Biases in LLM-Generated Musical Taste Profiles for Recommendation

Bruno Sguerra
Deezer Research
Paris, France
bmassonisguerra@deezer.com

Elena V. Epure
Deezer Research
Paris, France
eepure@deezer.com

Harin Lee
Max Planck Institute for Human Cognitive and Brain Sciences
Leipzig, Germany
Max Planck Institute for Empirical Aesthetics
Frankfurt am Main, Germany
hlee@cbs.mpg.de

Manuel Moussallam
Deezer Research
Paris, France
manuel.moussallam@deezer.com

Abstract

One particularly promising use case of Large Language Models (LLMs) for recommendation is the automatic generation of Natural Language (NL) user taste profiles from consumption data. These profiles offer interpretable and editable alternatives to opaque collaborative filtering representations, enabling greater transparency and user control. However, it remains unclear whether users consider these profiles to be an accurate representation of their taste, which is crucial for trust and usability. Moreover, because LLMs inherit societal and data-driven biases, profile quality may systematically vary across user and item characteristics. In this paper, we study this issue in the context of music streaming, where personalization is challenged by a large and culturally diverse catalog. We conduct a user study in which participants rate NL profiles generated from their own listening histories. We analyze whether identification with the profiles is biased by user attributes (e.g., mainstreamness, taste diversity) and item features (e.g., genre, country of origin). We also compare these patterns to those observed when using the profiles in a downstream recommendation task. Our findings highlight both the potential and limitations of scrutable, LLM-based profiling in personalized systems.

CCS Concepts

• **Information systems** → *Recommender systems*; • **Human-centered computing** → *User studies*.

ACM Reference Format:
Bruno Sguerra, Elena V. Epure, Harin Lee, and Manuel Moussallam. 2025. Biases in LLM-Generated Musical Taste Profiles for Recommendation. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25), September 22–26, 2025, Prague, Czech Republic.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3705328.3748030

RecSys '25, Prague, Czech Republic

ACM ISBN 979-8-4007-1364-4/25/09
https://doi.org/10.1145/3705328.3748030

1 Introduction

In 1967, the American writer John M. Culkin said "we shape our tools and thereafter our tools shape us", highlighting the feedback loops between human experience and technologies. Today, this dynamic is particularly evident in the rise of Large Language Models (LLMs), whose remarkable capabilities in text generation have far-reaching implications beyond academia and industry. Culkin's feedback loop also underscores a critical challenge: LLMs inherit human biases from their training data, which can result in discriminatory outcomes, especially when using sensitive attributes like gender [5]. As LLMs influence human decision-making and generate content that may be used for future training, there is a risk of reinforcing and amplifying existing biases over time [1].

In recommendation, a popular use of LLMs is to generate Natural Language (NL) user profiles from consumption data (item metadata and possibly explicit feedback) [19]. These profiles can then guide the recommendation, either via direct LLM prompting [20] or by embedding users and items in a shared space [10]. The resulting LLM textual profiles offer a scrutable alternative to latent user embeddings common in Collaborative Filtering (CF) [19, 20], allowing users to inspect and edit their profiles, thereby enhancing transparency and control. This approach also benefits cold-start cases, whereby new users can simply describe their preferences [19]. One might assume that, since LLMs are not primarily trained on user-item interactions, they might avoid data-driven biases such as popularity, exposure, and position bias, which are common in recommendation [3, 12, 15, 18, 25]. However, they introduce new challenges such as hallucinations [14], sensitivity to item order [26], and dependence on memorized content [2, 7]. Moreover, recent findings show that LLMs as recommenders tend to exhibit lower fairness and diversity than traditional CF methods, though fine-tuning and prompt design may help address these issues [5, 6].

Studies have shown that prompting LLMs with item metadata and feedback can yield coherent textual user summaries [10, 20, 28], with evaluation often focusing on downstream performance. However, it remains unclear to what extent users perceive these profiles

as accurate representations of themselves (i.e., their preferences). Moreover, whether (and how) profile representativeness and usability in recommendation vary with user attributes (e.g., mainstreamness) or item characteristics (e.g., music genre), particularly in combination with the social biases LLMs inherit from pre-training, remains an underexplored aspect.

In this work, we address this gap by evaluating profile quality along two dimensions: user-self identification, collected through a user study, and recommendation relevance, assessed in a downstream task. We argue that understanding how these dimensions relate to user and item characteristics is essential for assessing the feasibility of LLM-generated NL profiles, before optimizing them for downstream applications. In particular, such analysis is necessary to ensure that all user groups are fairly represented.

We address the following research questions: (1) To what extent users recognize their musical taste in automatically generated profiles? (2) Is profile quality biased towards user characteristics such as mainstreamness or consumption diversity, or item characteristics such as genre and country of origin? (3) Is user evaluation of profile quality aligned with relevance in recommendation settings?

We conduct our study in the context of music streaming, where modeling user preferences is especially challenging due to the scale and diversity of catalogs spanning genres, regions, and popularity levels. To support transparency and reproducibility, we release the source code of our experiments and the user study dataset (see Section 4).

2 Experimental Setup

A key lacking aspect of recent LLM summarization methods is human qualitative evaluation of their ability to correctly represent user preferences. Therefore, we conducted a user study to investigate potential biases in NL user profiles following two stages: (1) generating NL profiles in a systematic way to understand which features matter on the resulting quality; (2) conducting an online study (N participants = 64) to evaluate the generated profiles.

2.1 Building NL User Musical Profiles

Generating scrutable NL user profiles offers transparency but necessitates conciseness, a significant challenge in music recommendation. Unlike prior work often relying on explicit feedback [10, 20, 28], music recommendation frequently relies on implicit feedback, which is more noisy, making preference inference more uncertain [13, 22]. While using concatenated item attributes from consumption logs can help model preferences in implicit settings [27], in music consumption, users interact with vast catalogs [21], leading metadata-based representations to exceed LLM context limits. This makes item sampling strategies particularly important. Furthermore, musical taste blends long-term preferences with short-term exploration [24]. Thus, sampling only frequent tracks within short windows (e.g., one week) favors recency but is prone to noise. In contrast, longer time windows smooth out fluctuations but could miss more recent explorations. Also, since users listen to collections in the form of playlists and albums, simple frequency-based sampling can lead to redundant profiles due to repeated listens of the same collections.

To address these challenges, we propose a 2-step sampling strategy: (1) identify the top-n most played artists within a given time window; (2) select most played tracks per artist. This mitigates artist-level redundancy and allows the time window to control the balance between short- and long-term preference signals.

Figure 1 illustrates the increasing stability of sampled top artist-track sets across different time windows. Each point represents the mean Jaccard similarity across users (with standard deviation bars), comparing top tracks from a shorter window to those from a longer one. For example, the first point compares users' top tracks in a 7-day window ("7d") to those in a 30-day window ("30d"). The upward trend indicates that as the time windows increase, the sampled top items become more stable. In our study, we generate profiles using our two-step strategy across four time windows: two shorter ones (30 and 90 days), where preferences are still relatively unstable, and two longer ones (180 and 365 days), where preferences tend to plateau.

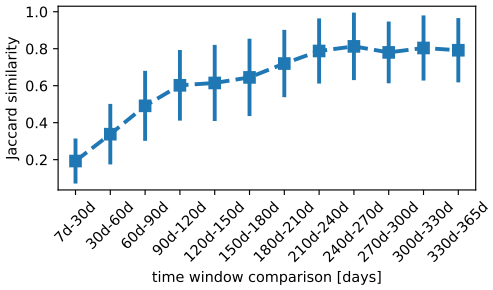

Figure 1: Mean and standard deviation of Jaccard similarity for users' top-15 tracks, across different time window lengths.

For NL summary generation, we follow prior work by prompting LLMs with a list of items described by textual metadata. For each user, we sample their top-15 artist–track pairs using our two-step strategy and collect metadata (track title, play count, artist name, album name, release year, artist main and secondary genres). The number of items was chosen to fit within a single-shot prompt, avoiding context length limitations. Prior work often uses fewer items (5 items in [10, 20]) and sometimes more (30 in [28]). However, our results show 15 items are enough for user recognition.

We do not provide example profiles in the prompt, allowing the model to propose its own structure. This choice enables us to assess which formats resonate best with users. Also, the prompt explicitly instructs the model not to overly rely on artist and track names, encouraging more abstract and high-level descriptions of preferences. We select two open-source LLMs, Llama 3.2 [23] and DeepSeek-R1 [11] (a reasoning-focused model), both available via the Ollama library [17], and Gemini 2.0 Flash [4], a proprietary model. We leave the output context window to their default values and set the temperature to 0.8. We include the full prompt and profile examples in the Appendix.

2.2 User Study

We recruited participants on a voluntary basis through internal communication channels within our organization (Deezer). Participation took place during paid working hours, and the study

was conducted online. Upon invitation, participants were explicitly informed that their streaming histories would be used to generate profiles for evaluation purposes.

A total of 64 participants took part in the study (40 males, 23 females, and one non-binary), aged between 22 and 48 years (33.73 ± 6.28). For them, we collected several high-level metrics, derived from 28-day consumption data on Deezer, a music streaming service: (1) Generalist-Specialist Score (GS-Score)[1], which captures whether a listener is a specialist (high GS-score) or a generalist (low GS-score); (2) the median rank of their listened tracks, relates to the popularity of the tracks based on the consumption of all users of Deezer is as a measure of mainstreamness, higher values indicating a preference for popular tracks; (3) mean age of songs indicates a preference for recent vs. older music.

In addition, we collected participants' consumption history for the year starting April 1, 2024, considering only streams with a listening time greater than 30 seconds[2]. From this data, we sampled items using the artist-track frequency method described in Section 2.1, across four time windows, and retrieved the corresponding metadata[3]. We then generated an NL profile using the three selected LLMs. Therefore, every user has 12 profiles.

The mean profile length (in characters) per model is: Gemini: 396.43 ± 70.62; DeepSeek: 543 ± 153.06; Llama: 638.16 ± 117.54. Gemini-based profiles being overall smaller than the others.

During evaluation, each participant was presented with 17 profiles: 12 personalized and five random profiles from other users serving as negative baselines. The evaluation begins with an introductory screen presenting an example profile to familiarize participants with the task. Then, the 17 profiles are presented in randomized order. For each profile, users are asked *"How well does the following description match your music taste?"*, and asked to respond with a 7-point Likert scale (1 being "not at all" and 7 being "very much"). After rating all profiles, users are asked about their musical engagement based on the Goldsmiths Musical Sophistication Index [16].

The user study was launched on April 9th 2025 and was concluded on the 11th. In the next section we discuss how the obtained user ratings relate to different user-item characteristics.

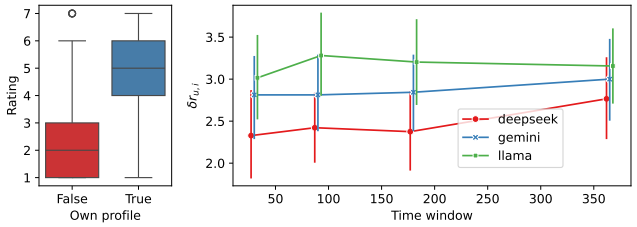

Figure 2: Left: box-plot of the user ratings for the true and false profiles. Right: the mean and 0.95 Confidence Interval (CI) for the profiles generated by each model at each time window.

[1]It quantifies the dispersion of songs users have listened to based on SVD embeddings derived from co-occurrence of tracks in playlists and personal collections.
[2]A threshold commonly used in the industry to define a positive interaction.
[3]Metadata sourced from Music Story: https://music-story.com/.

3 Biases in User-evaluated NL Profiles

To investigate how well the generated profiles represent users in relation to different user- and item- characteristics, for every user u, we compute the difference in rating between a true profile $p_i \in P_u^+$ and the median rating on their fake profiles P_u^- following: $\delta r_{u,i} = r_{u,i} - \text{median}\left(\{r_{u,j} : p_j \in \mathcal{P}_u^-\}\right)$, where $r_{u,i}$ is the rating user u gave to profile p_i. Therefore, instead of using absolute ratings, that vary depending on users, we use this difference as a measure of how well users believe a profile represents their taste in relation to the random baseline.

The left side of Figure 2 shows box plots of ratings for positive and negative profiles, indicating that users rated profiles based on their own consumption data higher than random ones. On the right, we report the mean $\delta r_{u,i}$ and 95% CI for each model across four time windows. Time window had only a marginal impact on profile ratings, and although confidence intervals overlap, Llama-based profiles consistently received higher scores than the other two. This suggests that, beyond content, the form, specific to each LLM, may play a greater role in self-recognition. Supporting this, a two-way ANOVA revealed a significant effect of model, $F(2, 1076) = 5.16$, $p = 0.0059$, but no effect of time window, $F(3, 1076) = 1.59$, $p = 0.191$, and no interaction, $F(6, 1076) = 0.28$, $p = 0.947$.

3.1 User-characteristics

Here we evaluate how the ratings relate to user characteristics that are not directly dependent on the tracks used for the profiles. Our goal is to assess whether the method exhibits biases towards specific user groups characterized on some dimension, therefore we investigate linear relationships.

Concerning gender, rating differences were small across all models. Female users gave slightly lower ratings than male users, with Cohen's d indicating a small effect for DeepSeek ($d = -0.15$) and Gemini ($d = -0.16$), and a negligible effect for Llama ($d = -0.07$).

To quantify the influence of other user characteristics on rating behavior, we perform 5K bootstrap simulations, where we fit linear regression models predicting $\delta r_{u,i}$ from each investigated characteristic (normalized) for each LLM separately. This process allows us to quantify the CI for the fitted linear coefficient. The orange bars on the top of Figure 3 show the mean coefficient with the black bar indicating the 95% CI. GS-score shows a significant positive association with $\delta r_{u,i}$, indicating that specialist users tend to rate the profiles more positively. This is consistent with the idea that specialist users, having narrower preferences, are more easily represented through the 15 sampled items, while generalist users may require more items for accurate representation. The mean age of consumed songs and user age also correlate positively with $\delta r_{u,i}$ across all models, suggesting that older users and users that listen to older songs tend to rate the generated profiles more favorably (these measures are both somewhat correlated Pearson = 0.22, p-value=0.048). User mainstreamness, measured via the median track rank, was not a significant predictor in any model (all CI go through 0), suggesting no clear relationship between mainstream taste and perceived profile quality, the same for active engagement.

3.2 Item-characteristics

We next investigate how sampled item characteristics relate to the quality of profiles rated by users, using the same bootstrapping approach described above.

We first examined whether the median play count of the sampled items relate to user ratings, Deepseek profiles showed a positive association ($mean(\beta) = 2.11$ $[0.22, 3.93]$), suggesting repeated interactions may enhance relevance. Llama profiles showed a slightly weaker effect ($mean(\beta) = 1.91$ $[0.31, 3.55]$), while no clear effect was observed for Gemini ($mean(\beta) = 0.83$ $[-0.92, 2.78]$).

Since each track is annotated with a single "main genre" (when available), we compute, for each profile, the proportion of tracks associated with each genre among the 15 sampled tracks. We also include analogous ratios based on country of origin, as well as the proportion of missing genre or country annotations. We compute entropy measures over these distributions. Together, these features enable us to investigate how the compositional profile characteristics relate to their perceived quality. Other than the entropy measures, in the following analysis, we consider only genres and countries that appear in the listening histories of at least 20% of the users. The set of investigated genres are: *electronic, jazz, metal, pop, rap, rock, soul, world music*. The set of investigated countries: *Belgium, Brazil, Canada, France, Germany, United Kingdom, United States*.

We perform 5K bootstrap simulations to estimate the linear association between these ratios and $\delta_{u,i}$. Figure 3 shows the bootstrapped mean coefficients and 95% confidence intervals. We exclude associations with intervals crossing zero or with width greater than 10, as these indicate high uncertainty and lack of a reliable effect. The results suggest that specific item characteristics are consistently linked to profile quality, for example, a higher proportion of rap tracks is negatively associated with ratings across all models, while a greater share of U.S.-origin tracks correlates positively.

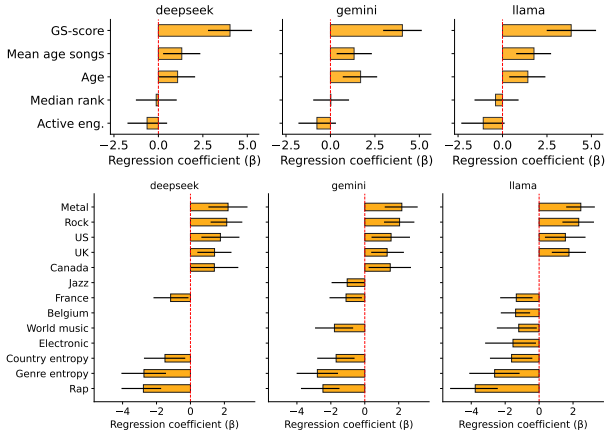

Figure 3: Bootstrapped estimated coefficients of user characteristics and genre and country composition on profile ratings.

3.3 Biases in NL profiles

While the linear coefficients shown above suggest that LLMs generate some profiles better than others over certain item features, this may reflect confounding factors, such as user preferences, rather than true model biases. For instance, profiles dominated by genres users already prefer (e.g., metal for a metal fan) may receive higher ratings simply due to taste alignment. In this case, being a metal fan acts as a confounder, influencing both the sampled items used to generate the profile and the user's evaluation of it. As a result, the observed correlation between genre ratio and rating may be spurious with respect to assessing the LLM's ability to generalize preferences. Additionally, users with a strong genre presence may be easier to summarize, particularly for specialists (top of Figure 3).

To disentangle these effects, we apply a Doubly Robust (DR) estimation of the Average Treatment Effect (ATE) [9]. This method combines outcome regression and propensity score modeling, yielding consistent estimates if either model is correctly specified. For a binary treatment variable T, the ATE is: $ATE = \frac{1}{N}\sum(\frac{T_i(Y_i - \hat{Y}_1(X))}{PS(X)} + \hat{Y}_1(X)) - \frac{1}{N}\sum(\frac{(1-T_i)(Y_i - \hat{Y}_0(X))}{1-PS(X)} + \hat{Y}_0(X))$, where, X is the covariates, Y is the outcome (in our case $\delta r_{u,i}$), $\hat{Y}_1$ and $\hat{Y}_0$ are predicted outcomes given X (via gradient boosting regression) under treatment and control, and $PS(X)$ is the propensity score from logistic regression.

To control for taste and exposure biases, we construct covariates X from each user's long-term genre ratios, the genre composition of profile items (to capture underrepresentation effects), user age, and GS-score. We restrict the analysis to the genres shown in Figure 3 to reduce collinearity from rare categories like country or blues. Treatment is defined as having a genre ratio above the median, and the ATE quantifies the effect of including more content from this genre on user ratings, while controlling for confounders. For each genre and model, we estimate ATE and 95% CI via 1K bootstrap simulations (Figure 4).

The bootstrapped ATE estimates in Figure 4 show that, even after controlling for our covariates, the presence of certain content types in the generated profiles still significantly influences user ratings. This is especially evident for rap and metal: across all models, rap-related content is associated with consistently negative ATEs, while metal shows positive ones. This suggests that profiles with a higher proportion of rap tracks tend to receive lower ratings. Sensitivity to content types also varies by model: for example, Gemini and DeepSeek are largely unaffected by the presence of Belgian and world music, while Llama see drops in profile quality. Also both DeepSeek and Llama see a positive effect from Canadian music that is not reliable for Gemini (CI crosses zero).

4 Downstream Evaluation

We assess whether profile ratings correlate with downstream recommendation performance. Investigating if higher-rated NL profiles also yield better recommendations, essential for ensuring both transparency and utility. We highlight that the goal is not to optimize recommendation performance, but rather to analyze its relation with user evaluations.

Following prior work, we adopt a recommendation approach by embedding NL profiles and item metadata into a shared latent

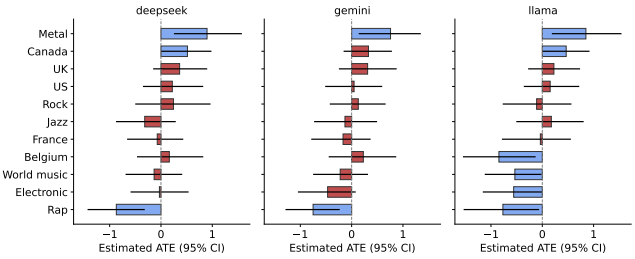

Figure 4: Estimated ATE of genre presence in profiles on ratings, controlling for covariates. Blue bars indicate significant effects; red bars have 95% CIs crossing zero, suggesting non-significance.

space [10, 19]. Profiles are used as-is, while each track has associated metadata textual representation, as used in the prompts. Given both items and users textual representations, we fine-tune the bi-encoder proposed in [8] on our dataset by using their music-specific ranker as a teacher[4].

We use all the generated profiles for training. For each profile, the sampled tracks for each user–time window pair serve as positive examples, while negative examples are sampled from tracks the user has not interacted with in the year considered. In total, the training set comprised 100K user-positive-negative triplets. We train the model for 1 epoch using an AdamW optimizer, a batch size of 10, and a maximum sequence length fixed to 512. To test the model, for each user, we sampled 10 items the user has interacted with (but not used for profile generation) and 1K negative items, all of which not seen in the training set with that specific user. Our analysis is based on recall@10 and ndcg@10 metrics from the model's similarity scores, averaged over 5 runs with different random seeds. The code for implementation is made public in our repository[5].

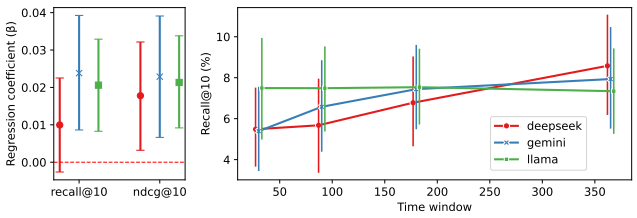

Figure 5: Left: Bootstrapped linear coefficients predicting recommendation scores from ratings. Right: Mean Recall@10 with 95% CIs per model and time window.

The left side of Figure 5 illustrates how $\delta r_{u,i}$ relates to recall@10 and NDCG@10, based on bootstrapped linear regression models that predict the normalized metrics from user ratings. The coefficient reflects the expected change in performance per one-point increase in rating. Except for Deepseek, all other metrics and models show a weak positive linear relationship. We provide more details on how recommendation performance relates to user and item characteristics in the Appendix. On the right side of Figure 5, we show the mean recall@10 with 95% CI for the different models and time windows. We note that the pattern is quite different from that in Figure 2. Except for Llama, the recommendation performance seems to improve with the longer time window, which can partially be explained by the fact that the test items were sampled from the entire consumption year.

[4]The ranker is implemented as a cross-encoders that take as input a music description concatenated with a set of music tags and produces as output a similarity score.
[5]https://github.com/deezer/recsys25_llm_biases

5 Discussion and Conclusion

This paper evaluated the quality of LLM-generated NL user profiles along two dimensions: user ratings and downstream recommendation performance. We compared profiles generated from items sampled over varying time windows in users' listening histories and found that window choice had limited impact on ratings. As long as key preferences were captured, users generally recognized themselves. While some noted missing aspects of their taste (see Appendix), variation in ratings was primarily driven by the choice of LLM. This likely reflects stylistic differences: Gemini favored concise, high-level summaries; Llama emphasized specific artists and tracks (despite prompt instructions), which some users preferred; DeepSeek occasionally included metadata like "remastered", which users found less meaningful. However, given the diversity of user preferences and model-specific tendencies, deeper qualitative analysis is needed to understand which profile aspects users favor.

We further examined how user and item characteristics influenced profile quality. Profiles of specialist users (those with high GS-scores) tended to receive higher ratings, suggesting their tastes may be easier to capture in NL. Item characteristics, such as genre and country ratios, were also correlated with ratings, both positively and negatively. To account for potential confounds such as taste alignment, we employed a DR framework. The resulting ATE values confirmed that some content types systematically influenced ratings: for instance, rap content tended to reduce scores, while metal increased them, within the limits of our small user sample. Although it remains unclear whether these effects stem from metadata quality or model behavior, the fact that different LLMs responded differently to the same metadata suggests model-specific biases in content representation. These findings raise fairness concerns, as some users may consistently receive more representative profiles than others. Given that LLMs can inherit biases from the textual data on which they are trained, communities that are underrepresented in digital corpora, particularly non-English-speaking ones, may be disadvantaged. We note, however, that our study included only three LLMs, and conclusions may not generalize broadly.

Moreover, user ratings and recommendation performance were only weakly correlated, revealing a misalignment between perceived representativeness and algorithmic utility. This is critical: if users do not recognize themselves in recommendation-optimized profiles, they may lose trust in the system. One potential solution is to decouple profile generation for user feedback from those optimized solely for recommendation.

Future work should explore targeted fine-tuning or debiasing strategies to reduce representation gaps. In production systems, LLMs could be refined using feedback from user-corrected profiles, aiming to better balance personalization and fairness.

References

[1] Levin Brinkmann, Fabian Baumann, Jean-François Bonnefon, Maxime Derex, Thomas F Müller, Anne-Marie Nussberger, Agnieszka Czaplicka, Alberto Acerbi, Thomas L Griffiths, Joseph Henrich, et al. 2023. Machine culture. *Nature Human Behaviour* 7, 11 (2023), 1855–1868.

[2] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646* (2022).

[3] 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.

[4] Google Cloud. 2025. Gemini 2.0 Flash: Generative AI on Vertex AI. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash. Accessed: 2025-04-30.

[5] Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. 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*. 6437–6447.

[6] Yashar Deldjoo. 2024. Understanding biases in chatgpt-based recommender systems: Provider fairness, temporal stability, and recency. *ACM Transactions on Recommender Systems* (2024).

[7] Dario Di Palma, Felice Antonio Merra, Maurizio Sfilio, Vito Walter Anelli, Fedelucio Narducci, and Tommaso Di Noia. 2025. Do llms memorize recommendation datasets? a preliminary study on movielens-1m. *arXiv preprint arXiv:2505.10212* (2025).

[8] Elena V. Epure, Gabriel Meseguer-Brocal, Darius Afchar, and Romain Hennequin. 2024. Harnessing High-Level Song Descriptors towards Natural Language-Based Music Recommendation. In *Proceedings of the 3rd Workshop on NLP for Music and Audio (NLP4MusA2024)*. Association for Computational Linguistics.

[9] Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. 2011. Doubly robust estimation of causal effects. *American journal of epidemiology* 173, 7 (2011), 761–767.

[10] Zhaolin Gao, Joyce Zhou, Yijia Dai, and Thorsten Joachims. 2024. End-to-end Training for Recommendation with Language-based User Profiles. *arXiv preprint arXiv:2410.18870* (2024).

[11] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[12] Jin Huang, Harrie Oosterhuis, Masoud Mansoury, Herke Van Hoof, and Maarten de Rijke. 2024. Going beyond popularity and positivity bias: Correcting for multifactorial bias in recommender systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 416–426.

[13] Gawesh Jawaheer, Peter Weller, and Patty Kostkova. 2014. Modeling user preferences in recommender systems: A classification framework for explicit and implicit user feedback. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 4, 2 (2014), 1–26.

[14] Chumeng Jiang, Jiayin Wang, Weizhi Ma, Charles LA Clarke, Shuai Wang, Chuhan Wu, and Min Zhang. 2024. Beyond Utility: Evaluating LLM as Recommender. *arXiv preprint arXiv:2411.00331* (2024).

[15] Kristina Matrosova, Lilian Marey, Guillaume Salha-Galvan, Thomas Louail, Olivier Bodini, and Manuel Moussallam. 2024. Do Recommender Systems Promote Local Music? A Reproducibility Study Using Music Streaming Data. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 148–157.

[16] Daniel Müllensiefen, Bruno Gingras, Jason Musil, and Lauren Stewart. 2014. The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PloS one* 9, 2 (2014), e89642.

[17] Ollama. 2025. Ollama: Open-Source Large Language Models. https://github.com/ollama/. Accessed: 2025-04-30.

[18] Pantelis Pipergias Analytis and Philipp Hager. 2023. Collaborative filtering algorithms are prone to mainstream-taste bias. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 750–756.

[19] Filip Radlinski, Krisztian Balog, Fernando Diaz, Lucas Dixon, and Ben Wedin. 2022. On natural language user profiles for transparent and scrutable recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2863–2874.

[20] Jerome Ramos, Hossein A Rahmani, Xi Wang, Xiao Fu, and Aldo Lipani. 2024. Transparent and Scrutable Recommendations Using Natural Language User Profiles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 13971–13984.

[21] Markus Schedl. 2019. Deep learning in music recommendation systems. *Frontiers in Applied Mathematics and Statistics* 5 (2019), 457883.

[22] Bruno Sguerra, Viet-Anh Tran, Romain Hennequin, and Manuel Moussallam. 2025. Uncertainty in Repeated Implicit Feedback as a Measure of Reliability. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*.

[23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[24] Viet-Anh Tran, Guillaume Salha-Galvan, Bruno Sguerra, and Romain Hennequin. 2024. Transformers Meet ACT-R: Repeat-Aware and Sequential Listening Session Recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 486–496.

[25] Robin Ungruh, Karlijn Dinnissen, Anja Volk, Maria Soledad Pera, and Hanna Hauptmann. 2024. Putting Popularity Bias Mitigation to the Test: A User-Centric Evaluation in Music Recommenders. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 169–178.

[26] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926* (2023).

[27] Lanling Xu, Junjie Zhang, Bingqian Li, Jinpeng Wang, Mingchen Cai, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Prompting large language models for recommender systems: A comprehensive framework and empirical analysis. *arXiv preprint arXiv:2401.04997* (2024).

[28] Joyce Zhou, Yijia Dai, and Thorsten Joachims. 2024. Language-Based User Profiles for Recommendation. *arXiv preprint arXiv:2402.15623* (2024).