PFx: Measuring the Trade-off Between Fairness and Personality in LLM-Based Recommendations

*Note: Author Rebuttal for Submission 89

I. Introduction

1) Fairness Definition.: In the context of large language model-based recommender systems (RecLLMs), we define fairness as the absence of systemic prejudice or favoritism toward user groups defined by sensitive attributes (e.g., gender, age, continent, occupation) in the generation of recommendations. Our focus is on user-side fairness, which ensures that users with comparable preferences and personalities receive equitable treatment—regardless of their demographic group—especially when such sensitive information is not explicitly present in the prompt. Furthermore, this work extends fairness analysis beyond traditional demographic parity to incorporate personality-aware fairness. We examine whether recommendations are equitably aligned with users' inferred psychographic traits (e.g., OCEAN dimensions), ensuring that LLMs do not inadvertently prioritize or marginalize certain personality types during generation. This dual view—demographic and psychographic—forms the basis of our proposed PFx framework.

2) Personality-Based Fairness Definition.: Although traditional fairness in recommendation systems emphasizes demographic parity (e.g., gender, age, occupation), it often neglects the role of psychographic factors. We define personality-based fairness as the consistent and equitable alignment of recommendation outputs with users' individual personality traits, as modeled by the OCEAN framework. This concept ensures that no specific personality profile (e.g., introverted vs. extroverted) is systematically advantaged or disadvantaged in how recommendations are generated. By incorporating cognitive, emotional, and behavioral diversity, this definition broadens the fairness paradigm to account for deeper dimensions of user identity.

II. 3.1.3 Benchmark Metrics.

We employ 10 key evaluation metrics to assess the fairness, personality alignment, and quality of LLM-based recommendations. These metrics are grouped into three categories:

a) Personality Fit.: Personality fit is assessed using Personality Alignment Score (PAS) and Genre-Personality Alignment (GPA), which together capture how well recommendations reflect the user's inferred psychological profile.

- **Personality Alignment Score (PAS)**: Measures the cosine similarity between the user's OCEAN vector $\vec{p}_u$ and the inferred genre vector $\vec{g}_u$ from the LLM's recommended items.

$$PAS(u) = \frac{\vec{p}_u \cdot \vec{g}_u}{\|\vec{p}_u\| \cdot \|\vec{g}_u\|} \tag{1}$$

  where $\vec{p}_u$ represents the OCEAN personality vector for user u, and $\vec{g}_u$ is the genre vector inferred from the LLM's recommendation. The dot product between these vectors captures how similar the user's personality is to the recommended genres, normalized by the magnitude of each vector(1).
- **Genre-Personality Alignment (GPA)**: Aggregates weighted overlaps between recommended genres and those linked to the user's dominant OCEAN traits.

$$GPA(u) = \sum_{g \in G_{rec}} \sum_{t \in OCEAN} \mathbb{I}_{g \in G_t} \cdot p_u^t \tag{2}$$

  where G_{rec} is the set of genres recommended to user u, and G_t denotes the set of genres associated with each of the OCEAN traits $t \in OCEAN$. The indicator function $\mathbb{I}_{g \in G_t}$ is 1 if genre g aligns with the trait t, and p_u^t is the user's score for that trait(2).

b) Fairness.: We evaluate fairness using Demographic Parity (DP) and Equal Opportunity (EO), which measure group-level disparities, along with Intra-list Fairness (ILF), which captures diversity within individual recommendation lists.

- **DP (Demographic Parity)**: Measures the difference in recommendation probability across sensitive groups.

$$DP = |P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1)| \tag{3}$$

  where $\hat{Y}$ is the recommendation outcome (1 if an item is recommended, 0 if not), and A is a sensitive attribute (e.g., gender, race). This metric compares the recommendation probability between two groups (e.g., male and female, or different races(3)).
- **EO (Equal Opportunity)**: Measures the fairness of recommendations given positive outcomes.

$$EO = |P(\hat{Y} = 1|Y = 1, A = 0) - P(\hat{Y} = 1|Y = 1, A = 1)| \tag{4}$$

  where Y is the true relevance (whether the item is actually relevant), and A represents sensitive attributes. This metric compares the true positive recommendation rates between two sensitive groups (e.g., how equally likely a relevant item is recommended for both genders(4)).

Identify applicable funding agency here. If none, delete this.

- **ILF (Intra-list Fairness)**: Measures the diversity of recommendations within a list.

$$ILF@K = -\sum_{g \in G} p(g) \cdot \log p(g) \tag{5}$$

where G is the set of items in the recommendation list, and $p(g)$ is the probability distribution of items. This metric evaluates how evenly items are distributed across different categories, ensuring that the recommendations don't favor certain types of items disproportionately (5).

c) Prompt Sensitivity.: We assess prompt sensitivity using SNSR@K (Sensitive-to-Neutral Similarity Range), SNSV@K (Similarity Variance), and Jaccard@K, which together quantify how much the model's recommendations change in response to personality-sensitive prompting.

- **SNSR@K (Sensitive-to-Neutral Similarity Range) (6)**: Measures the maximum variation in recommendation overlap between sensitive and neutral prompts.

$$SNSR@K = \max_{a \in A} \left| \frac{|R_a^K \cap R_n^K|}{K} \right| - \min_{a \in A} \left| \frac{|R_a^K \cap R_n^K|}{K} \right| \tag{6}$$

where R_a^K and R_n^K represent the top-K recommendations under sensitive and neutral prompts for group a. This metric measures how much the top-K recommendations fluctuate when switching between sensitive and neutral prompts.

- **SNSV@K (Sensitive-to-Neutral Similarity Variance)**: Measures the variance in recommendation overlap across sensitive groups.

$$SNSV@K = \mathrm{Var}_{a \in A} \left(\frac{|R_a^K \cap R_n^K|}{K} \right) \tag{7}$$

This metric calculates the variance in the recommendation overlap for sensitive groups under neutral and sensitive prompts. Higher variance suggests instability in the model's recommendations based on how the user profile is framed(6).

- **Jaccard@K(7)**: Measures the similarity of the top-K recommendations between neutral and sensitive prompts.

$$Jaccard@K = \frac{|R_{neutral}^K \cap R_{sensitive}^K|}{|R_{neutral}^K \cup R_{sensitive}^K|} \tag{8}$$

This metric calculates the Jaccard similarity, comparing the overlap of recommendations under neutral vs. sensitive prompts. A higher Jaccard score suggests that the prompt variation does not heavily impact the recommendations.

d) Standard Metrics.: We use Precision@K and Recall@K to evaluate the relevance and retrieval effectiveness of the top-K recommendations, providing a baseline for assessing overall recommendation quality.

- **Precision@K(8)**: Measures the proportion of top-K recommendations that are relevant.

$$Precision@K = \frac{|Rel \cap Rec@K|}{|Rec@K|} \tag{9}$$

where Rel is the set of relevant items, and $Rec@K$ is the top-K recommended items. This metric measures the accuracy of the top-K recommendations by calculating the proportion that is relevant.

- **Recall@K(9)**: Measures the proportion of relevant items in the top-K recommendations.

$$Recall@K = \frac{|Rel \cap Rec@K|}{|Rel|} \tag{10}$$

This metric captures how many of the relevant items are retrieved in the top-K list. It ensures that the recommender system isn't missing out on relevant items, even if they are not ranked at the top.

A. Dataset and Preprocessing

We evaluate our PFx framework on large-scale public datasets: **MovieLens 10M**, which provide complementary domains (movies) for assessing fairness and personalization in LLM-based recommendations. Table I summarizes key statistics and sensitive features for each dataset.

- **MovieLens 10M**[1] contains over 10 million explicit ratings from 71,567 users on 10,681 movies. Each user is tagged with demographic attributes such as gender, age group, and occupation. We preprocess `ratings.dat`, `movies.dat`, and `tags.dat` files to construct user-item interaction matrices, extract genre metadata, and infer user preferences.

For datasets, we perform the following preprocessing steps:

- **User Filtering:** We retain users with at least 200 interactions to ensure reliable personality inference and meaningful LLM-based evaluations.
- **Trait Inference via Genre Mapping:** Building on prior work in psychometric personalization, we map movie and music genres to the Big Five (OCEAN) traits. For example, genres such as Sci-Fi and Documentary are linked to high Openness, while Romance may correlate with Agreeableness. This mapping allows us to simulate personality profiles and evaluate alignment.

B. Personality Modeling.

We represent each user u as a five-dimensional vector grounded in the Big Five personality traits (OCEAN model), denoted by:

$$\vec{p}_u = [O_u, C_u, E_u, A_u, N_u] \in [0, 1]^5$$

where each dimension corresponds to Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Inspired by McCrae and John (10), we infer these traits using behavioral proxies extracted from user interaction data such as genre affinity distributions, rating dispersion, temporal activity, and catalog diversity.

To strengthen psychological grounding, we integrate trait-genre mappings aligned with prior work in music preference modeling (11). Specifically, traits like Openness correlate with

[1]https://grouplens.org/datasets/movielens/10m/

TABLE I: Key Dataset Statistics for MovieLens 10M.

Dataset	Female/Male	Senior/Young	Items	Records	Ratings	Sensitive Features
MovieLens 10M	17.4k / 43.6k	1.4k / 4.6k	10.0k	10,000.1k	1–5 stars	⚥ Gender, 🎂 Age, 💼 Occupation

TABLE II: Big Five trait dimensions and representative behavioral factors, adapted from (10; 11)

OCEAN Trait	Representative Characteristics
Openness	Artistic, curious, original, wide interest
Conscientiousness	Efficient, organized, reliable, responsible
Extraversion	Energetic, talkative, assertive, sociable
Agreeableness	Kind, sympathetic, generous, trusting
Neuroticism	Anxious, moody, unstable, tense

exploratory genres (e.g., Indie, Jazz), while traits like Conscientiousness align with structured content (e.g., Documentaries, Classical). These mappings help simulate a user's cognitive-emotional profile and enable trait-sensitive prompt generation and fairness evaluation.

C. Prompt Design.

We design two types of prompts to evaluate the impact of personality-sensitive conditioning on LLM-generated recommendations: a generic neutral prompt and a personality-aligned sensitive prompt. This prompt distinction is central to our framework and supports comparative evaluation of fairness and alignment.

Neutral Prompt:

> "Please recommend 15 popular movies/music suitable for a general audience."

Sensitive Prompt Example (for an Introverted User):

> "I am an introverted movie lover who prefers thoughtful, emotional stories. Please recommend 15 movies/music."

These prompts reflect the contrast between system-wide generalization and individualized personalization. The *neutral prompt* provides a baseline for group fairness, whereas the *sensitive prompt* aims to reflect users' psychological traits (e.g., introversion, agreeableness) inferred from their content preferences. This design choice allows us to evaluate whether personality-aware prompting leads to noticeable behavioral shifts in recommendations and to what extent these shifts remain equitable across demographic and personality subgroups. Figure 2 in main manuscript illustrates how these prompt types are embedded within the PFx evaluation pipeline.

REFERENCES

[1] E. Lex, D. Kowald, P. Seitlinger, T. N. T. Tran, A. Felfernig, M. Schedl *et al.*, "Psychology-informed recommender systems," *Foundations and trends® in information retrieval*, vol. 15, no. 2, pp. 134–242, 2021.

[2] R. Abolghasemi, P. Engelstad, E. Herrera-Viedma, and A. Yazidi, "A personality-aware group recommendation system based on pairwise preferences," *Information Sciences*, vol. 595, pp. 1–17, 2022.

[3] A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi *et al.*, "Fairness in recommendation ranking through pairwise comparisons," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2212–2220.

[4] H. Wu, B. Mitra, C. Ma, F. Diaz, and X. Liu, "Joint multisided exposure fairness for recommendation," in *Proceedings of the 45th International ACM SIGIR Conference on research and development in information retrieval*, 2022, pp. 703–714.

[5] M. Jiang, K. Bao, J. Zhang, W. Wang, Z. Yang, F. Feng, and X. He, "Item-side fairness of large language model-based recommendation system," in *Proceedings of the ACM Web Conference 2024*, 2024, pp. 4717–4726.

[6] J. Zhang, K. Bao, Y. Zhang, W. Wang, F. Feng, and X. He, "Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation," in *Proceedings of the 17th ACM Conference on Recommender Systems*, 2023, pp. 993–999.

[7] W. I. D. Mining, "Data mining: Concepts and techniques," *Morgan Kaufinann*, vol. 10, no. 559-569, p. 4, 2006.

[8] H. Lyu, S. Jiang, H. Zeng, Y. Xia, Q. Wang, S. Zhang, R. Chen, C. Leung, J. Tang, and J. Luo, "Llm-rec: Personalized recommendation via prompting large language models," *arXiv preprint arXiv:2307.15780*, 2023.

[9] S. Dai, C. Xu, S. Xu, L. Pang, Z. Dong, and J. Xu, "Bias and unfairness in information retrieval systems: New challenges in the llm era," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 6437–6447.

[10] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of personality*, vol. 60, no. 2, pp. 175–215, 1992.

[11] B. Ferwerda, E. Yang, M. Schedl, and M. Tkalcic, "Personality and taxonomy preferences, and the influence of category choice on the user experience for music streaming services," *Multimedia tools and applications*, vol. 78, pp. 20 157–20 190, 2019.