Harmonious Minds: Benchmarking Intertwined Reasoning of Human Personality and Musical Preference

Sayantan Pal, Souvik Das[†], Rohini K. Srihari

Department of Computer Science and Engineering,
State University of New York at Buffalo
[†]JPMorgan Chase & Co.[*]

Abstract

Understanding how large language models (LLMs) reason across semantically distinct domains remains an open challenge. In this work, we investigate whether LLMs can connect personality traits to musical preferences, specifically chord progressions. Drawing on psychological theory and symbolic music structure, we introduce a novel benchmark that evaluates two interdependent tasks: (1) inferring personality traits from a textual context and (2) selecting a musically appropriate chord progression aligned with the inferred trait. We release a synthetic, expert-guided dataset grounded in Cattell's 16 Personality Factors (PF16), genre-conditioned chord structures, and diverse situational contexts. We explore multiple learning strategies, including fine-tuning task-specific corpora, model merging with LoRA adapters, and advanced prompt-based reasoning techniques such as verbalization. Additionally, we propose a teacher-student framework to evaluate the quality of model-generated explanations using a five-dimensional rubric. Our findings show that verbalization outperforms standard reasoning methods, achieving up to 11% improvement over zero-shot baselines.

1 Introduction

Music is more than just entertainment; it is often an emotional extension of the self, subtly reflecting our moods, values, and personalities (Juslin, 2010; Flannery and Woolhouse, 2021; Chamorro-Premuzic and Furnham, 2007; Ferwerda et al., 2017). Across cultures, the music people enjoy and create provides a window into who they are, offering insights into their internal worlds. In recent years, conversational AI has seen a surge of interest in simulating human-like personas using datasets such as Persona Chat (PC) (Zhang et al., 2018) and Blended Skill Talk (BST) (Smith et al., 2020). These efforts have been further extended by creating a Synthetic Persona Chat (SPC) (Jandaghi et al., 2024) dataset, which generates artificial profiles to diversify persona-driven dialogue. Parallel to this, newer datasets like Journal Intensive Conversations (JIC) (Pal et al., 2025) aim to capture more intrinsic personality traits by grounding conversations in long-form, autobiographical texts. While such datasets have improved the personalization capabilities of large language models (LLMs) (Kasahara et al., 2022), they are primarily confined to understanding personality within the dialogue domain. What remains underexplored is whether LLMs can reason across domains (Simon et al., 2022; Akyürek et al., 2023), particularly in capturing the connection between personality traits (Mazaré et al., 2018; Lee et al., 2025) and musical preferences. *We hypothesize that the music individuals resonate with is not a random choice but a psychologically grounded expression*

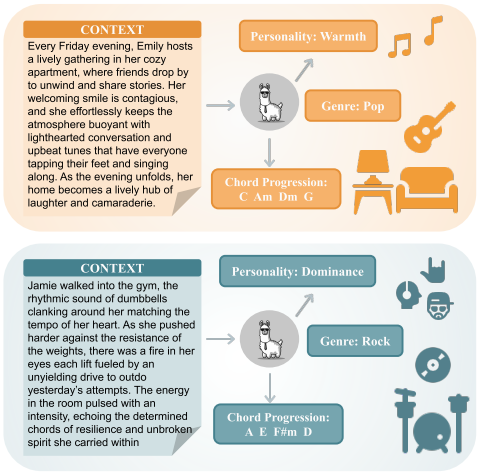

Figure 1: Illustrating our benchmark's core idea: *Can language models link personality traits to musical structures?* Given a context, the model infers personality and selects a matching genre and chord progression, reflecting distinct creative preferences.

[*]This work is independent research and does not involve JPMorgan Chase & Co.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2000–2018

of who they are. For instance, highly extroverted individuals may gravitate toward energetic genres like rock or hip-hop, while more introspective personalities may find alignment with genres such as jazz or pop. This connection is intuitive and supported by decades of psychological literature (Chamorro-Premuzic and Furnham, 2007; Ferwerda et al., 2017), yet remains largely untested in computational models. This raises important questions about whether LLMs can reason consistently when connecting personality traits to other domains like music, a low-resource domain with symbolic structure and limited textual data (Huang and Yang, 2020).

Personality refers to the characteristic patterns of thoughts, emotions, and behaviors that define how individuals perceive and interact with the world (Mairesse et al., 2007; McCrae and Costa Jr., 1999; Sanchez-Roige et al., 2018). In psychology, several models have been proposed to represent personality traits, including the Big Five (OCEAN (Hurtz and Donovan, 2000; Azucar et al., 2018)), the Myers-Briggs Type Indicator (MBTI)(Cohen et al., 2013), and Raymond Cattell's 16 Personality Factors (PF16) (Cattell and P. Cattell, 1995). On the other hand, chord progression in music theory refers to a sequence of chords that forms the harmonic foundation of a musical piece (Cho et al., 2016; Kawase, 2024). Different progressions evoke distinct emotional tones and are often linked to particular genres, moods, or artistic styles (Bakker and Martin, 2015). Music theory and chord reasoning are relatively low-resource and underrepresented in mainstream LLM training corpora, making it difficult for models to generalize across these modalities (Yuan et al., 2024). Understanding how LLMs interpret the relationship between personality traits and musical inclinations can open new possibilities for personalized content generation (Wang et al., 2024), human-centered creative tools (Spangher et al., 2025), and emotionally aware recommendation systems (Lyu et al., 2024). Addressing this cross-domain reasoning challenge is a step forward in improving LLM capabilities and is crucial for building AI that reflects the interconnectedness of human cognition, creativity, and identity.

The relationship between personality and music is not just theoretical; it is often explicitly articulated by artists when reflecting on their creative process. For example, Taylor Swift breaks down "Blank Space"[1], she explains how the song responds to the media's exaggerated portrayal of her dating life, crafting a fictionalized character based on public perception. Similarly, Charlie Puth breaks down "Attention"[2], describing how it evolved from a piano ballad into a layered pop track reflecting his emotional state and artistic confidence. He emphasizes his production choices, such as using acoustic guitar rhythms, vinyl textures, and deliberate drops as extensions of his personal identity and emotional landscape. Billie Eilish and Finneas, in their deconstruction of "Bad Guy"[3], highlight their spontaneous, personal approach to sound design, incorporating chaotic sonic quirks that mirror their creative personalities. These examples suggest that musical composition often encodes deeper personality traits and emotional intentions. This motivates our central question as shown in Figure 1: Can large language models, which perform well in dialogue personalization, also reason across domains to connect personality with musical structure?

In this work, (1) we release a **novel benchmark** that evaluates whether large language models can **reason across domains** by aligning personality traits with musical structure. (2) Our benchmark is designed with a **modular construction methodology**, allowing extensibility to more challenging variants. (3) We explore two key training strategies: **personality-grounded fine-tuning** and **joint adaptation across both tasks**. (4) To improve performance without training, we propose a **verbalization-based inference pipeline** that dynamically guides chord selection, yielding a consistent **11% improvement** over zero-shot baselines. (5) Finally, we introduce a **thinking evaluation protocol** to assess reasoning quality through **teacher-rated justifications**, enabling evaluation beyond output correctness. Our code and data is publicly available.[4]

2 Related Work

Personality Grounding in Conversational LLMs. Understanding and generating persona-consistent responses has been a long-standing goal in dialogue systems (Liu et al., 2016; Caron and Srivastava, 2023; Saha et al., 2022). Early approaches (Yamashita et al., 2023; Zhang et al., 2018) such

[1] Blank Space breakdown by Taylor Swift
[2] Attention breakdown by Charlie Puth
[3] Bad Guy breakdown by Billie Eilish and Finneas
[4] Code and Data

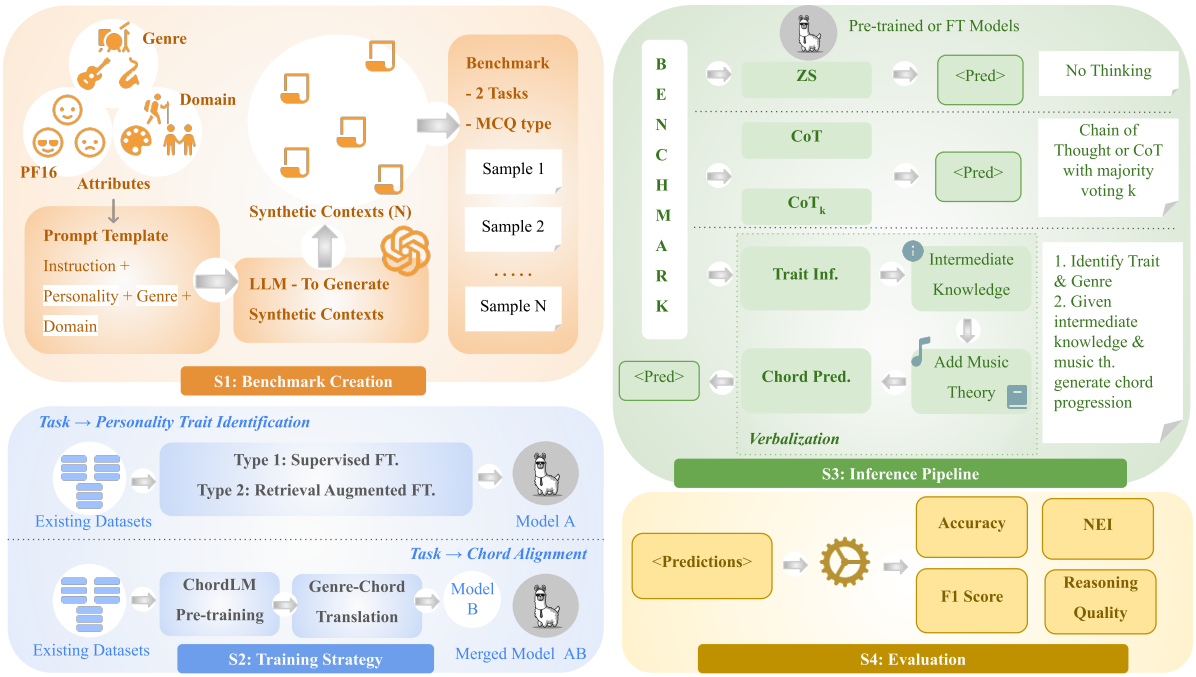

Figure 2: Overview of our benchmark design and methodology. **S1: Benchmark Creation**- We generate synthetic contexts using PF16 traits, genres, and domains as prompts, forming two multiple-choice tasks per sample. **S2: Training Strategy**- Task 1 models are trained on existing dialogue datasets, while Task 2 uses chord pretraining and genre-conditioned generation. **S3: Inference Pipeline**- We evaluate models under zero-shot, CoT, and verbalized prompting strategies. Verbalization chains trait inference and music theory to guide chord generation. **S4: Evaluation**- Outputs are assessed using Accuracy, F1, NEI, and reasoning quality.

as PersonaChat (Zhang et al., 2018) and Blended-SkillTalk (Smith et al., 2020) used manually authored or template-based persona attributes to drive more faithful conversations (Dušek and Jurčíček, 2016). Synthetic extensions like SPC (Jandaghi et al., 2024) diversified persona representation, while Journal Intensive Conversations (JIC) (Pal et al., 2025) introduced free-form autobiographical grounding to capture more nuanced traits. Beyond static attributes, recent research has explored integrating psychological theories such as the Big Five (OCEAN) (Azucar et al., 2018; Hurtz and Donovan, 2000) and PF16 (Cattell and P. Cattell, 1995) to enrich modeling of personality in LLMs. While fine-tuning LLMs on personalized data has led to more trait-aligned generations (Labruna et al., 2024; Pal et al., 2024), most work focuses exclusively on dialogue settings, without examining if such personalization can extend to semantically distant domains like music.

Language Models for Symbolic Music Understanding. Recent works (Copet et al., 2023) have explored LLM capabilities in music composition and understanding, primarily through symbolic representations such as chord labels or textual scores. ChatMusician (Yuan et al., 2024) introduced a large-scale dataset and training framework for music-related dialogue tasks, while Chordonomicon (Kantarelis et al., 2024) compiled a genre-specific chord progression database aimed at symbolic pattern learning. Models like MusicTransformer (Huang et al., 2018) and MusicNet (Thickstun et al., 2017) have focused on sequence modeling in music but often rely on specialized musical representations or piano-roll formats. In contrast, our work focuses on the textual reasoning behind chord choices, leveraging symbolic descriptors grounded in genre and personality. Unlike prior approaches that focus on chord generation or genre classification, we pose chord selection as a cross-domain reasoning task, situated at the intersection of psychological and creative semantics.

3 Personality-Music Alignment Bench (PMAB)

3.1 Motivation and Novelty

Recent advances in personalized dialogue and creative language generation have treated these domains separately. Yet, real-world reasoning is rarely so isolated. Our benchmark takes a first

step toward evaluating whether LLMs can connect personality traits expressed through language with corresponding musical choices like chord progressions. It challenges models to carry intent across symbolically and semantically distinct domains without training, supervision, or domain-specific fine-tuning. Designed purely for evaluation, this two-stage task tests how well LLMs generalize and align outputs across contexts that demand interpretation and creativity. Figure 2 outlines the overview of our benchmark design and methodology.

3.2 PMAB Design: Annotation, Terminology, and Mapping Strategy

Personality Trait Framework. We adopt Cattell's 16 Personality Factor (PF16) model[5] to represent fine-grained human traits across both ends of a bipolar scale. Each trait includes a high and low pole, capturing variations in behavior and emotional disposition. For instance, the trait *Warmth* spans from descriptors like "outgoing" and "participating" to "reserved" and "aloof."

Genre Alignment via Expert-Led Mapping. To connect personality traits to musical preferences, we curated a genre inventory consisting of 19 categories (Apdx. Table 8). Using GPT-4o under the guidance of psychology and music experts, we annotated each PF16 trait with *alike genres* (those resonating with the trait) and *different genres* (those misaligned). For example, high scorers on *Warmth* were mapped to genres like *Pop*, *Country*, and *Gospel*, reflecting expressive and socially oriented aesthetics. Conversely, low scorers were aligned with introspective or emotionally distant genres like *Rock*, *Electronic*, and *Metal*. These mappings form the backbone of our benchmark's trait-to-genre reasoning tasks and are detailed in Apdx. Table 9.

Chord Progression Design. To support musically grounded inference, we sourced chord progressions from the Chordonomicon dataset (Kantarelis et al., 2024), a structured repository of genre-labeled chord progression patterns. For each genre, we selected 10 most representative progressions based on the frequency and stylistic fit. These were then reviewed by musicians to ensure genre authenticity. A subset of 5 genres with 4 representative progressions each is presented in Apdx. Table 10, along with associated danceability and mood characteristics. This mapping enables precise chord inference aligned with inferred personality and genre labels.

Algorithm 1 Synthetic Context Generation for Personality-Music Benchmark

Input: PF16 trait key t, trait-to-genre mapping $\mathcal{G}$, domain set $\mathcal{D}$ with subdomains, descriptor set $\mathcal{H}_t$ (high polarity), sampling parameter K
Initialize: GPT-4o as context generator $\mathcal{M}$
Create two descriptor subsets: $\mathcal{H}_1, \mathcal{H}_2$ from $\mathcal{H}_t$ with $|\mathcal{H}_i| = 3$
for each domain $d \in \mathcal{D}$ **do**
  for each subdomain $s \in d$ **do**
    for each genre $g \in \mathcal{G}(t)$ **do**
      for each descriptor set $\mathcal{H}_i \in \{\mathcal{H}_1, \mathcal{H}_2\}$ **do**
        Generate K candidate contexts $\{c_1, c_2, \ldots, c_K\}$ using GPT-4o given prompt $(\mathcal{H}_i, g, d, s)$
        Use GPT-4o to select the most diverse and representative context $c^* = \arg\max_{c_k} \texttt{representativeness}(c_k)$
        Append context c^* to dataset $\mathcal{C}$
      end for
    end for
  end for
end for
Output: Synthetic context set $\mathcal{C}$ grounded in PF16 traits, musical genres, and domain diversity

Synthetic Domain and Subdomain Construction. To ensure wide topical coverage and contextual variability, we constructed 10 high-level domains (e.g., *Personal Development*, *Creative Expression*), each with 3 coherent subdomains. These were generated using GPT-4o with prompt templates encouraging socio-cognitive realism, emotional nuance, and stylistic diversity (Apdx. Table 11). Our method is inspired by the prompt diversification approach of the TÜLU dataset[6] from AllenAI, which showed that structured instruction generation improves coverage and generalization. This design simulates real-world contexts where personality traits manifest through narrative, behavior, and preference.

3.3 Synthetic Context and PMAB Construction

Context Generation. We generate compact, fictional scenarios that implicitly reflect specific personality traits and musical preferences. Each scenario is conditioned on: (i) a PF16 trait t, (ii) a descriptor subset $\mathcal{H}_t$, (iii) a genre g aligned with the trait, and (iv) a domain-subdomain pair (d, s). A generation model $\mathcal{M}$ (GPT-4o) receives the prompt:

$$\mathcal{P} = \texttt{prompt}(\mathcal{H}_t, g, d, s),$$

and generates K narrative candidates $\{c_1, \ldots, c_K\} = \mathcal{M}(\mathcal{P})$. The final context c^* is selected as:

$$c^* = \arg\max_{c_k} \texttt{representativeness}(c_k),$$

[5] Wiki-16PF-Questionnaire

[6] Tulu-v2-sft-mixture

Algorithm 2 Final Benchmark Construction for Personality-Music Alignment

Input: Generated contexts $\mathcal{C}$, PF16 types $\mathcal{T}$, chord progression map $\mathcal{M}$, embedding model $\mathcal{E}$
Output: Flat benchmark dataset $\mathcal{B}$ with dual-task annotations
Compute context embeddings $\mathcal{E}(\mathcal{C})$
Build descriptor pool $\mathcal{D}$ from all PF16 trait descriptors (high + low)
Compute descriptor embeddings $\mathcal{E}(\mathcal{D})$
for each context $c \in \mathcal{C}$ **do**
  // Task 1: Personality Identification
  Select correct descriptor $d_{gt} \in c[\texttt{high}]$
  Exclude descriptors belonging to the same PF16 type
  Select 3 distractors d_i with lowest cosine similarity to $\mathcal{E}(c)$
  Shuffle correct option with distractors to form Task 1 choices
  // Task 2: Chord Progression Matching
  Select genre $g = c[\texttt{alike\_genre}]$
  Choose one correct progression from $\mathcal{M}[g]$
  Choose 3 distractors from other genres in $\mathcal{M}$
  Shuffle correct option with distractors to form Task 2 choices
  Append entry to benchmark $\mathcal{B}$ with:
  context, domain, subdomain, PF16 type, genre,
  Task 1: question, options, correct label
  Task 2: question, options, correct label
end for
Return benchmark dataset $\mathcal{B}$

ensuring trait expressivity and musical alignment. Further details are provided in Algorithm 1 and Appendix A.1.

Benchmark Tasks. Each context is converted into two multiple-choice tasks: *(1) Personality Trait Identification* and *(2) Chord Progression Matching*. Both follow a 1-correct, 3-distractor format. Distractors for Task 1 are selected based on semantic distance in embedding space, while Task 2 distractors are sampled from genres unrelated to the aligned one. Algorithm 2 provides the full construction pipeline.

Modularity and Extensibility. The benchmark is designed to support harder variants: distractors can be made more semantically similar, genre-chord mappings can be made noisier, or new context types can be added. This modular design allows the benchmark to scale in complexity, enabling a more nuanced evaluation of cross-domain reasoning.

3.4 Quality Control and Evaluation Attributes

Automated Filtering during Context Generation. As part of the generation process, we sample $K = 1$ to 3 candidate contexts per configuration and retain the most representative sample, as selected by GPT-4o based on a diversity and relevance prompt. This initial filtering ensures each retained scenario reflects its intended personality-musical alignment.

Human and Model-Based Evaluation. To further verify quality, we randomly select 20 contexts per PF16 trait and 3 human judges and GPT-4o evaluate them across five qualitative dimensions: *personality alignment*, *musical coherence*, *naturalness*, *implicitness*, and *specificity*. Each context is rated independently by three expert annotators and GPT-4o. Appendix A.3 defines each evaluation dimension, and Table 12 has the mean and SD of the evaluation.

Reliability and Qualitative Evidence. Inter-rater reliability is assessed using Intraclass Correlation Coefficients (ICC), reported in Apdx. Table 13. Qualitative examples are included in Appendix A.2.

4 Benchmark Statistics

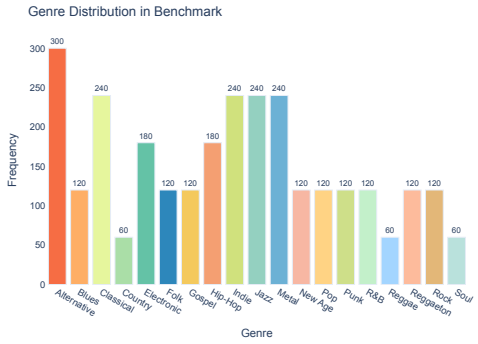

Figure 3: Genre distribution across the 2,880 benchmark contexts. Each bar represents the frequency of a genre used in the dataset. The genres are alphabetically sorted, and the distribution is designed to maintain stylistic and semantic diversity across personality-aligned contexts.

Our benchmark consists of a total of 2,880 context–question pairs. For each trait, we generate exactly 180 unique contexts, resulting in a balanced distribution across personality categories. The benchmark spans 10 high-level domains, each further subdivided into 3 subdomains. This leads to 288 data points per domain and 96 samples per subdomain, ensuring uniform representation across the entire context space.

In total, the benchmark covers 19 unique musical genres, with personality-genre mappings grounded in trait semantics and verified through expert feedback. We measure the diversity of the dataset using the number of unique categories and Shannon entropy for both personality traits and genres, summarized in Table 1. Genre distribution across the benchmark is visualized in the bar chart in Figure 3.

5 Experimentation

5.1 Training Strategy

We fine-tune both benchmark tasks using parameter-efficient fine-tuning (PEFT) with Low-Rank Adaptation (LoRA) (Hu et al., 2022) on

Table 1: Benchmark Composition and Diversity Statistics

Attribute	Unique Values	Entropy
Total contexts	2,880	–
Personality traits	16	2.773
Musical genres	19	2.844
Domains	10	–
Subdomains per domain	3	–

LLaMA 3 8B (AI@Meta, 2024) and Mistral 7B v0.3 (Jiang et al., 2023). Only the projection layers (W_q, W_k, W_v, W_o) are updated while all other weights remain frozen.

Let $\mathcal{D}^{(1)}_{\text{train}}$ and $\mathcal{D}^{(2)}_{\text{train}}$ denote the training sets for Task 1 (T1) and Task 2 (T2), respectively. The general training objective is to minimize the negative log-likelihood (NLL) loss:

$$\mathcal{L}_{\text{task}}(\theta) = -\frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \log p(y|x;\theta),$$

where x is the input (e.g., context, genre, or chord prefix) and y is the target label.

LoRA Optimization. Each attention projection matrix is updated using low-rank decomposition:

$$W'_q = W_q + A_q B_q, \quad A_q \in \mathbb{R}^{d\times r}, \quad B_q \in \mathbb{R}^{r\times d},$$

with r as the rank hyperparameter. The same applies to W_k, W_v, and W_o.

T1: Personality Trait Identification. We fine-tune both models using multiple dialogue-style datasets annotated with personality traits: PC, BST, SPC, and JIC following the strategy by (Pal et al., 2025). These datasets help models capture a wide range of personality-grounded linguistic cues. See Appendix B.1.1 for dataset statistics.

T2: Chord Alignment and Adapter Merging. For Task 2, we first pre-train a chord-focused language model on sequences from Chordonomicon, followed by genre-conditioned chord generation. We additionally perform inverse modeling to predict genre from chord sequences. Finally, we merge the personality and chord adapters using SVD-based LoRA merging from the PEFT library, enabling joint reasoning over dialogue and symbolic music. Detailed strategy, implementation, and training configurations are provided in Appendix B.1.

5.2 Inference Pipeline

We evaluate models under three setups: **(1) Zero-shot prompting**, and **(2) Guided prompting strategies**.

(1) Zero-shot Prompting. Given a context x, we prepend a task description and present four candidate options. The model selects the most probable output without training. This serves as our baseline. Additionally, we use the fine-tuned adapter $\mathcal{M}_{\text{JIC}}$ for Task 1 and a merged model $\mathcal{M}_{\text{SVD}}$ for Task 2. Outputs are compared to gold labels using classification metrics.

(2) Guided Prompting. To improve reasoning, we use structured prompting methods inspired by chain-of-thought (CoT) and staged inference.

(A) CoT Prompting. The model is encouraged to reason step by step using a modified prompt to think step by step.

(B) Ensemble CoT. We sample $k = 5$ completions $\{y_1, ..., y_k\}$ and take the majority vote:

$$\hat{y} = \texttt{mode}\{y_1, y_2, ..., y_k\}.$$

(C) Verbalization Pipeline (Task 2). This is our key reasoning enhancement method. We introduce a two-step prompt chaining approach to align personality traits with genres and then map those genres to suitable chords using music theory.

Step 1: Trait → Genre Reasoning. Given a context x, we prompt the model to infer the dominant trait and its associated genres using a structured prompt $\mathcal{P}_{\text{genre}}$:

$$\hat{g} = \mathcal{M}_{\text{verbal}}(\mathcal{P}_{\text{genre}}(x))$$

Step 2: Genre → Chord Reasoning with Music Theory. The next prompt $\mathcal{P}_{\text{chord}}$ is conditioned on the context x, predicted genre $\hat{g}$, and corresponding music theory explanation $\mu(\hat{g})$. The model generates a rationale and selects a suitable chord progression:

$$\hat{y} = \mathcal{M}_{\text{verbal}}(\mathcal{P}_{\text{chord}}(x, \hat{g}, \mu(\hat{g})))$$

The final prediction is extracted from the generated response using answer tags. Prompt formats are detailed in Appendix B.2.1.

5.3 Evaluation Strategy

Task-Level Performance Metrics. We evaluate both benchmark tasks using standard classification metrics: accuracy (Acc) and macro-averaged F1 score (F1). In addition, we report the *Not Enough Information* (NEI) count, which reflects model responses that are vague, contradictory, or insufficiently grounded in context. NEI is triggered when the model fails to select a valid option.

Reasoning Quality Assessment. To assess deeper model understanding, we evaluate reasoning quality for Task 2 on randomly selected 500 samples. For each model, we prompt a teacher model (GPT-4o) to generate ideal responses based on the correct chord label. The student model's explanation is then rated on a 1–5 Likert scale across five dimensions: *Personality-Musical Alignment (PMA)*, *Chordal Appropriateness (CA)*, *Causal Justification (CJ)*, *Specificity (S)*, and *Fluency and Clarity (F)*. Definitions of these dimensions and the full scoring rubric are provided in Appendix C.1. We report both the mean score and the percentage of high-quality responses (score ≥ 4) for each dimension.

Benchmark Validity. To validate Task 1, we compare model performance against the LM Evaluation Harness[7] OCEAN benchmark.

6 Results and Discussion

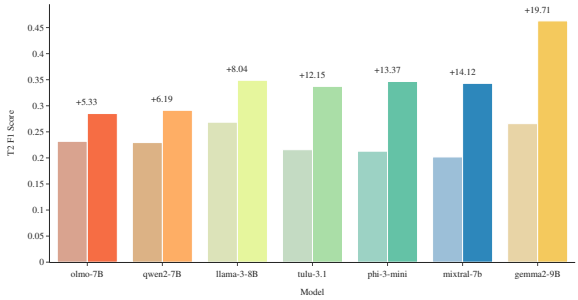

Figure 4: Task 2 F1 scores for baseline and verbalized setups across models, sorted by improvement.

Model	T2 F1 (B)	T2 F1 (V)	%Imp.
phi-3-mini	0.2129	0.3466	13.37
mixtral-7b	0.2018	0.3430	14.12
llama3-8B	0.2685	0.3489	8.04
tulu3.1-8B	0.2157	0.3372	12.15
qwen2-7B	0.2294	0.2913	6.19
olmo2-7B	0.2318	0.2851	5.33
gemma2-9B	0.2657	0.4628	**19.71**

Table 2: Comparison of Task 2 F1 scores between zero-shot baseline (B) and verbalized prompting (V) for open-source models. Verbalization consistently improves performance, with largest gains seen in `gemma2-9B`. Detailed results of all tasks in Table 14

[7] https://github.com/EleutherAI/lm-evaluation-harness

6.1 Inference-Only Evaluation

We begin by analyzing model performance under inference-only settings, which include zero-shot prompting, chain-of-thought (CoT), majority-vote ensembles (CoT$_k$), and our verbalization pipeline. Table 2 reports performance across second task. Detailed results in Apdx. Table 14

Task 1 (Personality Trait Identification). Zero-shot prompting yields strong performance across models, with F1 scores consistently above 0.81. Surprisingly, CoT reasoning does not always help: models like `mistral`, `llama3` and `gemma2` see degraded F1 with CoT$_k$, indicating overthinking or drift in trait inference. The best-performing model in Task 1 is `gemma2-9b-it` (F1: 0.8988), followed by `llama-3-8Bb` (F1: 0.8620).

Task 2 (Chord Progression Alignment). In contrast to Task 1, the baseline F1 scores for Task 2 are significantly lower (ranging from 0.18–0.27), highlighting the inherent difficulty of musical reasoning. Here, CoT strategies offer only marginal improvements (1–3%) in most models. However, our **verbalization pipeline** consistently provides the largest gains, improving F1 by over **13%** on `phi-3`, `mixtral`, and `Tulu-3.1`, and over **19%** on `gemma2-9b-it`. This validates the importance of modular reasoning and staged prompt design in music-theoretic alignment tasks. Additional insights in Appendix D.1

Commercial Models. Closed-source models like `GPT-4.1` and `GPT-4o` outperform open models, with `GPT-4.1-mini` scoring highest on Task 1. For Task 2, performance improves with newer and larger models. Due to API cost, we report only zero-shot results for these systems.

Overall Insights. Verbalization consistently outperforms unguided reasoning strategies such as CoT and CoT$_k$ by explicitly chaining personality inference to genre selection and symbolic chord prediction. This structured, staged inference enables better control and alignment than generic prompting methods. Notably, `deepseek-llama3.1`, a model pretrained to "think" by design, outperforms its supervised counterpart (`Tulu-3.1`) in Task 2, highlighting that models trained for reflective reasoning inherently generalize better than those forced into such behavior via prompts alone.

Further, the pretrained capabilities of open-source models reveal strong variance. As shown

in Figure 4 and Table 2, base performance on Task 2 varies significantly. `Llama3` and `gemma2` start strong, whereas `phi-3` and `mistral` begin weaker. However, the gains from guided prompting are also uneven: models like `phi-3` and `mistral` achieve large improvements post-verbalization, sometimes even surpassing stronger baselines like `llama3`. Meanwhile, `gemma2` maintains top performance throughout, demonstrating both high base capability and strong responsiveness to verbalized guidance.

6.2 Impact of Fine-Tuning and Merged Models.

Among all supervised setups, fine-tuning on JIC provides the largest gains for Task 1, improving F1 scores by 1.4% (Mixtral) and 2.4% (LLaMA-3), consistent with trends from the LM Harness benchmark (Pal et al., 2025). However, merged models fail to yield meaningful improvements on Task 2, likely due to the dataset gap: symbolic chord alignment may require dialogue-style data where musical preferences are explicitly discussed (e.g., artist interviews). Notably, the verbalized inference variant (`Merged (V)`) still delivers the best Task 2 performance without fine-tuning, reaffirming the strength of prompt chaining. Detailed results in Table 15.

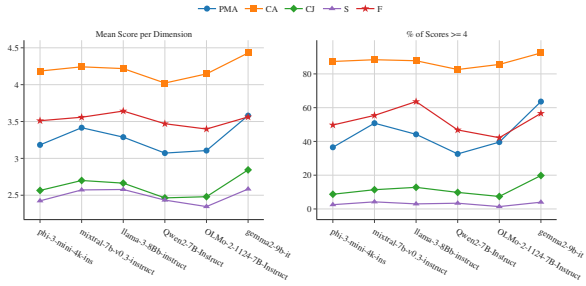

Figure 5: Reasoning quality across models on five dimensions: PMA, CA, CJ, S, and F. Left: mean Likert scores (1–5). Right: percentage of high-quality responses (≥ 4).

6.3 Reasoning Quality Across Models.

Figure 5 illustrates the reasoning quality of models along five dimensions. The left plot shows the mean Likert scores (1–5), while the right plot reports the percentage of responses rated ≥ 4. We observe in Table 3 that `gemma2-9b-it` achieves the highest scores across most dimensions, especially in PMA, CA, and CJ, indicating consistent alignment between personality and musical reasoning.

Model	PMA	CA	CJ	S	F
phi-3	3.18 / 36.5	4.19 / 87.4	2.57 / 8.7	2.42 / 2.5	3.51 / 49.7
mixtral	3.42 / 50.8	4.24 / 88.4	2.70 / 11.4	2.57 / 4.2	3.56 / 55.4
llama3	3.29 / 44.2	4.22 / 87.8	2.66 / 12.8	2.58 / 3.0	**3.64 / 63.6**
qwen2	3.07 / 32.6	4.02 / 82.6	2.46 / 9.8	2.43 / 3.4	3.47 / 46.8
olmo2	3.11 / 39.6	4.15 / 85.6	2.48 / 7.4	2.35 / 1.4	3.40 / 42.2
gemma2	**3.58 / 63.6**	**4.43 / 92.4**	**2.84 / 19.8**	**2.58 / 4.0**	3.56 / 56.6

Table 3: Mean (M) and % of high-quality (**%H**) teacher ratings (≥ 4) across five reasoning dimensions (Reported M/%H): PMA, CA, CJ, Specificity, and Fluency. Full table in Table 16.

`llama-3-8B` excels in Fluency, and `mixtral-7b` leads in Specificity. These trends reaffirm that larger models are generally more capable of producing coherent, human-aligned justifications, even when overall accuracy differences are moderate.

6.4 Ablation Study

Variant	Acc	F1	%Imp
Verbalized	**0.3521**	**0.3489**	**8.04**
CoT (Maj. Vote)	0.2874	0.2755	0.70
CoT	0.2872	0.2752	0.67
Baseline	0.2721	0.2685	–
Mis-verbalized	0.2282	0.2287	-3.98

Table 4: Inference-time ablation on `LLaMA-3-8B` for Task 2 (Chord Matching). Verbalized prompting shows the highest gains. See full results in Appendix 17.

We conduct an ablation to assess the impact of different reasoning strategies and training configurations on benchmark performance. Table 17 separates results into two blocks: *inference-only* and *trained model* settings, using the LLaMA-3-8B and Mistral-7B architectures. In inference mode, we vary prompting strategies (baseline, CoT, verbalization, etc.), while in training mode, we isolate the effect of verbalization by toggling it in merged models.

7 Conclusion

This work bridges a novel intersection between personality modeling and symbolic music reasoning by evaluating whether LLMs can align personality traits with musical preferences through a modular cross-domain benchmark. We evaluate fine-tuning and inference-time strategies, finding that while trait-specific training aids identification, verbalization-based prompting significantly improves chord alignment without supervision. Our reasoning evaluation further shows that larger models generate fluent, personality-consistent justifications. These findings open a new direction in evaluating LLMs' capacity to reason about identity

through language and music.

Limitations

While our benchmark and methods demonstrate promising results in aligning personality traits with musical structure, several limitations remain:

1. Limited Real-World Data. Our benchmark relies on synthetic contexts rather than real-world conversations or interviews about music preferences. While we mitigate this through careful prompt engineering and sampling, grounded datasets involving real individuals would provide more robust evaluation.

2. Symbolic Music Bias. Chord progression alignment captures only one facet of musical structure. Other dimensions such as rhythm, melody, or production aesthetics are not modeled, potentially underrepresenting a user's full musical identity.

3. Model Pretraining Bias. Open-source models vary widely in pretraining corpora and musical exposure, affecting baseline performance on symbolic reasoning tasks. Some gains may be due to prior exposure to musical texts or specific chord forms.

4. Limited Generalization to Other Domains. While our focus is on personality-music reasoning, it remains to be seen whether similar verbalization or alignment techniques generalize well to other symbolic or creative domains (e.g., art, literature).

Ethical Considerations

Our benchmark relies on synthetic data generated by GPT-4o, which may carry subtle biases from its pretraining. While personality and genre cues are indirectly embedded, unintended stereotypes could emerge. The benchmark is for research use only and is not intended for sensitive applications like diagnosis or profiling. All annotations were model-generated or anonymized, and no personal data was used. The dataset will be released under the CC BY-NC 4.0 license for non-commercial research use.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback on our manuscript. We are also grateful to the human judges who participated in evaluating the LLM-generated synthetic data. The human evaluation team was led by Anushka Tiwari (PhD student, Computational and Data-Enabled Sciences, University at Buffalo), and a small team of graduate students within the University at Buffalo research group. We had external contributions from Anisha Gupta and Shruti Kumar.

We gratefully acknowledge use of the research computing resources of the Empire AI Consortium, Inc., supported by Empire State Development of the State of New York, the Simons Foundation, and the Secunda Family Foundation (Bloom et al., 2025).

This work was partly supported by the National Science Foundation (NSF) under Grant No. 2214070: *III: Small: Purposeful Conversational Agents based on Hierarchical Knowledge*. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the NSF.

References

AI@Meta. 2024. Llama 3 model card.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2023. What learning algorithm is in-context learning? investigations with linear models. *Preprint*, arXiv:2211.15661.

Danny Azucar, Davide Marengo, and Michele Settanni. 2018. Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and individual differences*, 124:150–159.

David Radford Bakker and Frances Heritage Martin. 2015. Musical chords and emotion: Major and minor triads are processed for emotion. *Cognitive, Affective, & Behavioral Neuroscience*, 15:15–31.

Stacie Bloom, Joshua C. Brumberg, Ian Fisk, Robert J. Harrison, Robert Hull, Melur Ramasubramanian, Krystyn Van Vliet, and Jeannette Wing. 2025. Empire AI: A new model for provisioning AI and HPC for academic research in the public good. In *Practice and Experience in Advanced Research Computing (PEARC '25)*, page 4, Columbus, OH, USA. ACM.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *NeurIPS 2018*.

Graham Caron and Shashank Srivastava. 2023. Manipulating the perceived personality traits of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2370–2386, Singapore. Association for Computational Linguistics.

Raymond B Cattell and Heather E P. Cattell. 1995. Personality structure and the new fifth edition of the

16pf. *Educational and Psychological Measurement*, 55(6):926–937.

Tomas Chamorro-Premuzic and Adrian Furnham. 2007. Personality and music: Can traits explain how people use music in everyday life? *British journal of psychology*, 98(2):175–185.

Yong-Hun Cho, Hyunki Lim, Dae-Won Kim, and In-Kwon Lee. 2016. Music emotion recognition using chord progressions. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 002588–002593. IEEE.

Yuval Cohen, Hana Ornoy, and Baruch Keren. 2013. Mbti personality types of project managers and their success: A field survey. *Project Management Journal*, 44(3):78–87.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ondřej Dušek and Filip Jurčíček. 2016. A context-aware natural language generator for dialogue systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 185–190, Los Angeles. Association for Computational Linguistics.

Bruce Ferwerda, Marko Tkalcic, and Markus Schedl. 2017. Personality traits and music genres: What do people prefer to listen to? In *Proceedings of the 25th conference on user modeling, adaptation and personalization*, pages 285–288.

Maya B. Flannery and Matthew H. Woolhouse. 2021. Musical preference: Role of personality and music-related acoustic features. *Music & Science*, 4:20592043211014014.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music transformer. *Preprint*, arXiv:1809.04281.

Yu-Siang Huang and Yi-Hsuan Yang. 2020. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1180–1188, New York, NY, USA. Association for Computing Machinery.

Gregory M Hurtz and John J Donovan. 2000. Personality and job performance: the big five revisited. *Journal of applied psychology*, 85(6):869.

Pegah Jandaghi, Xianghai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2024. Faithful persona-based conversational dataset generation with large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15245–15270, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Patrik N. Juslin. 2010. *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford University Press.

Spyridon Kantarelis, Konstantinos Thomas, Vassilis Lyberatos, Edmund Dervakos, and Giorgos Stamou. 2024. Chordonomicon: A dataset of 666,000 songs and their chord progressions. *Preprint*, arXiv:2410.22046.

Tomohito Kasahara, Daisuke Kawahara, Nguyen Tung, Shengzhe Li, Kenta Shinzato, and Toshinori Sato. 2022. Building a personalized dialogue system with prompt-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 96–105, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Satoshi Kawase. 2024. Is happier music groovier? the influence of emotional characteristics of musical chord progressions on groove. *Psychological Research*, 88(2):438–448.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *Preprint*, arXiv:1909.05858.

Carol L Krumhansl. 1995. Music psychology and music theory: Problems and prospects. *Music Theory Spectrum*, pages 53–80.

Tiziano Labruna, Sofia Brenna, and Bernardo Magnini. 2024. Dynamic task-oriented dialogue: A comparative study of llama-2 and bert in slot value generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 358–368, St. Julian's, Malta. Association for Computational Linguistics.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. 2025. Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics.

In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8397–8437, Albuquerque, New Mexico. Association for Computational Linguistics.

Fei Liu, Julien Perez, and Scott Nowson. 2016. A recurrent and compositional model for personality trait recognition from short texts. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 20–29, Osaka, Japan. The COLING 2016 Organizing Committee.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, and Jiebo Luo. 2024. LLM-rec: Personalized recommendation via prompting large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 583–612, Mexico City, Mexico. Association for Computational Linguistics.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

Robert R. McCrae and Paul T. Costa Jr. 1999. *A Five-Factor theory of personality.*, pages 139–153. Handbook of personality: Theory and research, 2nd ed. Guilford Press, New York, NY, US.

Sayantan Pal, Souvik Das, Rohini Srihari, Jeff Higginborham, and Jenna Bizovi. 2024. Empowering AAC users: A systematic integration of personal narratives with conversational AI. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 12–25, Miami, Florida, USA. Association for Computational Linguistics.

Sayantan Pal, Souvik Das, and Rohini K. Srihari. 2025. Beyond discrete personas: Personality modeling through journal intensive conversations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7055–7074, Abu Dhabi, UAE. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Sougata Saha, Souvik Das, and Rohini Srihari. 2022. Stylistic response generation by controlling personality traits and intent. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 197–211, Dublin, Ireland. Association for Computational Linguistics.

Sandra Sanchez-Roige, Joshua C Gray, James MacKillop, C-H Chen, and Abraham A Palmer. 2018. The genetics of human personality. *Genes, Brain and Behavior*, 17(3):e12439.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *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)*, pages 1702–1723.

Christian Simon, Masoud Faraki, Yi-Hsuan Tsai, Xiang Yu, Samuel Schulter, Yumin Suh, Mehrtash Harandi, and Manmohan Chandraker. 2022. On Generalizing Beyond Domains in Cross-Domain Continual Learning . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9255–9264, Los Alamitos, CA, USA. IEEE Computer Society.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Alexander Spangher, Tenghao Huang, Philippe Laban, and Nanyun Peng. 2025. Creative planning with language models: Practice, evaluation and applications. In *Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 1–9, Albuquerque, New Mexico. Association for Computational Linguistics.

John Thickstun, Zaid Harchaoui, and Sham Kakade. 2017. Learning features of music from scratch. *Preprint*, arXiv:1611.09827.

Danqing Wang, Kevin Yang, Hanlin Zhu, Xiaomeng Yang, Andrew Cohen, Lei Li, and Yuandong Tian. 2024. Learning personalized alignment for evaluating open-ended text generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13274–13292, Miami, Florida, USA. Association for Computational Linguistics.

Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. 2023. RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*,

pages 852–861, Hong Kong, China. Association for Computational Linguistics.

Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, Liumeng Xue, Ziyang Ma, Qin Liu, Tianyu Zheng, Yizhi Li, Yinghao Ma, Yiming Liang, Xiaowei Chi, Ruibo Liu, and 13 others. 2024. ChatMusician: Understanding and generating music intrinsically with LLM. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6252–6271, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

A PMA Bench Details

A.1 Context Generation Prompt

The following prompt was used with GPT-4o to generate realistic personality-conditioned musical scenarios. Each instance was conditioned on a PF16 descriptor set $\mathcal{H}_t$, an aligned genre g, and a domain-subdomain pair (d, s). No trait or genre was explicitly named in the output.

> You specialize in crafting realistic, engaging scenarios that subtly reflect individuals' personality traits and musical preferences without explicitly mentioning them.
>
> Generate a short, structured context (50–70 words) that describes an individual's character and preferred music style in a natural, real-world scenario. Do not explicitly mention the personality trait or the music genre.
>
> **Instructions:**
> 1. The individual exhibits qualities such as: {descriptor list}.
> 2. Their musical preference subtly aligns with a style reminiscent of {genre}.
> 3. Incorporate contextual details from the domain "{domain}" and subdomain "{subdomain}".
>
> Please produce a plain text description.

A.2 Qualitative Analysis of Personality-Musical Alignment

We present illustrative examples from our benchmark to highlight how personality traits and musical preferences are naturally encoded in context and whether models can reason through this mapping. GPT-4o rated each example below along five dimensions: Personality Alignment, Musical Coherence, Naturalness, Implicitness, and Specificity. Scores range from 1–5.

Example 1: Joyful Domesticity (Pop – Liveliness)

> *Sophie twirled into the living room, her laughter echoing through the house as she playfully swung her younger brother around in a dance. The sound of catchy, upbeat tunes streamed from the kitchen, where her mom clapped along while preparing dinner. Sophie loved these impromptu moments, turning everyday routines into lively celebrations, her energy infectious, setting a joyful rhythm in their home.*

This context captures a high-liveliness personality through spontaneous, energetic behavior and social warmth. The selected trait, *animated*, and the genre-aligned chord progression (`C G Am F`, common in uplifting pop music) match the tone and rhythm of the narrative. GPT-4o rated this scenario with perfect scores across Personality Alignment, Musical Coherence, and Naturalness, noting slightly lower scores (4) for Implicitness and Specificity, indicating the cues were vivid but somewhat overt.

Example 2: Reflective Absorption (Jazz – Abstractedness)

> *At the monthly community art exhibit, Ella wandered, her eyes lingering on abstract canvases. Lost in thought, she barely noticed the faint strains of saxophone drifting from the corner stage. As locals gathered, swapping stories of the night, she found herself nodding in rhythm to the seamless improvisation, her fingers unconsciously sketching patterns on the program in hand.*

This example showcases introspective, imaginative tendencies associated with high Abstractedness. The trait *absentminded* aligns with the narrative's detached tone, while the chosen genre (Jazz) and progression (`Dm7 Em7 A7 Dm7`) reinforce the reflective, improvisational mood. All five reasoning dimensions received top scores, validating both the subtle personality encoding and genre integration.

A.3 Evaluation Dimensions for Context Quality

Human judges evaluated a randomly sampled subset of synthetic contexts to assess their quality and faithfulness to intended personality-music mappings. Each context was independently rated on a 5-point Likert scale (1 = poor, 5 = excellent) across five qualitative dimensions listed in Table 5. Annotators were instructed to read each context carefully and assign scores reflecting how well it satisfied each criterion.

Instruction to Human Annotators:

For each given context, please rate it from **1 (poor)** to **5 (excellent)** on the following five aspects:

1. Personality Alignment: Does the context express the intended PF16 trait?
2. Musical Coherence: Is the implied musical preference plausible for that trait?
3. Naturalness: Is the text fluent and realistic in everyday language?
4. Implicitness: Are the cues subtle rather than explicit?
5. Specificity: Does the description include concrete, non-generic detail?

Provide your numeric ratings only; no written explanations are required.

Dimension	Explanation
Personality Alignment (PA)	Measures how well the context reflects the high-polarity descriptors of the intended PF16 trait.
Musical Coherence (MC)	Evaluates whether the scenario plausibly implies a genre preference aligned with the given trait.
Naturalness (N)	Assesses the realism and fluency of the scenario in everyday language.
Implicitness (I)	Captures whether personality and genre cues are subtly embedded rather than explicitly stated.
Specificity (S)	Judges the level of concrete detail and avoidance of generic phrasing.

Table 5: Qualitative dimensions used to evaluate sampled contexts across traits.

B Experiments

B.1 Training Strategy Detailed

T2: Chord Progression Alignment. This task involves symbolic reasoning over music theory, conditioned on inferred personality traits. We approach it in two steps:

(i) Chord-LM Pre-training (Unsupervised): We first train a causal decoder-only model on 30K+ chord sequences from the Chordonomicon corpus.

Attrib	PC	SPC	BST	JIC
# of Conversations	18,878	10,905	6,808	418,476
Tot. # of Turns	120,361	152,945	44,959	3,347,808
Avg. # of Turns	6.38	14.03	6.60	8.00
Tot. # of Utterances	259,600	310,874	89,918	6,695,616
Avg. Utt. (conv)	13.75	28.51	13.21	16.00
Avg. Words (u)	11.24	8.75	13.46	15.48
Avg. Conv. Length (w)	154.56	249.53	177.83	247.61
Longest Conv. (u)	49	117	28	16
Shortest Conv. (u)	11	6	4	16
Longest Conv. (w)	477	637	422	581
Shortest Conv. (w)	41	60	24	16
Avg. Topic Consistency (u)	0.50	0.57	0.55	0.53
Avg. Semantic Similarity (u)	0.31	0.39	0.36	0.36

Table 6: Comparison of various datasets across several attributes. Here PC is Persona Chat, SPC is Synthetic Persona Chat, BST is Blended Skill Talk, (u) means per utterance, and (w) means per word.

Given a sequence $C = (c_1, c_2, \ldots, c_n)$, the model is trained to maximize the likelihood of predicting the next chord:

$$\mathcal{L}_{\text{chord}}(\theta) = -\sum_{i=2}^{n} \log p(c_i | c_{1:i-1}; \theta).$$

(ii) Genre-Chord Translation (Supervised): We treat genre-chord reasoning as a translation task. The model is trained in both directions: (a) generate chords $c_{1:n}$ from a genre label g and related attributes (e.g., subgenre, era), (b) generate the most likely genre label g given a chord progression. This bidirectional fine-tuning improves both generation and reverse classification:

$$\mathcal{L}_{\text{genre}}(\theta) = -\sum_{i=1}^{n} \log p(c_i | c_{1:i-1}, g; \theta)$$

$$\mathcal{L}_{\text{genre-inv}}(\theta) = -\log p(g | c_{1:n}; \theta)$$

This formulation allows the model to learn the genre-conditioned musical structure and improves generalization to unseen chord forms and personality mappings.

Model Merging via SVD. To combine the personality and chord reasoning capabilities, we merged the adapters trained on T1 and T2 using SVD-based adapter merging from the PEFT library. This method approximates a shared low-rank subspace across LoRA modules, allowing efficient integration.

B.1.1 Personality Trait Identification - Dataset stats.

Table 6 shows the dataset statistics.

B.1.2 Training Arguments and GPU

All the models were trained on a single A100 80 GB. Table 7 shows the Training Args used to train

all the Models. The batch size default was set to 4 but was reduced to 2 when A100 80GB was unavailable (used A100 40 GB). LoRA hyper parameters (r = 64, $alpha$ = 16, dropout = 0.1) were most significant.

Argument	Value
lora_r	64
lora_alpha	16
lora_dropout	0.1
bf16	True
learning_rate	2.0e-05
gradient_accumulation_steps	128
gradient_checkpointing	True
logging_strategy	Steps
logging_steps	1
save_strategy	Steps
save_steps	100
eval_steps	100
per_device_train_batch_size	4
per_device_eval_batch_size	4
max_seq_length	2048
lr_scheduler_type	Cosine
early_stopping_patience	4

Table 7: Trainer Arguments

B.2 Inference Pipeline

B.2.1 Verbalization Prompt

Prompt Templates for Verbalization Pipeline

Step 1: Personality-to-Genre Prompt

> You are a music expert and songwriter who knows how personality drives musical taste.
>
> Below is a mapping of personality traits to genres:
> - Warmth: Pop, Country, Gospel
> - Reasoning: Jazz, Alternative, Classical ...
>
> Context: {context}
>
> Task: In no more than 30 words, identify the dominant personality traits from the context and recommend 2–3 genres they are most likely to enjoy, based on the mapping above.

Step 2: Genre-to-Chord Prompt

> You are a music expert and songwriter who knows how personality influences musical preferences and harmonic structure.
>
> Below is a mapping of genres to music theory knowledge:
> - Pop: Uses the I–V–vi–IV major-key cycle with simple triads. Uplifting, highly danceable.
> - Rock: Drives on I–IV–V or I–V–vi–IV power chords. Bold, energetic tone.
> - Jazz: Rich in 7th chords and ii–V–I cadences. Complex, smooth. ...
>
> Context: {context}
>
> Personality and Genre Knowledge: {predicted trait and genre rationale from Step 1}
>
> Question: Which chord progression best matches the user's music preference?
>
> Options: A: {prog1} B: {prog2} C: {prog3} D: {prog4}
>
> Please analyze how the context, predicted genre, and music theory knowledge align to justify your selection.
>
> Limit your answer to 100 words. Conclude with ONLY the letter of your choice inside <answer> tags. Example: <answer>C</answer>.

C Evaluation Strategy

C.1 Reasoning Evaluation Dimensions

To evaluate the quality of model-generated reasoning in Task 2 (chord progression alignment), we define five key dimensions, each rated on a 1–5 Likert scale. These dimensions are adapted from prior work on explanation quality, natural language generation, and interpretability in NLP models.

1. **Personality-Musical Alignment:** Measures how well the explanation connects the inferred personality trait to the selected chord progression or genre. Inspired by trait-grounded reasoning evaluation in personality modeling (Mairesse et al., 2007) and controllable generation (Keskar et al., 2019).

2. **Chordal Appropriateness:** Assesses whether the described chord progression is musically coherent with the inferred genre or emotional tone. Draws from music theory alignment evaluation and symbolic music generation benchmarks (Krumhansl, 1995).

3. **Causal Justification:** Evaluates whether the explanation includes a clear, logical cause-effect rationale for why the chosen progression fits the described personality or mood. Builds on explanation plausibility and justification work in NLI and commonsense reasoning (Camburu et al., 2018; Rajani et al., 2019).

4. **Specificity:** Measures the level of detail in the reasoning, including references to actual chord functions, emotional textures, or subtraits. Inspired by dialogue specificity metrics (See et al., 2019).

5. **Fluency and Clarity:** Captures the grammatical correctness, coherence, and readability of the explanation.

Each reasoning sample is first evaluated by a teacher model (GPT-4o) that generates a gold explanation from the context and ground-truth label. It then scores the student model's explanation on each dimension, optionally providing a short justification.

D Results and Discussions

D.1 Verbalization Success and NEI Trends

Verbalized prompting substantially reduces NEI rates for most models, notably dropping from 187 to 1 for `OLMo`, indicating improved grounding. While `mixtral` shows a minor NEI increase, which indicates the model is overthinking, models like `phi-3`, `llama3`, and `gemma2` remain consistently low, reflecting better alignment and control.

E Miscellaneous

Contains tables and extra information

Genre	Description
Alternative	A broad genre encompassing experimental and non-mainstream rock/pop styles.
Blues	A soulful genre rooted in African American history, emphasizing emotion and improvisation.
Classical	Western art music tradition featuring orchestral, chamber, and solo compositions.
Country	Narrative-driven music with acoustic instrumentation and Southern U.S. cultural roots.
Electronic	Synthesized, beat-driven music spanning ambient, techno, and dance substyles.
Folk	Acoustic storytelling music rooted in cultural traditions and social commentary.
Gospel	Spiritually expressive music grounded in Christian themes and vocal harmony.
Hip-Hop	Rhythm-centric genre combining rap, beats, and urban cultural expression.
Indie	Independent music often characterized by artistic freedom, introspection, and lo-fi sounds.
Jazz	Improvisational and harmonic genre blending swing, blues, and complex instrumentation.
Metal	High-intensity genre marked by distorted guitars, aggression, and darker themes.
New Age	Meditative, atmospheric music focused on relaxation, spirituality, and soundscapes.
Pop	Mainstream music emphasizing catchy melodies and broad appeal across demographics.
Punk	Raw, fast-paced genre with rebellious lyrics and stripped-down instrumentation.
R&B	Rhythm and blues style blending soulful vocals with groovy and romantic instrumentation.
Reggae	Jamaican-born genre known for offbeat rhythms and themes of resistance and peace.
Reggaeton	Latin fusion of reggae, hip-hop, and dancehall with rhythmic Spanish vocals.
Rock	Guitar-driven genre ranging from classic rock to modern subgenres like alt-rock.
Soul	Emotionally rich music emphasizing vocal power and themes of love and struggle.

Table 8: Descriptions of the 19 musical genres used in the benchmark. These span a range of emotional, structural, and cultural attributes relevant to personality alignment.

Trait	High Descriptors	Low Descriptors	Alike Genres	Different Genres
Warmth	Warm, outgoing, attentive to others, kindly, easygoing	Impersonal, distant, cool, detached, aloof	Pop, Country, Gospel	Rock, Electronic, Metal
Reasoning	Abstract-thinking, intelligent, bright, fast-learner	Concrete-thinking, less intelligent, unable to abstract	Jazz, Alternative, Classical	Pop, Country, Hip-Hop
Emotional Stability	Emotionally stable, adaptive, mature	Reactive, changeable, easily upset	New Age, Reggae, Folk	Metal, Punk, Rock
Dominance	Dominant, assertive, competitive, bossy	Deferential, cooperative, submissive	Rock, Metal, Punk	Pop, Country, Gospel
Liveliness	Lively, animated, spontaneous, cheerful	Serious, restrained, taciturn	Pop, Hip-Hop, Electronic	Jazz, Blues, Folk
Rule-Consciousness	Rule-bound, dutiful, moralistic	Expedient, nonconforming, self-indulgent	Classical, Jazz, R&B	Hip-Hop, Punk, Electronic
Social Boldness	Socially bold, venturesome, uninhibited	Shy, timid, hesitant	Hip-Hop, Electronic, Reggaeton	Jazz, Classical, New Age
Sensitivity	Sensitive, sentimental, refined	Tough-minded, objective, rough	Soul, Gospel, Blues	Metal, Punk, Rock
Vigilance	Vigilant, skeptical, oppositional	Trusting, unsuspecting, accepting	Alternative, Metal, Punk	Pop, Country, Reggae
Abstractedness	Imaginative, absentminded, absorbed in ideas	Grounded, practical, conventional	Jazz, New Age, Indie	Country, Hip-Hop, Reggae
Privateness	Discreet, shrewd, diplomatic	Forthright, genuine, open	Classical, Indie, Folk	Pop, Hip-Hop, Electronic
Apprehension	Apprehensive, insecure, self-blaming	Self-assured, secure, confident	Blues, Alternative, Indie	Pop, Country, Gospel
Openness to Change	Experimental, analytical, freethinking	Traditional, conservative	Electronic, Alternative, Reggaeton	Country, Blues, Gospel
Self-Reliance	Self-reliant, solitary, individualistic	Group-oriented, dependent	Indie, Alternative, Metal	Pop, Country, Reggae
Perfectionism	Perfectionistic, organized, self-disciplined	Tolerates disorder, lax, impulsive	Classical, Jazz, R&B	Hip-Hop, Punk, Electronic
Tension	Tense, driven, frustrated	Relaxed, tranquil, composed	Rock, Hip-Hop, Metal	New Age, Folk, Reggae

Table 9: PF16 traits, their descriptors (high and low range), and genre associations.

Genre	Chord Progressions	Dancability	Mood
Pop	C Am F G C G Am F Am F C G C Am Dm G	8	Uplifting
Rock	G D Em C D A Bm G E B C#m A A E F#m D	7	Energetic
Jazz	Dm7 G7 Cmaj7 Am7 Dm7 G7 Cmaj7 Dm7 Em7 A7 Dm7 Gm7 C7 Fmaj7	6	Sophisticated
Hip-Hop	Am7 Dm7 G7 Cmaj7 Am F C G Em Am Dm G Am7 Em7 Dm7 G7	9	Groovy
Metal	Em C Am B Em G D A Em D C B Em C B A	4	Intense

Table 10: Representative chord progressions for five popular genres, along with dancability scores (1–10) and associated mood labels. These examples demonstrate genre-specific harmonic structures used in our benchmark.

Domain	Subdomains
Social Dynamics	Friendships, Romantic Relationships, Community Engagement
Personal Development	Self-Improvement, Emotional Resilience, Identity Exploration
Lifestyle & Routine	Daily Routines, Leisure & Recreation, Health & Wellness
Professional Environment	Workplace Culture, Career Growth, Leadership Dynamics
Creative Expression	Artistic Pursuits, Music Exploration, Literary Interests
Cultural Engagement	Heritage & Tradition, Global Influences, Local Community
Urban Living	City Life, Public Transit, Neighborhood Vibes
Technology & Media	Digital Connectivity, Social Media Trends, Innovative Consumption
Family & Home	Domestic Life, Family Bonds, Home Environment
Adventure & Exploration	Travel Experiences, Outdoor Activities, Culinary Journeys

Table 11: Synthetic domains and subdomains used for context generation.

Model	PA		MC		N		I		S	
	M	SD	M	SD	M	SD	M	SD	M	SD
GPT-4o	4.537	0.631	4.531	0.617	4.544	0.504	4.147	0.513	4.309	0.462
Judge 1	4.503	0.657	4.509	0.642	4.516	0.564	4.150	0.572	4.297	0.533
Judge 2	4.513	0.652	4.528	0.632	4.522	0.542	4.134	0.595	4.303	0.535
Judge 3	4.487	0.647	4.522	0.642	4.506	0.565	4.141	0.555	4.322	0.518

Table 12: Mean (M) and standard deviation (SD) of the five evaluation dimensions for each model/variant.

Metric	Personality Alignment	Musical Coherence	Naturalness	Implicitness	Specificity
Pearson_LLM_J1	0.923	0.919	0.881	0.852	0.833
Spearman_LLM_J1	0.912	0.917	0.910	0.862	0.859
Pearson_LLM_J2	0.918	0.924	0.860	0.858	0.847
Spearman_LLM_J2	0.905	0.927	0.877	0.866	0.874
Pearson_LLM_J3	0.912	0.918	0.877	0.839	0.837
Spearman_LLM_J3	0.912	0.908	0.905	0.842	0.858
Pearson_J1_J2	0.843	0.816	0.776	0.785	0.693
Spearman_J1_J2	0.833	0.834	0.830	0.796	0.751
Pearson_J1_J3	0.879	0.864	0.828	0.710	0.730
Spearman_J1_J3	0.867	0.864	0.865	0.720	0.783
Pearson_J2_J3	0.845	0.861	0.771	0.690	0.764
Spearman_J2_J3	0.846	0.868	0.833	0.695	0.811
ICC	0.969	0.968	0.951	0.935	0.933

Table 13: Pearson(P), Spearman(S), and Intraclass Correlation Coefficients (ICC) between the language model (LM: GPT-4o) and human annotations for the 5 dimensions. High ICC values indicate strong agreement between LLM and human evaluations.

Model	Params.(B)	Config.	Metric							
			T1_acc	T1_f1	%Imp.(T1_f1)	T1_NEI	T2_acc	T2_f1	%Imp.(T2_f1)	T2_NEI
Open Source Models ($\leq$10B)										
phi-3-mini-4k-ins	3.8	baseline	0.8597	0.8599	–	0	0.2358	0.2129	–	0
		cot	0.8649	0.8641	0.4200	0	0.2500	0.2264	1.3500	0
		cot_maj_vote	0.8681	0.8653	0.5400	0	0.2543	0.2275	1.4600	0
		verbalized	–	–	–	–	0.3462	0.3466	13.3700	2
mixtral-7b-v0.3-instruct	7	baseline	0.8240	0.8240	-	0	0.2448	0.2018	–	1
		cot	0.8104	0.8108	-1.3200	1	0.2462	0.2131	1.1300	2
		cot_maj_vote	0.8123	0.8128	-1.1200	0	0.2498	0.2173	1.5500	0
		verbalized	–	–	–	–	0.3420	0.3430	14.1200	21
llama-3-8Bb-instruct	8	baseline	0.8615	0.8620	–	0	0.2721	0.2685	–	0
		cot	0.8420	0.8426	-1.9400	0	0.2872	0.2752	0.6700	2
		cot_maj_vote	0.8425	0.8428	-1.9200	0	0.2874	0.2755	0.7000	0
		verbalized	–	–	–	–	0.3521	0.3489	8.0400	0
Llama-3.1-Tulu-3.1-8B	8	baseline	0.8632	0.8641	–	0	0.2576	0.2157	–	
		cot	0.8719	0.8716	0.7500	0	0.2681	0.2288	1.3100	8
		cot_maj_vote	0.8723	0.8777	1.3600	0	0.2657	0.2254	0.9700	8
		verbalized	–	–	–	–	0.3451	0.3372	12.1500	0
deepseek_llama_3.1_8B	8	thinking	0.8125	0.8188	-4.5300	45	0.2896	0.2893	7.3600	17
Qwen2-7B-Instruct	7	baseline	0.8177	0.8185	-	0	0.2549	0.2294	–	0
		cot	0.7858	0.7860	-3.9707	0	0.2812	0.2700	4.0600	4
		cot_maj_vote	0.7834	0.7840	-4.2150	0	0.2812	0.2700	4.0600	4
		verbalized	–	–	–	–	0.3028	0.2913	6.1900	9
OLMo-2-1124-7B-Instruct	7	baseline	0.8340	0.8323	–	0	0.2524	0.2318	–	0
		cot	0.8149	0.8150	-2.0786	0	0.2476	0.1839	-4.7900	187
		cot_maj_vote	0.8128	0.8136	-2.2468	0	0.2466	0.1819	-4.9900	187
		verbalized	–	–	–	–	0.3097	0.2851	5.3300	1
gemma2-9b-it	9	baseline	**0.8990**	**0.8988**	–	0	0.2840	0.2657	–	0
		cot	0.8573	0.8679	-3.4379	69	0.3059	0.2933	2.7600	4
		cot_maj_vote	0.8598	0.8685	-3.3712	46	0.3088	0.2967	3.1000	1
		verbalized	–	–	–	–	**0.4632**	**0.4628**	**19.7100**	0
Meta-Llama-3-70B-Instruct	70	baseline	0.8792	0.8858	–	43	0.3122	0.2984	–	0
Closed Source (commercial) Models										
gpt-3.5-turbo	N/A	baseline	0.8948	0.8950	–	1	0.2444	0.2333	–	142
gpt-4o-mini	N/A	baseline	**0.9142**	**0.9150**	–	5	0.3354	0.3358	–	0
gpt-4o	N/A	baseline	0.9007	0.9007	–	0	0.3660	0.3658	–	0
gpt-4.1-mini	N/A	baseline	0.9066	0.9066	–	0	0.3427	0.3391	–	0
gpt-4.1	N/A	baseline	0.9073	0.9074	–	0	**0.3750**	**0.3721**	–	0

Table 14: Performance of open-source and closed-source models across both tasks under inference-only settings. Baseline denotes the zero-shot prompting setup, while *cot*, *cot_maj_vote*, and *verbalized* represent guided prompting strategies. We report accuracy (Acc), macro F1 score (F1), improvement over baseline (%Imp), and Not Enough Information (NEI) rates for both personality trait identification (T1) and chord progression alignment (T2). Closed models (e.g., GPT-4.1) generally outperform open models, but verbalization provides substantial gains in T2 across most systems.

Model	Config	Metric							
		T1_acc	T1_f1	%Imp.(T1_f1)	T1_NEI	T2_acc	T2_f1	%Imp.(T2_f1)	T2_NEI
mixtral-7b-v0.3-instruct	baseline	0.8240	0.8240	-	0	0.2448	0.2018	-	1
	Ft. PC	0.8211	0.8214	-0.2600	0	0.2436	0.2003	-0.1500	1
	Ft. SPC	0.8256	0.8259	0.1900	0	0.2440	0.2013	-0.0500	0
	Ft. BST	0.8241	0.8236	-0.0400	0	0.2448	0.2018	0.0000	1
	Ft. JIC	0.8378	0.8379	1.3900	0	0.2573	0.1262	-7.5600	0
	Merged	0.8375	0.8375	1.3500	0	0.2531	0.1387	-6.3100	2
	Merged (V)	-	-	-	-	0.2437	0.2701	**6.8300**	572
llama-3-8Bb-instruct	baseline	0.8615	0.8620	-	0	0.2721	0.2685	-	0
	Ft. PC	0.8578	0.8586	-0.3400	0	0.2719	0.2682	-0.0300	0
	Ft. SPC	0.8618	0.8629	0.0900	0	0.2720	0.2689	0.0400	0
	Ft. BST	0.8579	0.8589	-0.3100	0	0.2708	0.2666	-0.1900	0
	Ft. JIC	**0.8858**	**0.8861**	**2.4100**	0	0.2615	0.2364	-3.2100	1
	Merged	0.8822	0.8834	2.1400	0	0.2633	0.2366	-3.1900	2
	Merged (V)	-	-	-	-	**0.3149**	**0.3124**	4.3900	29

Table 15: Supervised fine-tuning results for both tasks across LLaMA-3 and Mixtral models. JIC-based adapters consistently yield the best gains for Task 1 (personality reasoning). Verbalization (`Merged (V)`) continues to outperform others on Task 2 (chord reasoning)

Model	Metric														
	PMA			CA			CJ			S			F		
	M	SD	%H	M	SD	%H	M	SD	%H	M	SD	%H	M	SD	%H
phi-3-mini-4k-ins	3.183	0.777	36.52	4.185	0.738	87.36	2.565	0.678	8.71	2.424	0.558	2.53	3.511	0.548	49.72
mixtral-7b-v0.3-instruct	3.416	0.718	50.80	4.242	0.784	88.40	2.700	0.668	11.40	2.572	0.580	**4.20**	3.558	0.606	55.40
llama-3-8Bb-instruct	3.288	0.744	44.20	4.218	0.922	87.80	2.664	0.715	12.80	2.578	0.576	3.00	**3.642**	0.553	**63.60**
Qwen2-7B-Instruct	3.072	0.819	32.60	4.022	0.981	82.60	2.464	0.741	9.80	2.434	0.615	3.40	3.470	0.617	46.80
OLMo-2-1124-7B-Instruct	3.106	0.843	39.60	4.146	0.953	85.60	2.480	0.694	7.40	2.346	0.585	1.40	3.398	0.579	42.20
gemma2-9b-it	**3.580**	0.724	**63.60**	**4.430**	0.818	**92.40**	**2.844**	0.743	**19.80**	**2.584**	0.599	4.00	3.562	0.618	56.60

Table 16: Teacher-rated (Teacher is GPT-4o) reasoning quality across five dimensions: Personality-Musical Alignment (PMA), Chordal Appropriateness (CA), Causal Justification (CJ), Specificity (S), and Fluency & Clarity (F). Metrics include mean (M), standard deviation (SD), and percentage of high-quality ratings (%H) with a score $\geq$ 4.

Model	Variant	Task 1 (Personality ID)			Task 2 (Chord Matching)		
		Acc	F1	%Imp	Acc	F1	%Imp
Inference-Only							
LLaMA-3-8B	Verbalized	-	-	-	**0.3521**	**0.3489**	**8.04**
	CoT (majority vote)	0.8425	0.8428	-1.92	0.2874	0.2755	0.70
	CoT	0.8420	0.8426	-1.94	0.2872	0.2752	0.67
	Baseline	0.8615	0.8620	–	0.2721	0.2685	–
	Mis-verbalized	-	-	-	0.2282	0.2287	-3.98
Mistral-7B	Verbalized	-	-	-	**0.3420**	**0.3430**	**14.12**
	CoT (majority vote)	0.8123	0.8128	-1.12	0.2498	0.2173	1.55
	CoT	0.8104	0.8108	-1.32	0.2462	0.2131	1.13
	Baseline	0.8240	0.8240	–	0.2448	0.2018	–
	Mis-verbalized	-	-	-	0.2010	0.1854	-1.64
Training-Time Variants							
LLaMA-3-8B	Merged + Verbalized	-	-	-	**0.3149**	**0.3124**	**4.39**
	Merged (no verbalize)	0.8822	0.8834	2.14	0.2633	0.2366	-3.19
	Baseline	0.8615	0.8620	–	0.2721	0.2685	–
Mistral-7B	Merged + Verbalized	-	-	-	**0.2437**	**0.2701**	**6.83**
	Merged (no verbalize)	0.8375	0.8375	1.35	0.2531	0.1387	-6.31
	Baseline	0.8240	0.8240	–	0.2448	0.2018	–

Table 17: Ablation study comparing inference-only and fine-tuned configurations across two tasks. Verbalization consistently improves Task 2 (chord matching) performance. Merged models trained without verbalization show limited gains, highlighting the importance of prompt chaining and reasoning. Mis-verbalized means the intermediate knowledge about personality traits was randomly mapped to any genre.