

LLM Inferred Personality Traits for Calibrated Music Recommendation

Yuvisa Quispe Palomino
yuvisa@usp.br
Universidade de São Paulo
São Carlos, SP, Brazil

Paul Atauchi
paul.atauchi@usp.br
Universidade de São Paulo
São Carlos, SP, Brazil

Luan Soares de Souza
luanssouza@usp.br
Universidade de São Paulo
São Carlos, SP, Brazil

Dietmar Jannach
dietmar.jannach@aau.at
University of Klagenfurt
Klagenfurt, Austria

Marcelo Garcia Manzato
mmanzato@icmc.usp.br
Universidade de São Paulo
São Carlos, SP, Brazil

ABSTRACT

Recommender Systems (RS) help manage information overload by generating personalized recommendations from user data. However, feedback loops can introduce and amplify biases, such as popularity and exposure bias, systematically favoring certain items or groups. Calibration has been proposed as a post-processing strategy to align recommendation lists with user preferences while improving beyond-accuracy objectives such as diversity and fairness. In this work, we investigate personality-aware calibration for music recommendation when explicit personality traits are unavailable. We infer users' Big Five personality traits from music-consumption signals using Large Language Models (LLMs), specifically GPT-4.1-mini and Gemini-2.5-flash, and compare them with questionnaire-based personality data. Our results show that LLMs can approximate personality traits better than random and permutation baselines, with the best performance achieved by GPT using summarized profiles enriched with lyrical semantic and emotional features. Experiments in the music domain further show that the proposed approach can improve diversity and fairness in several configurations while maintaining competitive accuracy. Although gains are not uniform across all algorithms and calibration settings, the results suggest that LLM-inferred personality traits can provide useful signals for calibrated recommendation when explicit personality information is unavailable.

KEYWORDS

Personality Traits, Inferred Personality Traits, Big Five, Calibration, Fairness, Large Language Models (LLMs)

1 INTRODUCTION

In the era of information overload, Recommender Systems (RS) have become essential tools, helping users make decisions by filtering relevant content and being widely used across domains such as movies, music, games, and e-commerce [19].

RS primarily rely on historical interactions between users and items (e.g., music, videos, or products) [19]. Although these interactions allow the system to learn user preferences, the exposure mechanism of RS influences user behavior, where user interactions are continuously recorded and utilized to train the recommendation model and produce future recommendations. Thus, feedback loops not only introduce biases but also amplify them over time, leading to a “rich-get-richer” effect. For example, popular artists tend to be recommended more frequently than less popular ones, which increases their visibility and reinforces their popularity. In contrast, less popular artists receive limited exposure, not necessarily because of lower quality, but because the algorithm does not provide them with equal opportunities to be recommended [7, 21].

These issues contribute to broader problems, such as the formation of filter bubbles and reduced diversity in recommendations. As a result, user may be exposed to a limited range of content, restrict discovery, ultimately affecting fairness and user satisfaction [25]. To address these challenges, several bias mitigation strategies have been proposed that go “beyond-accuracy”, promoting the discovery of a wider range of items, which is essential for improving the long-term user experience. Among these approaches, model-agnostic techniques applied in the post-processing stage have gained increasing attention.

In particular, Calibrated Recommendations, introduced by [23], aim to align the distribution of recommended items with users' historical preferences. For example, if a user has interacted with 70% romance movies and 30% action movies, the recommendation list should ideally reflect a similar proportion. However, in practice, recommender systems often deviate from this distribution due to their strong focus on accuracy.

Calibration relies on users' past preferences to guide the recommendation process, ensuring that the resulting list reflects the distribution of their interests. When a user exhibits a diverse profile, this diversity is preserved in the recommendations, avoiding overly narrow suggestions and providing a more balanced and personalized representation of the user's preferences [9]. In this sense, calibration goes beyond relevance by explicitly considering how well recommendations reflect the overall user profile.

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2026). Lavras/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2026.

© 2026 SBC – Brazilian Computing Society.
ISSN 2966-2753

Moreover, prior work has shown that calibration can enhance diversity, expand item coverage, and introduce more novel content [14]. It has also been applied to mitigate biases in recommender systems – particularly popularity bias – by promoting a more balanced exposure of items beyond the most frequently recommended ones [1, 22].

Despite these advantages, traditional calibration approaches estimate the target distribution from the genres represented in the user's interaction history [23], which may not fully capture their underlying preferences, especially in cases of sparse or limited data. To address this limitation, [16] proposed incorporating personality traits as a complementary source of information to the calibration process. The proposed approach combines users' historical interactions with personality traits to capture preference patterns that may not be fully reflected in past behavior. By integrating both behavioral and personality information, the calibration target provides a more comprehensive representation of users' preferences instead of relying exclusively on the diversity observed in their historical consumption.

However, personality-aware calibration approaches depend on explicit personality information collected for example through questionnaires [16]. Although questionnaires provide structured and interpretable user traits, they are costly to collect, may interrupt the user experience, and are rarely available in operational recommender systems. This limits the practical adoption of personality-aware calibration, especially in domains where user profiles are sparse or where only implicit behavioral signals are available.

In this context, this work investigates whether Large Language Models (LLMs) can infer useful personality signals from music-consumption data and whether these inferred traits can support personality-aware calibration. In particular, we estimate users' Big Five personality traits [11], from two types of music-based representations: full listening histories composed of songs, artists, and genres, and summarized user profiles enriched with semantic and emotional features extracted from song lyrics. The inferred traits are then incorporated into a calibrated recommendation framework and compared against questionnaire-based personality traits. This work is guided by two research questions:

- RQ1:** *How well can LLMs predict users' personality traits from their musical preferences and the emotional content of song lyrics?*
- RQ2:** *How consistent is personality-aware calibration when questionnaire-based personality traits are replaced by LLM-inferred traits in music recommendation models?*

2 RELATED WORK

Personality has been widely investigated in recommender systems as a complementary signal to enhance user modeling [15, 24]. By capturing stable behavioral characteristics, personality-aware approaches enable the modeling of long-term preferences beyond interaction data. Empirical evidence shows that personality traits are significantly associated with user behaviors such as engagement, content preferences, and

rating patterns [13]. Moreover, prior work has highlighted that incorporating personality traits can improve recommendation performance, particularly in cold-start and data sparsity scenarios [10]. These findings suggest that personality can act as a complementary signal to interaction history, especially when historical data are sparse, narrow, or biased.

This is particularly relevant in the context of calibration and fairness, RS significantly influence item exposure, raising concerns about bias and fairness. Traditional algorithms optimized for accuracy often lead to popularity bias, where popular items are over-represented while less popular ones receive limited exposure, negatively affecting diversity and user experience [1]. Calibration has been proposed as a post-processing strategy to address this issue by aligning recommendation outputs with users' preference distributions. Steck introduced a calibration method based on Kullback–Leibler divergence and re-ranking, aiming to reduce the overrepresentation of dominant interests while preserving relevance [23]. Recent work has extended calibration to explicitly address fairness and bias. For example, a two-stage calibration approach balances relevance and fairness by adjusting item exposure [22], while other studies show that calibration can improve diversity and mitigate popularity bias [20].

However, traditional calibration relies primarily on interaction data, limiting its effectiveness for users with narrow or biased profiles. To overcome this limitation, recent approaches incorporate additional user information, such as personality traits, to estimate latent preferences and promote more balanced recommendations [16]. However, this strategy still depends on the availability of personality traits, which are usually collected through questionnaires and may not be available. So recent studies have investigated the use of transformer-based models to infer user personality traits from textual data, demonstrating their ability to capture semantic and linguistic cues. Nevertheless, these approaches often face limitations related to domain dependency, as models trained on specific types of data may not generalize effectively across domains [2].

In the music domain, personality traits have been shown to be associated with users' preferences for lyrical and emotional content. For example, [17] investigated the relationship between the Big Five personality traits and linguistic patterns present in individuals' favorite songs. Using LIWC-based text analysis and correlational methods, the authors demonstrated that personality traits are predictive of specific linguistic features, suggesting that music consumption reflects stable psychological characteristics.

To extract such linguistic and emotional signals from textual data, lexicon-based approaches have been commonly used. Among these, Empath [12] provides a scalable and flexible alternative to traditional manually curated lexicons. Empath dynamically generates semantic categories from seed words using neural embeddings and validates them through crowd-based evaluation. This approach enables the analysis of a wide range of emotional, social, and topical dimensions in text, making it suitable for capturing affective signals from song lyrics.

Although previous studies have shown that personality traits can improve recommendations, collecting them through questionnaires filled out by users is costly in terms of time, requires ethics committee approval, and is often difficult to obtain in practice. On the calibration side, most approaches still rely solely on the proportion of categories based on the user's interaction history, as proposed by Steck [23], which limits their effectiveness when user profiles are sparse or biased. Combining personality-aware calibration with automatically inferred traits remains an open challenge in the literature.

3 CALIBRATION BASED ON PERSONALITY TRAITS

Figure 1 shows the extended personality-aware calibration approach, which builds on the method proposed by Palomino [16]. While the original approach relies on explicitly personal traits to estimate a personality category distribution, this extension incorporates implicit personality signals. Explicit traits are obtained through questionnaires, whereas implicit traits are inferred using LLMs from user interactions and behavioral data. This allows the estimation of personality traits even when explicit input is not available. Moreover, questionnaire responses may not always reflect users' true personality because self-report measures are susceptible to social desirability and response biases [6]. In contrast, implicit signals derived from actual listening behavior provide complementary behavioral evidence that can improve personality inference beyond self-reported questionnaires [5].

The predicted personality profile is then combined with user category preferences derived from interaction history to construct an affinity matrix. This matrix is built by computing Spearman's correlation between each Big Five trait and the genre consumption frequency in the user's history [16]. The resulting distribution is incorporated into the calibration stage together with the history-based distribution, allowing the recommendation list to be calibrated using history-based, personality-based, or hybrid strategies.

4 LLM-BASED PERSONALITY INFERENCE

To make personality-aware calibration more practical in scenarios where questionnaire data are unavailable, we infer users' Big Five personality traits from music-consumption signals using LLMs. Figure 2 summarizes the personality prediction pipeline. The input consists of each user's listening behavior, which is transformed into two alternative representations. The first representation uses the full listening history, including song titles, artists, and genres. The second representation uses a summarized user profile enriched with semantic and emotional features extracted from song lyrics using Empath [12].

We use two instruction-following LLMs for personality inference: GPT-4.1-mini and Gemini-2.5-flash. Both models receive the same structured prompt and are asked to estimate the Big Five personality traits – Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism – on a scale

from 1.0 to 5.0. These inferred traits are then used in the same calibration framework as questionnaire-based traits, enabling a direct comparison between explicit and LLM-inferred personality signals.

4.1 User Data Representation

4.1.1 Representation 1: Listening History.

For this representation, each user profile is constructed from the user's complete listening history. Specifically, we concatenate the metadata of the tracks listened to by the user, including the artist name, song title, and genre of each track. This representation provides the LLM with a direct behavioral summary of the user's music preferences.

4.1.2 Representation 2: Summary Profile.

To construct a psychologically enriched representation of each user, we construct a user-level Empath profile based on their Top-20 most frequently listened tracks from their complete listening history, following a similar methodology to prior personality-aware music studies such as [17]. However, instead of relying on *Linguistic Inquiry and Word Count (LIWC)*¹, we employ Empath, an open and extensible lexical analysis tool. To obtain the textual content required for lexical analysis, we retrieve song lyrics using the *Genius API*².

Lyrics were retrieved and processed within the same pipeline: they were temporarily kept in memory and immediately analyzed to extract Empath-based features. The motivation for using lyrics is that listening history alone mainly captures behavioral patterns, such as genre variety, artist preferences, or repeated listening. These signals may be useful for traits such as Openness and Extraversion, but they may provide limited affective or semantic evidence for other traits, such as Neuroticism, Agreeableness, and Conscientiousness. To enrich the user representation, we therefore extract lexical, emotional, social, and semantic features from song lyrics using the full set of Empath categories. For each user, the values of each category are averaged across the lyrics of the listened tracks, producing an aggregated lyrical profile. Categories with zero values were removed before constructing the final input representation used in the prompt. This process provides the LLM with compact affective and semantic cues from lyrical content, complementing the behavioral information captured by the listening history.

4.2 Prompt Design for Big Five Personality Inference

Figure 3 presents the prompt used to estimate users' Big Five personality traits from their music-listening data using GPT and Gemini models. The prompt receives one of the two user representations described above: either the full listening history, composed of songs, artists, and genres, or the summarized profile, which includes music-consumption information enriched with Empath-based semantic features. Based on

¹<https://www.liwc.app/>

²<https://genius.com>

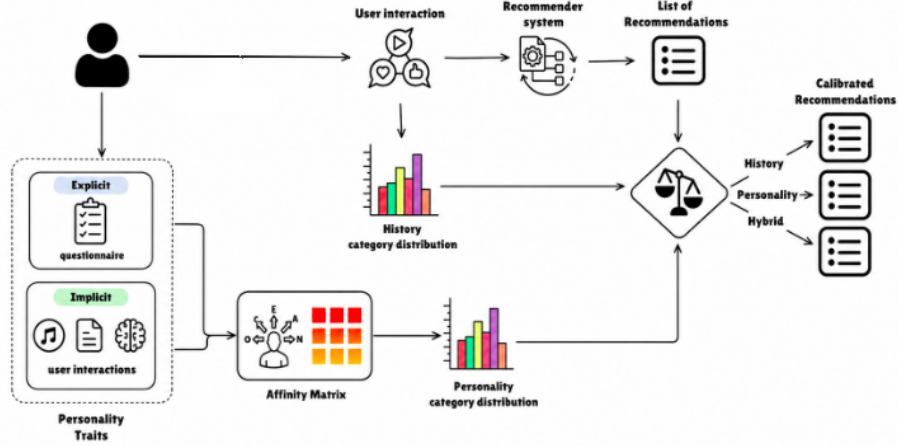


Figure 1: Overview of the extended personality-aware calibration approach.

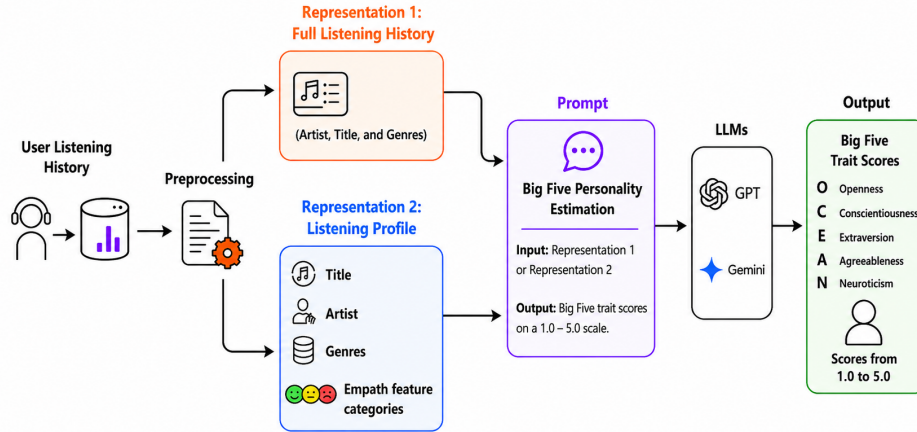


Figure 2: Overview of the personality prediction pipeline using LLMs.

this input, the model estimates five personality scores, Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, each ranging from 1.0 to 5.0. During prompt development, several preliminary executions were conducted to refine the prompt.

For the final experiments, each user was processed once for each LLM and input representation, using the default API settings

5 EVALUATION METHODOLOGY

5.1 Dataset

We rely on the music recommendation dataset provided by Alves [4], which was obtained from an online user study conducted in 2024. Table 1 shows the statistics of the dataset

Table 1: Music Dataset used in the experiment

	Music
Users	348
Items	1766
Interactions	4185

after preprocessing, including users, songs, and interactions. The dataset also contains Big Five personality-trait information, which was used in the original study to support the generation of personality-aware explanations for music recommendations. In our work, we leverage these personality traits to apply our personality-aware calibration approach.

```

""" You are a computational personality modeling expert.
You will receive music-consumption data for one user. The input may be either:
1. a full listening history with songs, artists, and genres, or
2. an aggregated profile with artist, genre, and semantic/empath statistics.
Task:
Estimate the user's Big Five personality trait scores from the provided music-consumption signals.
Important:
Music-consumption data should be treated as indirect behavioral evidence, not as deterministic evidence of personality.
Rules:
- Use only the information provided in the input.
- Do not infer demographic or sensitive attributes.
- Avoid cultural, artist-based, genre-based, or popularity-based stereotypes.
- Do not overinterpret isolated songs, artists, or genres.
- If evidence for a trait is weak, ambiguous, or missing, return a neutral score close to 3.0.
- Scores must be floats between 1.0 and 5.0.
- If semantic/empath coverage is low or absent, reduce the weight of semantic/empath signals.
Interpretation guidelines:
- Openness: may be reflected by diversity, variety, and exploratory listening patterns.
- Conscientiousness: may be weakly reflected by stable or concentrated patterns; infer conservatively.
- Extraversion: may be reflected by energetic, upbeat, social, or dance-oriented patterns when supported by the data.
- Agreeableness: may be weakly reflected by warm, positive, social, or harmony-oriented signals when supported by the data.
- Neuroticism: may be reflected by negative, sad, anxious, or emotionally intense semantic/empath signals when supported by the data.
User music-consumption input JSON:
{USER_INPUT_JSON}
Return only valid JSON in this exact key order:
{
  "openness": <float 1.0-5.0>,
  "conscientiousness": <float 1.0-5.0>,
  "extraversion": <float 1.0-5.0>,
  "agreeableness": <float 1.0-5.0>,
  "neuroticism": <float 1.0-5.0>
}
"""

```

Figure 3: Prompt used to infer Big Five personality traits from music-consumption data.

5.2 Algorithms and Baseline

As illustrated in Figure 1, each recommender first generates an initial list of $N = 50$ candidate items. The calibration step is then applied to this candidate list, re-ranking the items according to the selected target distribution. After this process, only the top-10 items are returned as the final recommendation list.

Unlike previous personality-aware calibration work [16], which relied on default hyperparameter configurations, in this work we first performed hyperparameter optimization for each algorithm before applying calibration. This optimization was performed by evaluating predefined hyperparameter combinations on the validation set, selecting the configuration that produced the best performance. The selected hyperparameters were then fixed for the subsequent evaluation. This procedure allowed us to assess whether personality-aware calibration remains effective when applied to competitive initial recommendation rankings.

The baseline recommenders were implemented using CaseRecommender³, an open source Python framework for RS [8]. For hyperparameter optimization, the dataset was initially partitioned using a user-wise (80%) train-test split. The training set was further divided into a train split (80%) and a validation set (20%), the latter used exclusively for hyperparameter selection. All algorithms were executed with their optimized hyperparameter configurations and a ranking length of $N = 50$ to produce the candidate list required by the calibration strategy.

We considered four algorithms: **BPRMF** (*Bayesian Personalized Ranking Matrix Factorization*) [18], **UserKNN** and

ItemKNN (*User-based and Item-based K-Nearest Neighbors*) [3], and **ItemAttributeKNN**.

The optimized hyperparameters were defined as follows: for **BPRMF**, we used 64 latent factors, a learning rate of 0.05, and 70 training epochs; for **UserKNN**, we used Jaccard similarity with 40 neighbors; for **ItemKNN**, we used Jaccard similarity with 70 neighbors; and for **ItemAttributeKNN**, we used Jaccard similarity with 10 neighbors.

As a *baseline*, our proposal was compared with the traditional history-based calibration method proposed by Steck [23].

Thus, we have:

- **History-based Calibration:** Target distribution calculated from the user's consumption history, following Steck's approach in Figure 1.
- **Personality-based Calibration:** Target distribution derived from the affinity matrix obtained through Spearman's correlation between personality traits and music genres, as shown in Figure 1.
- **Hybrid-based Calibration:** The hybrid calibration approach combines historical and personality-based information through a weighted aggregation of their respective divergence terms, assigning equal weights (50% each) to both information sources.

5.3 Metrics

We evaluate the recommendation results from multiple perspectives, considering accuracy, fairness, diversity, popularity bias, and overall utility, following prior personality-aware calibration work [16]. Accuracy is measured using Mean Average Precision (MAP), which evaluates the quality of the ranked

³<https://github.com/caserec/CaseRecommender>

recommendation list by considering the position of relevant items. Diversity is measured using Intra-List Diversity (ILD), which captures the average dissimilarity among the items recommended to the same user.

To evaluate calibration-oriented fairness, we use $F1_{MRMC}$, which combines genre-based and popularity-based miscalibration into a single score. Higher $F1_{MRMC}$ values indicate that the recommendation list is better aligned with the user profile while reducing miscalibration across item groups. To capture popularity bias, we report $RMSE_{GAP}$, which measures the discrepancy between the target distribution and the distribution observed in the recommendation list; lower values indicate less bias.

Finally, Multi-Attribute Utility Theory (MAUT) aggregates the evaluated metrics into a single utility score for global comparison. All metrics are assigned equal weight, so accuracy, fairness, diversity, and calibration error contribute equally to the final score.

5.4 Execution Protocol

The evaluation of the proposed approach was conducted comparatively, considering three calibration strategies: (i) the traditional strategy, based on the user's interaction history, (ii) the personality-based strategy, whose target distribution is derived from genre affinity estimated through the Big Five model traits, and (iii) the hybrid strategy, which combines both user's interaction history and traits to calibrate the recommended list.

In addition, for the personality-based and hybrid strategies, we compare two types of personality traits: explicit traits, obtained from questionnaire responses, and implicit traits, inferred by the LLM from music-consumption data. This comparison allows us to analyze whether LLM-inferred personality traits can replace questionnaire-based traits in the calibration process without substantially changing the behavior of the recommendation results.

After hyperparameter optimization, the selected hyperparameters were fixed for the final evaluation. Ten evaluation folds were generated using independent user-wise 80/20 train-test splits. For each fold, 80% of the interactions were used to train the recommender and 20% for testing.

The recommender models **UserKNN**, **ItemKNN**, **BPRMF**, and **ItemAttributeKNN** were trained on each fold and used to generate an initial list of $N = 50$ candidate items. Each calibration approach was then applied to these candidate lists, and the final top-10 recommendation lists were evaluated using the metrics described in the previous subsection. The final reported results correspond to the mean across the 10 folds. Statistical significance was assessed using the Wilcoxon signed-rank test.

6 RESULTS

6.1 RQ1: Personality Prediction Using LLMs

Table 2 reports the RMSE for each Big Five trait, Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, against questionnaire based ground truth, comparing

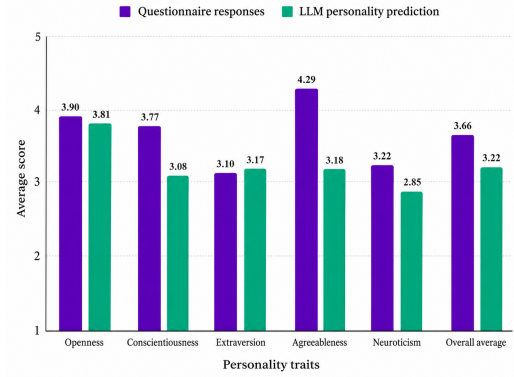


Figure 4: Overall average is a derived measure across all Big Five traits. Likert scale: 1 = strongly disagree, 5 = strongly agree.

LLM-based inference against Random and Permutation baselines. Overall, all LLM configurations outperform both baselines, with Summary Profile (GPT) achieving the best Total RMSE of 0.958, compared with 1.039 for Permutation and 1.590 for Random. These results provide evidence of the predictive performance of the LLM-based approaches compared with the reference baselines considered in this work.

For GPT, the summarized profile approach outperforms History, improving O, C, E, and A, but with a higher error for N. For Gemini, History performs better overall (1.007 vs. 1.037), with the summarized profile improving only Agreeableness, while resulting in higher errors for the other traits.

Across traits, Openness has the lowest error, while Agreeableness and Extraversion remain more challenging. Additionally, Figure 4 compares the average personality trait values obtained from the ground truth with those derived from the LLM-based predictions using the best-performing configuration (Summary Profile with GPT).

The predicted traits follow a similar overall distribution to the ground truth for most dimensions; however, notable differences are observed for Agreeableness and Conscientiousness, where the LLM predictions deviate more substantially from the questionnaire-based values.

Based on these results, the Summary Profile with GPT configuration, which achieved the lowest overall RMSE (0.958) among the evaluated LLM configurations, was selected for the subsequent calibration experiments in RQ2.

Answer to RQ1: LLMs can approximate personality traits from music data. The best performance is achieved by GPT with the summarized profile, indicating that combining listening behavior with lyrical features can improve the personality prediction. However, effectiveness varies by model and trait, suggesting that representation choice and personality dimension both influence prediction quality.

The comparison with Random and Permutation provides a reference point for the predictive performance of the proposed LLM-based approaches.

Table 2: RMSE for Big Five personality trait prediction. Lower values indicate better performance. O, C, E, A, and N denote Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, respectively.

Method	O	C	E	A	N	Avg.
Random	1.617	1.452	1.552	1.801	1.531	1.590
Permutation	1.001	0.765	1.424	0.719	1.287	1.039
History (GPT)	0.770	0.929	1.056	1.276	0.915	0.989
History (Gemini)	0.791	0.923	1.113	1.225	0.981	1.007
Summary Profile (GPT)	0.702	0.876	1.020	1.218	0.976	0.958
Summary Profile (Gemini)	0.871	0.938	1.171	1.074	1.133	1.037

6.2 RQ2: Robustness of Calibration to LLM-Predicted Personality Traits

We compare the results obtained with questionnaire-based traits and GPT-inferred personality traits, reported in Table 3, respectively. The goal is to verify whether personality-aware calibration preserves its behavior when explicit personality information is replaced by inferred personality traits.

Overall, the results show that GPT-predicted personality traits can be used effectively in the calibration process. The personality-based strategy (+per) produces results similar to those obtained with questionnaire-based traits, especially for diversity, fairness, and global utility. This indicates that LLM-inferred traits provide useful signals for guiding calibration, even when explicit questionnaire data are not available.

For UserKNN and ItemKNN, the personality-based strategy consistently improves $F1_{MRMC}$ and ILD compared with history-based calibration. This means that GPT-predicted traits help generate recommendation lists that are more calibrated and more diverse. These improvements are also reflected in MAUT, where +per often obtains higher values than +his, showing a better overall trade-off among the evaluated metrics.

For BPRMF, the improvements are more limited. Although +per improves $F1_{MRMC}$ in several cases, the gains in diversity and MAUT are less consistent than in the neighborhood-based algorithms. This suggests that the effect of personality-aware calibration depends on the recommendation model used to generate the initial ranking.

For ItemAttributeKNN, the results show that GPT-predicted traits are also useful in the content-based setting. The +per strategy improves MAUT in several configurations and can improve diversity and fairness, although the effects are not as stable as in UserKNN and ItemKNN.

Regarding accuracy, MAP is generally preserved at low and moderate values of λ . However, when $\lambda = 1.00$, the personality-based strategy tends to reduce MAP. This happens because the calibration gives maximum importance to the personality-derived distribution, which increases diversity but may reduce relevance. The hybrid strategy (+hyb) helps reduce this effect by combining historical and personality-based information.

Answer to RQ2: Personality-aware calibration remains consistent when questionnaire-based traits are replaced by GPT-inferred personality traits. The results show improvements in fairness, diversity, and overall utility, especially for UserKNN

and ItemKNN, in optimized recommendation settings as well, under competitive initial rankings. However, the benefits depend on the algorithm and on the calibration strength. Low and moderate values of λ provide a better balance, while strong calibration may reduce accuracy.

7 CONCLUSIONS

This paper investigated LLM-inferred personality traits for calibrated music recommendation, analyzing whether automatically inferred traits can replace questionnaire-based personality information in personality-aware calibration. We evaluated GPT and Gemini for Big Five personality prediction from music-consumption data and incorporated the inferred traits into history-based, personality-based, and hybrid calibration strategies. LLMs can approximate personality traits from music-consumption signals, with GPT using the summarized profile achieving the lowest RMSE. Despite variation across traits and models, the inferred traits provide useful calibration signals that improve diversity, fairness, and utility, complementing strong initial rankings by balancing relevance and beyond-accuracy objectives.

The impact of calibration depends on the recommendation algorithm, the source of personality information, and the calibration strength. Low and moderate values of λ tend to provide a better trade-off between accuracy and beyond-accuracy dimensions. Stronger calibration may increase diversity and fairness but can reduce MAP, especially when the personality-based distribution dominates the re-ranking process. These findings highlight the importance of carefully controlling the calibration trade-off to avoid excessive deviation from the original relevance ranking. Future research may explore additional LLMs and prompting strategies, evaluate the approach in other domains beyond music, user studies, automatic methods for selecting λ , and hybrid approaches combining implicit and explicit personality traits to further personalize diversity and fairness objectives.

ETHICAL CONSIDERATIONS

This study uses secondary data from a previous user study by Alves [4], whose participants provided informed consent. The dataset was made available through the study repository. No new data collection was conducted. In real-world applications, users should be informed when behavioral data are used to infer psychological traits, as this may raise privacy and transparency concerns, particularly in sensitive domains.

Table 3: Explicit vs. implicit (GPT-inferred) personality trait calibration on the Music dataset, reported for Historical (+his), Personality (+per), and Hybrid (+hyb) calibration across four recommendation models. ▲, ●, and ▼ indicate the personality/hybrid version is statistically superior, not significantly different, or inferior to the historical baseline, respectively. * denotes identical value to historical.

Model	Calibration λ	Explicit personality traits					Implicit GPT-inferred traits				
		MAP	F1 _{MRMC}	ILD	RMSE _{GAP}	MAUT	MAP	F1 _{MRMC}	ILD	RMSE _{GAP}	MAUT
BPRMF											
	Baseline (0.00)	0.052	0.222	0.570	3.166	0.463	0.052	0.222	0.570	3.166	0.457
	+his (0.10)	0.053	0.211	0.572	3.176	0.431	0.053	0.211	0.572	3.176	0.426
	+per (0.10)	0.052🟡	0.222🟢	0.570🔴	3.166🟡	0.463🟢	0.052🔴	0.222🟢	0.570🔴	3.166🟢	0.457🟢
	+hyb (0.10)	0.052🔴	0.216🟢	0.571🔴	3.175🟡	0.441🟢	0.052🔴	0.216🟢	0.571🔴	3.175🟡	0.436🟢
	+his (0.40)	0.062	0.185	0.579	3.119	0.386	0.062	0.185	0.579	3.119	0.384
	+per (0.40)	0.058🟡	0.221🟢	0.573🔴	3.145🟡	0.491🟢	0.058🟡	0.221🟢	0.573🔴	3.147🔴	0.486🟢
	+hyb (0.40)	0.060🔴	0.197🟢	0.576🔴	3.137🟡	0.413🟢	0.060🔴	0.197🟢	0.576🔴	3.137🟡	0.409🟢
	+his (0.70)	0.060	0.165	0.584	3.028	0.329	0.060	0.165	0.584	3.028	0.329
	+per (0.70)	0.058🟡	0.220🟢	0.576🔴	3.143🔴	0.492🟢	0.058🟡	0.220🟢	0.576🔴	3.146🔴	0.486🟢
	+hyb (0.70)	0.061🟡	0.174🟢	0.584🟡	3.070🟡	0.360🟢	0.062🟡	0.174🟢	0.584🟡	3.073🔴	0.358🟢
	+his (1.00)	0.052	0.158	0.551	2.063	0.427	0.052	0.158	0.551	2.063	0.434
	+per (1.00)	0.020🔴	0.209🟢	0.768🟢	2.004🟡	0.699🟢	0.021🔴	0.218🟢	0.765🟢	2.092🔴	0.722🟢
	+hyb (1.00)	0.051🟡	0.158🟡	0.557🟢	2.054🟡	0.433🟡	0.051🟡	0.159🟢	0.556🟢	2.052🟡	0.442🟡
UserKNN											
	Baseline (0.00)	0.142	0.203	0.569	1.423	0.497	0.142	0.203	0.569	1.423	0.484
	+his (0.10)	0.136	0.172	0.578	1.283	0.400	0.136	0.172	0.578	1.283	0.400
	+per (0.10)	0.140🟢	0.201🟢	0.590🟢	1.317🔴	0.553🟢	0.140🟢	0.201🟢	0.590🟢	1.318🟡	0.547🟢
	+hyb (0.10)	0.140🟢	0.180🟢	0.579🟡	1.292🟡	0.445🟢	0.139🟢	0.180🟢	0.579🟡	1.295🟡	0.441🟢
	+his (0.40)	0.117	0.156	0.583	1.231	0.313	0.117	0.156	0.583	1.231	0.318
	+per (0.40)	0.138🟢	0.199🟢	0.597🟢	1.308🔴	0.553🟢	0.139🟢	0.199🟢	0.597🟢	1.311🔴	0.546🟢
	+hyb (0.40)	0.123🟢	0.160🟢	0.586🟢	1.258🔴	0.338🟢	0.123🟢	0.160🟢	0.586🟢	1.250🟡	0.345🟢
	+his (0.70)	0.098	0.155	0.570	1.154	0.289	0.098	0.155	0.570	1.154	0.295
	+per (0.70)	0.135🟢	0.196🟢	0.613🟢	1.281🔴	0.560🟢	0.135🟢	0.195🟢	0.613🟢	1.286🔴	0.546🟢
	+hyb (0.70)	0.105🟢	0.155🟡	0.582🟢	1.201🔴	0.295🟡	0.105🟢	0.155🟡	0.582🟢	1.203🔴	0.299🟡
	+his (1.00)	0.067	0.158	0.563	0.993	0.303	0.067	0.158	0.563	0.993	0.312
	+per (1.00)	0.029🔴	0.203🟢	0.798🟢	0.903*	0.730🟢	0.029🔴	0.207🟢	0.793🟢	0.943🟡	0.722🟢
	+hyb (1.00)	0.065🟡	0.157🔴	0.568🟢	0.993🟡	0.305🟡	0.064🔴	0.158🟡	0.568🟢	0.989🟡	0.313🟡
ItemKNN											
	Baseline (0.00)	0.116	0.210	0.614	0.273	0.773	0.116	0.210	0.614	0.273	0.773
	+his (0.10)	0.112	0.182	0.625	0.270	0.645	0.112	0.182	0.625	0.270	0.645
	+per (0.10)	0.114🟡	0.207🟢	0.656🟢	0.271🟡	0.794🟢	0.114🟡	0.206🟢	0.656🟢	0.270🟡	0.792🟢
	+hyb (0.10)	0.116🟢	0.190🟢	0.625🟡	0.270🟡	0.688🟢	0.116🟢	0.189🟢	0.625🟡	0.270🟡	0.688🟢
	+his (0.40)	0.097	0.161	0.630	0.279	0.504	0.097	0.161	0.630	0.279	0.505
	+per (0.40)	0.113🟢	0.206🟢	0.658🟢	0.271🟡	0.788🟢	0.114🟢	0.205🟢	0.658🟢	0.270🟡	0.787🟢
	+hyb (0.40)	0.103🟢	0.167🟢	0.631🟡	0.276🟡	0.551🟢	0.103🟢	0.167🟢	0.631🟡	0.276🟡	0.552🟢
	+his (0.70)	0.086	0.157	0.620	0.323	0.416	0.086	0.157	0.620	0.323	0.418
	+per (0.70)	0.111🟢	0.203🟢	0.664🟢	0.271🟢	0.776🟢	0.111🟢	0.202🟢	0.663🟢	0.270🟢	0.772🟢
	+hyb (0.70)	0.091🟢	0.158🟢	0.632🟢	0.297🟢	0.464🟢	0.091🟢	0.158🟢	0.631🟢	0.297🟢	0.465🟢
	+his (1.00)	0.060	0.158	0.574	0.512	0.153	0.060	0.158	0.574	0.512	0.165
	+per (1.00)	0.028🔴	0.195🟢	0.823🟢	0.572🔴	0.442🟢	0.026🔴	0.195🟢	0.822🟢	0.591🔴	0.440🟢
	+hyb (1.00)	0.059🟡	0.157🟡	0.579🟢	0.523🟡	0.148🟡	0.059🟡	0.158🟡	0.578🟢	0.524🔴	0.157🔴
ItemAttributeKNN											
	Baseline (0.00)	0.026	0.295	0.255	0.404	0.565	0.026	0.295	0.255	0.404	0.556
	+his (0.10)	0.028	0.267	0.302	0.393	0.592	0.028	0.267	0.302	0.393	0.585
	+per (0.10)	0.027🟡	0.287🟢	0.293🔴	0.369🟢	0.652🟢	0.028🟡	0.285🟢	0.309🟢	0.376🟢	0.677🟢
	+hyb (0.10)	0.028🟡	0.275🟢	0.299🔴	0.392🟡	0.616🟢	0.028🟡	0.274🟢	0.299🔴	0.391🟡	0.611🟡
	+his (0.40)	0.027	0.234	0.329	0.398	0.523	0.027	0.234	0.329	0.398	0.515
	+per (0.40)	0.027🟡	0.288🟢	0.293🔴	0.369🟢	0.652🟢	0.028🟡	0.285🟢	0.310🔴	0.376🟢	0.677🟢
	+hyb (0.40)	0.028🟡	0.247🟢	0.315🔴	0.393🟡	0.551🟡	0.028🟡	0.246🟢	0.315🔴	0.393🟡	0.554🟡
	+his (0.70)	0.027	0.213	0.357	0.415	0.459	0.027	0.213	0.357	0.415	0.446
	+per (0.70)	0.027🟡	0.286🟢	0.295🔴	0.369🟢	0.644🟢	0.028🟡	0.284🟢	0.311🔴	0.375🟢	0.676🟢
	+hyb (0.70)	0.027🟡	0.222🟢	0.343🔴	0.399🟢	0.478🟡	0.027🟡	0.213🟢	0.357*	0.415*	0.446🟢
	+his (1.00)	0.025	0.200	0.433	0.547	0.304	0.025	0.200	0.433	0.547	0.283
	+per (1.00)	0.021🔴	0.220🟢	0.502🟢	0.431🟢	0.513🟢	0.021🔴	0.222🟢	0.520🟢	0.442🟢	0.502🟢
	+hyb (1.00)	0.024🟡	0.200🟡	0.435🟢	0.546🟡	0.299🟡	0.025🟡	0.200🟡	0.435🟢	0.548🟡	0.289🟡

ACKNOWLEDGMENTS

This work was supported by the Brazilian funding agencies FAPESP, CNPq and CAPES, as well as INCT-TILD-IAR

(process number 408490/2024-1). The authors acknowledge the use of AI tools for language editing and proofreading.

REFERENCES

- [1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward C. Malthouse. 2021. User-centered Evaluation of Popularity Bias in Recommender Systems. In *Proceedings of the 29th ACM Conference UMAP 2021, Utrecht*. ACM, 119–129. doi:10.1145/3450613.3456821
- [2] Somdeep Acharyya and Nargis Pervin. 2025. Enhancing cross-domain recommendations: Leveraging personality-based transfer learning with probabilistic matrix factorization. *Expert Systems with Applications* 263 (2025), 125667.
- [3] Charu C Aggarwal et al. 2016. *Recommender systems*. Vol. 1. Springer.
- [4] Gabrielle Alves, Dietmar Jannach, Luan Soares de Souza, and Marcelo Garcia Manzato. 2025. Towards Personality-Aware Explanations for Music Recommendations Using Generative AI. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. 684–689.
- [5] Ian Anderson, Santiago Gil, Clay Gibson, Scott Wolf, Will Shapiro, Oguz Semerci, and David M Greenberg. 2021. “Just the way you are”: Linking music listening on Spotify and personality. *Social Psychological and Personality Science* 12, 4 (2021), 561–572.
- [6] Henry I Braun, Douglas N Jackson, and David E Wiley. 2001. Socially desirable responding: The evolution of a construct. In *The role of constructs in psychological and educational measurement*. Routledge, 61–84.
- [7] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems* 41, 3 (2023), 1–39.
- [8] Arthur Da Costa, Eduardo Fressato, Fernando Neto, Marcelo Manzato, and Ricardo Campello. 2018. Case recommender: a flexible and extensible python framework for recommender systems. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 494–495.
- [9] Diego da Silva and Dietmar Jannach. 2026. Calibrated recommendations: Survey and future directions. *ACM Transactions on Recommender Systems* 4, 3 (2026), 1–32.
- [10] Sahraoui Dhelim, Nyothiri Aung, Mohammed Amine Bouras, Huansheng Ning, and Erik Cambria. 2022. A survey on personality-aware recommendation systems. *Artificial Intelligence Review* 55, 3 (2022), 2409–2454.
- [11] John M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology* 41, 1 (1990), 417–440.
- [12] Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 4647–4657.
- [13] Raghav Pavan Karumur, Tien T Nguyen, and Joseph A Konstan. 2018. Personality, user preferences and behavior in recommender systems. *Information Systems Frontiers* 20, 6 (2018), 1241–1265.
- [14] Mohammadmehdi Naghiaei, Mahdi Dehghan, Hossein A Rahmani, Javad Azizi, and Mohammad Aliannejadi. 2024. Personalized beyond-accuracy calibration in recommendation. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*. 107–116.
- [15] Huansheng Ning, Sahraoui Dhelim, and Nyothiri Aung. 2019. PersoNet: Friend recommendation system based on big-five personality traits and hybrid filtering. *IEEE Transactions on Computational Social Systems* 6, 3 (2019), 394–402.
- [16] Yuvisa Quispe Palomino, Rodrigo Ferrari de Souza, Paul Dany Flores Atauchi, and Marcelo Garcia Manzato. 2026. Calibration of Personality-Based Recommender Systems: Balancing Relevance and Diversity. In *Proceedings of the 41st ACM/SIGAPP Symposium on Applied Computing*. 2118–2126.
- [17] Lin Qiu, Jiayu Chen, Jonathan Ramsay, and Jiahui Lu. 2019. Personality predicts words in favorite songs. *Journal of Research in Personality* 78 (2019), 25–35.
- [18] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (Montreal, Quebec, Canada) (UAI '09). AUAI Press, Arlington, Virginia, USA, 452–461.
- [19] Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). 2022. *Recommender Systems Handbook*. Springer US. doi:10.1007/978-1-0716-2197-4
- [20] Andre Sacilotti, Rodrigo Souza, and Marcelo G. Manzato. 2023. Counteracting Popularity-Bias and Improving Diversity Through Calibrated Recommendations. In *Proceedings of the 25th International Conference on Enterprise Information Systems - Volume 1: ICEIS*. INSTICC, SciTePress, 709–720. doi:10.5220/0011846000003467
- [21] Rafael Santos and Giovanni Comarella. 2025. FaDeRS: Fairness and Depolarization in Recommender Systems. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*. SBC, 248–257.
- [22] Rodrigo Souza and Marcelo Manzato. 2024. A two-stage calibration approach for mitigating bias and fairness in recommender systems. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*. 1659–1661.
- [23] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems*. 154–162.
- [24] Hsin-Chang Yang and Zi-Rui Huang. 2019. Mining personality traits from social messages for game recommender systems. *Knowledge-Based Systems* 165 (2019), 157–168.
- [25] Yuying Zhao, Yu Wang, Yunchao Liu, Xueqi Cheng, Charu C. Aggarwal, and Tyler Derr. 2025. Fairness and Diversity in Recommender Systems: A Survey. *ACM Trans. Intell. Syst. Technol.* 16, 1, Article 2 (Jan. 2025), 28 pages. doi:10.1145/3664928