80-20 split of training-testing in the dataset and use the features in the training set to train the calibration network. We use stratified sampling when creating the splits to ensure that the distribution of labels is preserved across the two dataset parts. We perform 5-fold cross-validation for tuning the hyperparameters of the calibration network. When training the calibration network on the EvocativeLines *(small)* dataset, we create a 75-25 split of training-testing and then do a 5-fold cross-validation.

Choice of calibration network. We explore multiple deep-learning (DL) and machine-learning (ML) based architectures for the calibration network. Namely, we try **a) LSTM+Attn** ($\mathbf{S_{all}}$): a bi-directional classification network with Long Short Term Memory (LSTM) cells (Hochreiter and Schmidhuber 1997) coupled with an attention mechanism (Bahdanau 2014). We use an LSTM network as all of our features are temporal, and LSTMs are adept at modeling long-range temporal dependencies. We also perform ablation studies where we train the classifier on a subset of $\mathbf{S_{all}}$ to test the effect of the individual features; **b) XGBoost** (Chen and Guestrin 2016): We train a standard XGBoost model on our features; however, we *flatten* $\mathbf{S_{all}}$ as XGBoost does not inherently handle sequential data with a timestep dimension. Specifically, we train XGBoost (and all the other ML-based classifiers) on the following flattened input feature: $\mathbf{S_{flat}} = [\dot{\mathbf{S}}_{\mathbf{imagery}} \oplus \cdots \oplus \dot{\mathbf{S}}_{\mathbf{banality}}]$, where $\dot{\mathbf{S}}_{(\cdot)}$ is the flattened version of $\mathbf{S}_{(\cdot)}$ and $\oplus$ is the concatenation operation. $\mathbf{S_{flat}} \in \mathbb{R}^{Tn}$, where $n = 8T + m$; **c) Other ML-based Classifiers.** In addition to XGBoost, we train multiple other ML-based classifiers on $\mathbf{S_{flat}}$. Specifically, we train Logistic Regression **(LR)**, Support Vector Machine **(SVM)**, Decision Trees **(DT)**, and Random Forest **(RF)** models (Nelder and Wedderburn 1972; Cortes and Vapnik 1995; Quinlan 1986; Breiman 2001).

Other Frameworks. We compare our framework with multiple strong baselines. Specifically, we compare the following: **a) LSTM+Attn (SBERT):** we train an LSTM+Attn classifier that maps the Sentence-BERT (Reimers and Gurevych 2019) embeddings of poetic lines to the final rating. SBERT is a popular text embedding model that modifies the BERT architecture with siamese (Bromley et al. 1993) and triplet networks (Schroff, Kalenichenko, and Philbin 2015) to obtain semantically meaningful sentence representations. Note that in this scenario, there will be only one "time step"; **b) BERT** (Devlin et al. 2019): a state-of-the-art transformer-based language model pre-trained on large textual corpora. It is widely adopted for various natural language processing tasks, such as classification, question answering, and named entity recognition. We fine-tune a BERT-base model, which has 110M parameters, on the EvocativeLines dataset, which serves as a strong baseline comparing the performance of mapping the sentence embeddings of poetic lines directly with the final prediction; **c) LLaMA-3:** a prompting-based k-shot LLaMA classifier, where, for a given test line, we prompt a LLaMA-3-70b model with k "inspiring" and k "not inspiring" examples from the dataset and ask it to classify a test sentence. We use cosine similarity to find the k examples that are (semantically) closest to the test sentence in the SBERT embedding space. We find that this approach of selecting examples performs better than random sampling.

Evaluation of Different Frameworks. We adapt each framework on the EvocativeLines dataset and use test accuracy and area under the curve (AUC) scores as primary indicators of the predictive power of different models. We report these results in Table 1. Further, we test our proposed framework on the EvocativeLines *(small)* dataset with 10 diverse sets of preference profiles, and report results in Table 2.

Implementation Details. We implement the LSTM classifiers in PyTorch (Paszke et al. 2019) and train them for 500 epochs with early stopping (we stop a training run if the validation performance does not improve for 10 consecutive epochs). We perform grid search to find optimal values over the parameters listed in Figure S.9 on the project website, and find a *one*-layer *bi*-directional LSTM classifier trained

Method	$\mathcal{P}_1$		$\mathcal{P}_2$	
	Accuracy	**AUC**	**Accuracy**	**AUC**
Majority baseline	*51.5*	*50.0*	*53.0*	*50.0*
LSTM+Attn ($\mathbf{S_{all}}$)	**79.5**$^{(0.4)}$	**79.7**$^{(0.1)}$	**78.3**$^{(0.3)}$	**80.1**$^{(0.3)}$
LSTM+Attn (SBERT)	71.6$^{(0.3)}$	71.7$^{(0.1)}$	69.9$^{(0.2)}$	70.1$^{(0.3)}$
BERT	77.6$^{(2.1)}$	78.8$^{(1.7)}$	75.2$^{(2.3)}$	76.1$^{(1.1)}$
RF	80.5$^{(1.7)}$	80.6$^{(1.7)}$	72.4$^{(2.0)}$	73.0$^{(1.2)}$
SVM	70.5$^{(1.2)}$	70.5$^{(1.0)}$	67.3$^{(2.3)}$	63.1$^{(1.1)}$
LR	71.1$^{(1.9)}$	71.1$^{(1.8)}$	64.3$^{(1.9)}$	62.0$^{(1.4)}$
DT	84.2$^{(2.5)}$	84.3$^{(2.3)}$	84.1$^{(1.0)}$	80.0$^{(1.7)}$
XGBoost	**92.2**$^{(1.7)}$	**92.2**$^{(1.6)}$	**87.32**$^{(1.4)}$	**83.9**$^{(1.2)}$
LLaMA (200-shot)	73.3$^{(1.7)}$	73.3$^{(1.6)}$	62.3$^{(3.6)}$	65.1$^{(1.6)}$
LLaMA (300-shot)	73.9$^{(2.1)}$	74.2$^{(1.8)}$	64.1$^{(1.2)}$	65.8$^{(1.4)}$
LLaMA (450-shot)	**74.8**$^{(3.1)}$	**74.8**$^{(3.0)}$	**69.5**$^{(3.5)}$	**70.5**$^{(1.6)}$

Table 1: Comparison of different frameworks on *Evocative-Lines*. The first section shows the performance of DL-based methods, the second section shows the performance of ML-based methods, and the last section shows the performance of LLM-based methods.

User	Majority		LSTM+Attn		XGBoost	
	Accuracy	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
EvocativeLines (small)						
$\mathcal{P}_1$	50.0	50.0	75.6$^{(1.2)}$	64.9$^{(4.0)}$	70.7$^{(5.3)}$	70.2$^{(5.3)}$
$\mathcal{P}_2$	70.5	50.0	80.5$^{(1.2)}$	62.3$^{(6.1)}$	69.3$^{(9.5)}$	57.1$^{(8.4)}$
$\mathcal{P}_3$	58.5	50.0	**94.1**$^{(1.2)}$	**92.3**$^{(2.1)}$	63.4$^{(11.1)}$	62.6$^{(11.7)}$
$\mathcal{P}_4$	60.5	50.0	68.7$^{(3.1)}$	61.4$^{(4.5)}$	59.5$^{(4.5)}$	56.4$^{(6.6)}$
$\mathcal{P}_5$	58.0	50.0	80.4$^{(2.5)}$	84.8$^{(5.1)}$	71.7$^{(6.7)}$	**70.5**$^{(6.6)}$
$\mathcal{P}_6$	64.5	50.0	75.6$^{(3.1)}$	75.4$^{(2.8)}$	68.8$^{(3.2)}$	62.8$^{(3.0)}$
$\mathcal{P}_7$	58.0	50.0	75.6$^{(4.3)}$	82.3$^{(4.2)}$	61.0$^{(4.4)}$	60.3$^{(4.2)}$
$\mathcal{P}_8$	62.5	50.0	80.5$^{(5.4)}$	78.2$^{(5.1)}$	**72.2**$^{(5.0)}$	68.5$^{(7.1)}$
$\mathcal{P}_9$	70.0	50.0	76.5$^{(4.7)}$	72.4$^{(3.6)}$	70.7$^{(4.9)}$	57.3$^{(5.1)}$
$\mathcal{P}_{10}$	58.5	50.0	82.3$^{(2.3)}$	84.2$^{(3.5)}$	62.4$^{(5.9)}$	60.7$^{(5.0)}$
Avg.	61.1	50.0	**79.0**$^{(6.3)}$	**75.8**$^{(9.9)}$	67.0$^{(4.6)}$	62.6$^{(5.1)}$

Table 2: Performance of the proposed framework with LSTM+Attn and XGBoost calibration networks on the *EvocativeLines (small)* dataset.

with a learning rate of 0.0001 on $\mathbf{S_{all}}$ input features performs best.

We use the `sentence-transformers`[5] python package to obtain SBERT embeddings of sentences and use Huggingface's `transformers`[6] library to implement the finetuning of BERT classifier. Specifically, we finetune the BERT classifier for 500 steps with a weight decay of 0.01. We use the `sklearn` python package (Pedregosa et al. 2011) to implement all the ML-based classifiers. We use the default hyperparameters for all the models. Finally, we use a batch size of 32 for all the models.

For the LLaMA-3 classifier, we use the categorization of "inspiring" and "not inspiring" from the annotations from $\mathcal{P}_1$ and $\mathcal{P}_2$ (we randomly select which participant to choose for every test sentence).

Finally, we perform 5-fold cross-validation (except for the LLaMA-3 classifier), repeat all our experiments for 5 random seeds, and report the mean and std values in our results tables.

[5]`https://www.sbert.net/`

[6]`https://huggingface.co/docs/transformers/`

Results and Analysis

Proposed Linguistic and Poetic Features Have High Predictive Power

Our experimental results demonstrate that specialized feature engineering significantly outperforms even the largest language models in predicting artistic preferences. From Table 1, we observe that our framework with ML-based and DL-based calibration networks consistently surpasses a 450-shot LLaMA-3-70b classifier across participant profiles.

Specifically, our XGBoost model—containing just 163 trainable parameters—achieved test accuracies of 92.2% for $\mathcal{P}_1$ and 87.3% for $\mathcal{P}_2$. This substantially outperforms the 450-shot LLaMA-3 classifier with 70 billion parameters, which only reached accuracies of 74.8% and 69.8% for the same participants. Note that 450 examples represented the maximum number of positive and negative instances we could include in the LLaMA-3-70b prompt due to context length limitations.

Among deep learning approaches, our LSTM+Attention model (approximately 110k parameters) outperforms both a fine-tuned BERT model (110M parameters) and the LSTM+Attention model trained on SBERT embeddings. This further validates the superiority of our proposed features over general-purpose text representations.

The higher performance of our targeted feature extraction approach compared to using LLMs directly reveals an important insight: while we can leverage LLMs to compute established linguistic and stylistic characteristics, these models struggle to capture the nuanced differences between what individual users find "inspiring" versus "not inspiring." This highlights how our framework's focus on specific linguistic and poetic features is more effective for modeling personal artistic preferences than using much larger general-purpose models.

The Proposed Framework Adapts Robustly to Diverse User Preferences

Our framework demonstrates strong adaptability across varied artistic sensibilities, as shown in Table 2. When tested on the *EvocativeLines (small)* dataset containing 10 distinct preference profiles, both calibration networks effectively learned individual preferences despite their differences.

The LSTM+Attention variant achieved an average test accuracy of 79.0% ($\pm$6.3%) across all 10 participants, while the XGBoost variant reached 67.0% ($\pm$4.6%). This represents a performance gap of approximately 25 percentage points for XGBoost compared to its results on the full dataset. This decline likely stems from XGBoost's tendency toward overfitting in data-scarce scenarios, despite its general flexibility.

Overall, the results highlight the robustness of our framework given the high degree of variation across preference profiles. We can, therefore, conclude that our framework can successfully model and predict artistic preferences across diverse individuals, even with limited training data, by selecting an appropriate calibration network.

Feature	$\mathcal{P}_1$		$\mathcal{P}_2$		$\mathcal{P}_3$		$\mathcal{P}_6$		$\mathcal{P}_7$	
	coef	p-value	coef	p-value	coef	p-value	coef	p-value	coef	p-value
$\mathbf{S}_{\text{surprisal}}$	-0.4149	0.107	0.288	0.404	0.4351	0.233	0.416	0.242	-0.2176	0.518
$\mathbf{S}_{\text{entropy}}$	-0.8377	0.034	0.0534	0.889	0.02	0.958	-0.0632	0.873	-0.0817	0.822
$\mathbf{S}_{\text{npmi,uni}}$	7.2427	0.000	9.7483	0.068	59.1835	0.003	75.9894	0.018	21.0617	0.152
$\mathbf{S}_{\text{npmi,bi}}$	-6.2588	0.002	-9.8577	0.552	-55.0718	0.007	-73.9488	0.024	-19.2738	0.224
$\mathbf{S}_{\text{abs}}$	0.5402	0.298	0.6791	0.047	0.7423	0.045	0.9499	0.012	0.7226	0.03
$\mathbf{S}_{\text{vbin}}$	0.3281	0.001	-103.4697	0.556	-197.0881	0.028	-264.4955	0.158	9.1933	0.957
$\mathbf{S}_{\text{imagery}}$	14.8346	0.000	0.6861	0.048	2.065	0.057	0.4584	0.643	1.1321	0.241
$\mathbf{S}_{\text{energy}}$	-0.0543	0.985	0.9405	0.031	1.3163	0.423	2.4855	0.086	1.8852	0.191
$\mathbf{S}_{\text{banality}}$	-2.8474	0.128	0.5269	0.636	-0.2816	0.804	0.1945	0.866	3.2007	0.003
$\mathbf{S}_{\text{I1}}$	0.0549	0.084	-0.1213	0.609	-0.0655	0.745	-0.2414	0.255	0.3546	0.116
$\mathbf{S}_{\text{I2}}$	-0.0188	0.499	13.8702	0.027	14.5632	0.142	6.5436	0.438	3.1841	0.72
$\mathbf{S}_{\text{all}}$	0.0760	0.025	-1.0131	0.022	-0.3216	0.31	-0.2537	0.306	1.3457	0.145

Table 3: Results of interaction testing. I1 denotes the interaction between linguistic features (surprisal, entropy, banality, NPMI), and I2 denotes the interaction between stylistic features (imagery, word energy, abstraction, valence).

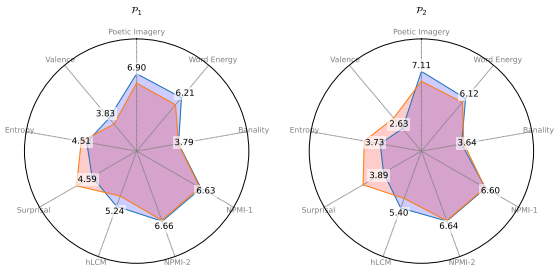

Figure 3: Radar charts showing preference profiles for two participants in the *EvocativeLines* dataset. Blue and red webs denote positive and negative lines, respectively. Figure S.3 on the project website shows a complete profile.

Identifying Significant Features Across Preference Profiles

To provide deeper insights into what drives individual preferences, we conducted statistical testing to determine the correlations between specific features and user judgments. We fit a logistic regression model with interaction terms to evaluate the significance of each feature on whether a participant found a poetic line inspiring.

For this analysis, we simplified our temporal features by averaging across the timestep dimension, converting each into a single score. While this representation sacrifices some temporal information, it allows for clearer interpretation of feature importance. Table 3 presents results for five selected participants: $\mathcal{P}_1$ and $\mathcal{P}_2$ (selected for their data-rich annotations) and $\mathcal{P}_3$, $\mathcal{P}_6$, and $\mathcal{P}_7$ (selected as a triplet with particularly low inter-participant agreement).

Our analysis reveals distinct preference patterns across participants. For $\mathcal{P}_1$, contextual entropy ($p = 0.034$) shows a significant negative correlation, indicating a preference for lines with higher unpredictability. Unidirectional NPMI ($p < 0.001$) shows a strong positive correlation, while bidirectional NPMI ($p = 0.002$) shows a significant negative correlation. Combined with a strong preference for poetic imagery (coefficient=14.83, $p < 0.001$), this suggests $\mathcal{P}_1$ values lines with clear linear progression that maintain creative impact when read in the intended direction.

For $\mathcal{P}_2$, we observe significant preferences for word energy ($p = 0.031$), level of abstraction ($p = 0.047$), and poetic imagery ($p = 0.048$), along with a significant interaction effect between stylistic features ($p = 0.027$). This indicates $\mathcal{P}_2$ values the combined impact of multiple poetic elements rather than any single dimension.

Each participant shows a unique signature of feature preferences—$\mathcal{P}_3$ values abstraction and unidirectional word associations, $\mathcal{P}_6$ prefers causality and abstraction, while $\mathcal{P}_7$ values originality (low banality, $p = 0.003$) and abstraction ($p = 0.03$).

These results demonstrate that our framework not only predicts preferences accurately but also provides interpretable insights into the specific linguistic and poetic elements that resonate with different individuals.

Qualitative Analysis

We visualize preference profiles as radar charts (See Figure 3) that display average feature values across dimensions for inspiring (blue) and non-inspiring (red) poetic lines. These visualizations reveal consistent distinctions between inspiring and non-inspiring examples across all participants, confirming our features effectively capture meaningful differences in artistic preferences. Notably, all participants prefer lines with higher poetic imagery and word energy, suggesting these elements form an important foundation for evocative content. However, each profile exhibits unique patterns in how other features (abstraction, surprisal, entropy) contribute to inspiration, highlighting the personalized nature of artistic preferences.

Individual example visualizations (Figure S.4 on the project web site) reveal complex feature interactions. For instance, "all the world is the fire" maintains high inspirational quality despite neutral emotional valence (0) through elevated imagery and energy values. This demonstrates how strength in certain dimensions can compensate for neutrality in others. Similarly, we observe that high abstraction paired with moderate surprisal creates lines that balance philosophical depth with cognitive accessibility.

Rarely does a single feature determine a line's inspirational quality. Instead, specific combinations—functioning as coherent stylistic signatures—resonate with individual preferences, with particularly strong correlations observed between inspiration and both word energy and abstraction across profiles.

Conclusion

In this work, we propose an interpretable and flexible framework for computational modeling of artistic inspiration in AI-generated poetic lines. Our approach assists artists in the Seed phase by identifying potentially "inspiring" lines from larger pools of candidates. We identified key linguistic and stylistic features that accurately forecast aesthetic preferences across diverse individuals. Our experiments demonstrate that our framework significantly outperforms state-of-the-art language models like LLaMA-3-70b classifiers. Crucially, our approach extends beyond poetry analysis. It can be adapted with different feature sets to predict other subjective aesthetic

preferences—for comparing sonnets from different poets or analyzing prose stylistics. In the context of creating art, it is essential to measure how well AI systems support individual creative processes that span a wide spectrum of artistic preferences and skill levels.

References

Althuizen, N. 2021. Revisiting Berlyne's inverted U-shape relationship between complexity and liking: The role of effort, arousal, and status in the appreciation of product design aesthetics. *Psychology & Marketing* 38(3):481–503.

Amabile, T. M. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of personality and social psychology* 43(5):997.

Baer, J. 2011. How divergent thinking tests mislead us: Are the Torrance tests still relevant in the 21st century? the division 10 debate. *Psychology of Aesthetics, Creativity, and the Arts* 5(4):309–313.

Bahdanau, D. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Banerjee, S., and Lavie, A. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Berlyne, D. E. 1975. Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation. *Journal of Aesthetics and Art Criticism* 34(1):86–87.

Bouma, G. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL* 30:31–40.

Breiman, L. 2001. Random forests. *Machine learning* 45:5–32.

Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; and Shah, R. 1993. Signature verification using a "siamese" time delay neural network. In *Proceedings of the 7th International Conference on Neural Information Processing Systems*, NIPS'93, 737–744. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Brooks, C., and Warren, R. P. 1976. *Understanding Poetry*. New York: Holt, Rinehart and Winston, 4th edition.

Brysbaert, M.; Warriner, A. B.; and Kuperman, V. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods* 46:904–911.

Burroughs, W. 2012. *The Job: Interviews with William S. Burroughs*. Penguin Modern Classics. Penguin Books Limited.

Cassirer, E. 1946. *Language and myth*, volume 51. Courier Corporation.

Chakrabarty, T.; Laban, P.; Agarwal, D.; Muresan, S.; and Wu, C.-S. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–34.

Charland, L. C. 2005. The heat of emotion: Valence and the demarcation problem. *Journal of consciousness studies* 12(8-9):82–102.

Chen, T., and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine Learning* 20:273–297.

Cross, C. R. 2001. *Heavier Than Heaven: A Biography of Kurt Cobain*. New York: Hyperion.

Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; and Ravi, S. 2020. GoEmotions: A dataset of fine-grained emotions. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4040–4054. Online: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dunnigan, S. M. 2014. *7 Poetic Imagery*. Edinburgh: Edinburgh University Press. 67–77.

Fu, J.; Ng, S.-K.; Jiang, Z.; and Liu, P. 2024. GPTScore: Evaluate as you desire. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6556–6576. Mexico City, Mexico: Association for Computational Linguistics.

Gold, B. P.; Pearce, M. T.; Mas-Herrero, E.; Dagher, A.; and Zatorre, R. J. 2019. Predictability and uncertainty in the pleasure of music: A reward for learning? *Journal of Neuroscience* 39(47):9397–9409.

Hashemi, H.; Eisner, J.; Rosset, C.; Van Durme, B.; and Kedzie, C. 2024. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13806–13834.

Hashimoto, T. B.; Zhang, H.; and Liang, P. 2019. Unifying human and statistical evaluation for natural language generation. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1689–1701. Minneapolis, Minnesota: Association for Computational Linguistics.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Johnson-Grey, K. M.; Boghrati, R.; Wakslak, C. J.; and Dehghani, M. 2020. Measuring abstract mind-sets through syntax: Automating the linguistic category model. *Social Psychological and Personality Science* 11(2):217–225.

Jones, D. 2017. *David Bowie: A Life*. New York: Crown Archetype, first edition.

Kao, J., and Jurafsky, D. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature*, 8–17.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2511–2522. Singapore: Association for Computational Linguistics.

Liu, Y. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lowder, M. W.; Choi, W.; Ferreira, F.; and Henderson, J. M. 2018. Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive Science* 42(S4):1166–1183.

Nelder, J. A., and Wedderburn, R. W. 1972. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society* 135(3):370–384.

Palahniuk, C. 2020. *Consider This: Moments in My Writing Life after Which Everything Was Different*. Grand Central Publishing.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research* 12:2825–2830.

Pillutla, K.; Swayamdipta, S.; Zellers, R.; Thickstun, J.; Welleck, S.; Choi, Y.; and Harchaoui, Z. 2024. Mauve: measuring the gap between neural text and human text using divergence frontiers. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21. Red Hook, NY, USA: Curran Associates Inc.

Pound, E. 1913. A few don'ts by an imagiste. *Poetry* 1(6):200–206.

Quinlan, J. R. 1986. Induction of decision trees. *Machine learning* 1:81–106.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8):9.

Reimers, N., and Gurevych, I. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Rubin, R. 2023. *The creative act: a way of being*. Penguin.

Scherer, K. R. 1984. On the nature and function of emotion: a component process approach. approaches to emotion.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Semin, G. R., and Fiedler, K. 1988. The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of personality and Social Psychology* 54(4):558.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal* 27(3):379–423.

Vechtomova, O.; Sahu, G.; and Kumar, D. 2021. Lyricjam: A system for generating lyrics for live instrumental music. In *Proceedings of the 12th Conference on Computational Creativity*.