A Study of Biases in LLM-Generated Musical Taste Profiles for Recommendation

BRUNO SGUERRA, Research, Deezer, Paris, France
ELENA EPURE, Deezer Research, Paris, France and Idiap Research Institute, Martigny, Switzerland
HARIN LEE, University of Cambridge, Cambridge, United Kingdom of Great Britain and Northern Ireland
MANUEL MOUSSALLAM, Research, Deezer, Paris, France

Large Language Models (LLMs) offer a promising approach to recommendation by enabling the generation of user profiles in Natural Language (NL) form. When used as summarization devices, LLMs can produce interpretable and editable alternatives to opaque collaborative filtering representations, potentially increasing transparency and user control. However, it remains unclear whether users perceive these profiles as accurate representations of their preferences, which is key 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 investigate these issues in the context of music streaming, where personalization is challenged by large and culturally diverse catalogs. We conduct a user study in which participants evaluate NL profiles generated from their own listening histories.

We analyze whether user identification with these profiles is biased by user attributes, such as mainstreamness and taste diversity, and by item features, including genre and country of origin. We further assess the usefulness of the generated profiles in a downstream recommendation task by analyzing their representations in a shared embedding space. Our results reveal systematic differences across models and user groups, highlighting both the potential and the limitations of scrutable, LLM-based profiling for personalized systems.

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

1 Introduction

LLMs are trained on vast amounts of human-written text, absorbing not only its knowledge but also its biases and cultural blind spots. As LLMs are increasingly integrated into everyday applications, from information access and writing assistance to decision-making and recommender systems, these inherited limitations risk being carried over into generated content, potentially leading to uneven or unfair outcomes across users and domains [6]. And since LLM-generated content may itself become part of future training corpora, this creates a feedback loop that can reinforce and amplify existing limitations over time [2]. Understanding and measuring these biases is therefore essential for building fair and accountable LLM-based systems.

One particularly promising application of LLMs is the generation of Natural Language (NL) user profiles from consumption data, including item metadata and, when available, explicit feedback [22]. These profiles can guide recommendations either through direct prompting of LLMs [23] or by embedding users and items in a shared representation space [12]. Compared to the latent user embeddings commonly used in Collaborative Filtering (CF) approaches [22, 23], LLM-generated textual profiles provide a more scrutable representation of user

Authors' Contact Information: Bruno Sguerra, Research, Deezer, Paris, France; e-mail: bmassonisguerra@deezer.com; Elena Epure, Deezer Research, Paris, France and Idiap Research Institute, Martigny, VS, Switzerland; e-mail: eepure@deezer.com; Harin Lee, University of Cambridge, Cambridge, Cambridgeshire, United Kingdom of Great Britain and Northern Ireland; e-mail: hl744@cam.ac.uk; Manuel Moussallam, Research, Deezer, Paris, France; e-mail: manuel.moussallam@deezer.com.

ACM 2770-6699/2026/6-ART
https://doi.org/10.1145/3815194

preferences. This allows users to inspect and edit their profiles, thereby enhancing transparency and control. Such representations are also well suited to cold-start scenarios, as new users can directly describe their interests [22].

One might expected that, since LLMs are not primarily trained on user–item interaction data, they would be less affected by traditional data-driven biases such as popularity, exposure, and position bias [4, 14, 17, 21, 31]. In practice, however, LLM-based recommenders introduce new challenges, including hallucinations [16], sensitivity to item ordering [33], and reliance on memorized content [3, 9]. Recent studies further indicate that LLM-based recommender systems tend to exhibit lower fairness and diversity than traditional CF methods, although mitigation strategies such as fine-tuning and careful prompt design may partially address these issues [6, 7].

In the specific context of NL user profiles, previous work has shown that prompting LLMs with item metadata and user feedback can yield coherent textual user summaries [12, 23, 37], with evaluation typically focusing on downstream performance. However, it remains unclear to what extent users perceive these profiles as accurate representations of themselves. This concern is particularly salient in music, where textual coverage is shaped by cultural visibility, market prominence, and mainstream validation: globally popular artists and genres tend to be surrounded by richer descriptions, while niche, local, or country-specific repertoires may be underrepresented in online text. As a result, LLMs may represent some forms of musical consumption more accurately than others. Whether and how profile representativeness varies with user characteristics (e.g., mainstreamness) or item properties (e.g., genre or country of origin), particularly in combination with biases inherited from pre-training, remains largely unexplored.

In this work, we address this gap by evaluating profile quality through user self-identification, collected in a dedicated user study. We argue that understanding how user assessments relate to user and item characteristics is essential for establishing the feasibility of LLM-generated NL profiles before optimizing them for downstream tasks. Such analysis is necessary to ensure that all user groups are fairly represented and accounted for. In a second step, we employ the generated profiles in a downstream recommendation task by encoding both user profiles and item metadata into a shared latent space. This allows us to examine whether user recognition aligns with recommendation performance. By analyzing these representations, we gain deeper insight into the properties of the generated profiles and the specific behaviors of different LLMs.

Accordingly, we address the following research questions:

(1) To what extent do users recognize their musical taste in automatically generated profiles?
(2) Is profile quality biased toward user characteristics such as mainstreamness or consumption diversity, or toward 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 particularly 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 many recent works when leveraging LLMs for recommendation is human qualitative evaluation. This is particularly critical in settings where LLMs are used to generate user descriptions, in order to assess 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.

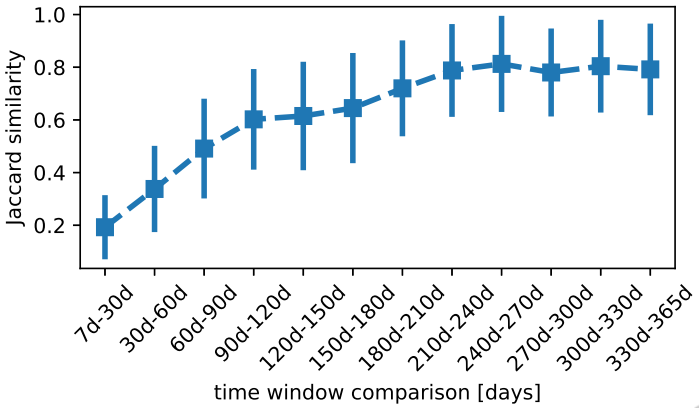


Fig. 1. Mean and standard deviation of Jaccard similarity for users' top-15 tracks, across different time window lengths. Note that the increasing similarity across longer windows is driven by the accumulation of listening history: longer windows filter out short-term fluctuations, producing more stable top-track sets.

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 [12, 23, 37], music recommendation frequently relies on implicit feedback, which is more noisy, making preference inference more uncertain [15, 28]. While using concatenated item attributes from consumption logs can help model preferences in implicit settings [35], in music consumption, users interact with vast catalogs [25], 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 [30]. 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.

For NL summary generation, we follow prior work by prompting Large Language Models (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 [12, 23]) and sometimes more (30 in [37]). However, our results show 15 items are enough for user recognition.

We select two open-source LLMs, Llama 3.2 [29] and DeepSeek-R1 [13] (a reasoning-focused model), both available via the Ollama library [20], and `Gemini 2.0 Flash` [5], a proprietary model. We leave the output context window to their default values and set the temperature to 0.8. We give more details concerning the prompt in the next section.

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, which relates to the popularity of the tracks based on the consumption of all users of Deezer, 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].

For generating the profiles, we used the same prompt template with different lists of items. All prompts consisted of a fixed preamble, followed by item metadata, and ended with the instruction "Only return the textual summary — no explanation, no lists." 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.

[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/.

Prompt example

```
Here is a list of items consumed by a single user. From this list, generate a short and scrutable
summary of their musical taste for music recommendation purposes. The summary should be a single,
coherent, and informative paragraph that clearly conveys the user's preferences in a natural,
human-readable way.
Base the description on the following list of consumed items and their metadata, without focusing
on the names of tracks and artists. The items are presented in descending order:

title: This is a Mood
playcount: 19
artist: Franc Moody
origin country: United Kingdom
album release date: 2020
main genre: Electronic music
secondary genre: Nu-disco, Synthpop, Electronica

title: CAFé CON RON
playcount: 11
artist: Bad Bunny
origin country: Puerto Rico
album release date: 2025
main genre: Rap music
secondary genre: Caribbean music, Dancehall, Dance rap, Latin hip hop, Latin
. . .
Only return the textual summary — no explanation, no lists.
```

For illustration purposes, from the complete prompt above, the following profiles were generated:

- **DeepSeek** (user rating: 4.0.): *The user has a strong preference for electronic music, particularly in genres like nu-disco, synthpop, and EDM, supported by notable tracks such as "Mood" and "Feel So Close." They also enjoy pop music with elements of Latin influence, as seen in works by Shakira and others. Additionally, their tastes include rap and rock, showcasing a diverse yet cohesive musical style that blends electronic, pop, and Latin flavors.*

- **Gemini** (user rating: 7.0.): *This user enjoys a diverse mix of music, with a strong preference for electronic music, particularly nu-disco and synthpop from the United Kingdom. They also have a significant interest in Latin music, including Latin pop, reggaeton, and Latin hip hop from various Latin American countries and Puerto Rico, alongside some mainstream pop and occasional soul and rock influences. They also seem to use music for relaxation or background ambiance, as indicated by tracks explicitly intended for sleep or relaxation.*

- **Llama** (user rating: 4.0.): *This user's musical taste leans towards a blend of electronic and pop elements, with a strong appreciation for global influences, particularly from Latin America and the Caribbean. They seem to enjoy upbeat and energetic tracks, often incorporating tropical rhythms and melodies, as evidenced by their fondness for artists like Bad Bunny, Shakira, and Daddy Yankee. There's also a noticeable interest in calming and atmospheric soundscapes, such as the instrumental cover of "Feel So Close" and the ambient piece "Dog Music". This user appears to have a diverse musical palette, but their overall taste tends towards lively and danceable tracks with Latin flair.*

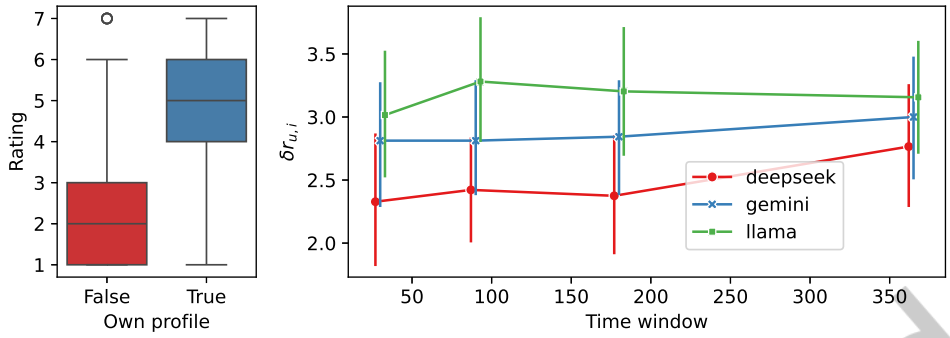


Fig. 2. Left: box-plot of the user ratings for the false and true profiles. Right: the mean and 0.95 Confidence Interval (CI) for the profiles generated by each model at each time window.

The mean profile length (in characters) per model is as follows: `Gemini`: 396.43±70.62; DeepSeek: 543.00±153.06; and Llama: 638.16±117.54. Profiles generated by `Gemini` are, on average, substantially shorter than those produced by the other models.

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 Goldsmiths Musical Sophistication Index (Gold-MSI) [19].

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.

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. Figure 3 shows the distribution of the considered user characteristics of the participants. We recall GS-score to be a measure of how specialist or generalist a user is (higher values, more specialist), median

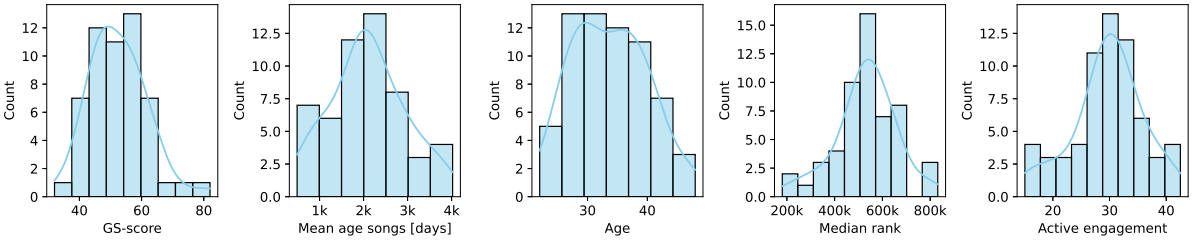


Fig. 3. Distribution of the users' characteristics.

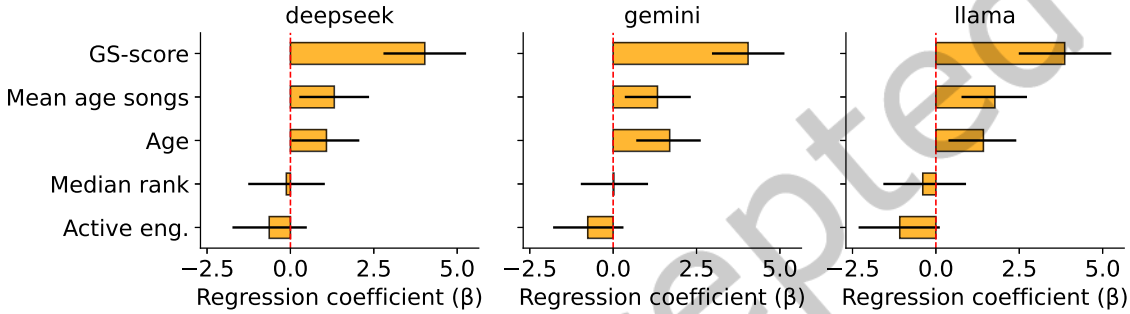


Fig. 4. Bootstrapped estimated coefficients of user characteristics on profile ratings.

rank of the listened tracks refers to a measure of track popularity (high rank signifies a preference for more popular tracks). We note that the active engagement of our participants $mean = 29.44 \pm 5.86$ is notably lower than the one reported for a larger population in [19] $mean = 41.52 \pm 10.36$, which suggests that although the participants are employees of a music streaming platform, they do not necessarily interact more intensively with music. Since our goal is to assess whether the method exhibits biases towards specific user groups characterized on some dimension, 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 Figure 4 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, $p = 0.048$). 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.

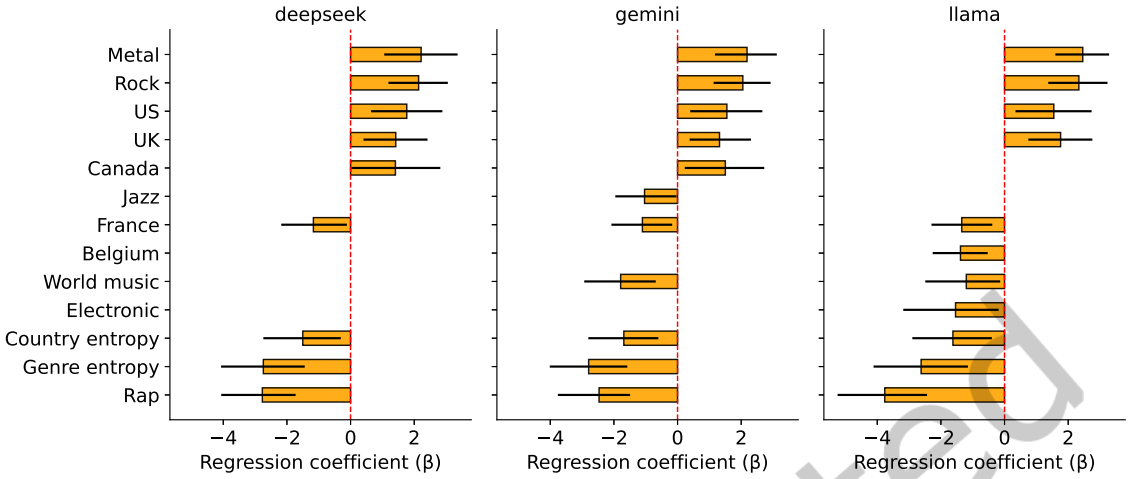


Fig. 5. Bootstrapped estimated coefficients of genre and country composition on profile ratings.

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 5 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.

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

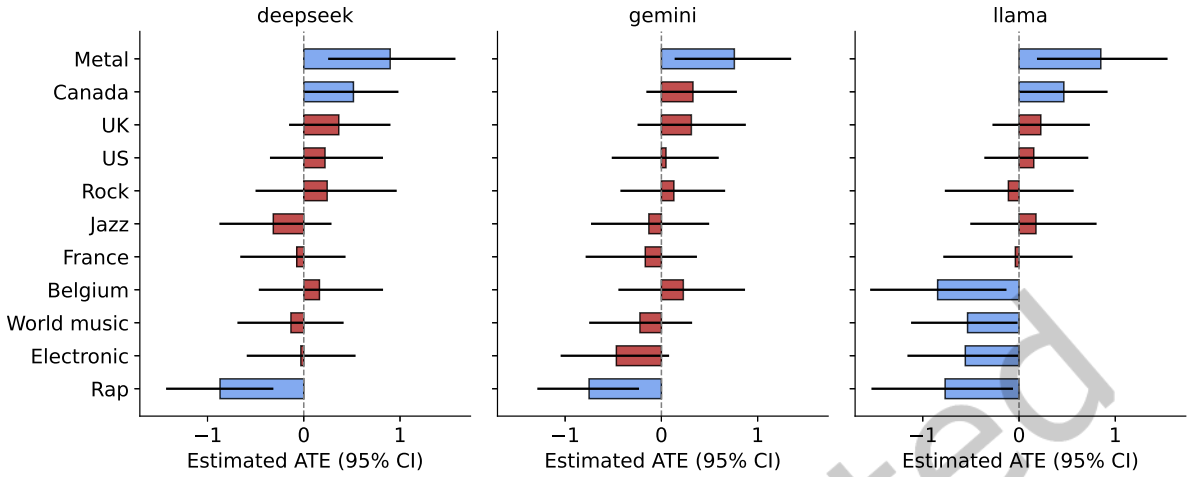


Fig. 6. Estimated ATE of genre presence in profiles on ratings, controlling for covariates. Blue bars indicate significant effects; red bars have 95% CIs crossing zero, suggesting non-significance.

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 (Figure 4).

To disentangle these effects, we apply a Doubly Robust (DR) estimation of the Average Treatment Effect (ATE) [11]. 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:

$$ATE = \frac{1}{N}\sum\left(\frac{T_i(Y_i - \hat{Y}_1(X))}{PS(X)} + \hat{Y}_1(X)\right) - \frac{1}{N}\sum\left(\frac{(1-T_i)(Y_i - \hat{Y}_0(X))}{1-PS(X)} + \hat{Y}_0(X)\right),$$

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 users' long-term genre ratios, the genre composition of profile items (capturing underrepresentation effects), user age, and GS-score. To reduce collinearity from rare categories (e.g., country or blues), we restrict the analysis to the genres shown in Figure 5. Treatment is defined as having a genre ratio above the median. The ATE measures the impact of the inclusion of a given genre on the user ratings while controlling for confounders. For each genre and model, we estimate the ATE and its 95% confidence interval using 1K bootstrap simulations (Figure 6).

The bootstrapped ATE estimates in Figure 6 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).

3.4 Qualitative Analysis: Rap vs Metal

Given the opposing treatment effects observed for metal and rap in the analysis above, we investigate these profiles in more detail. We selected profiles with above-median genre ratio for each genre and ran a bootstrap analysis (10K iterations) comparing lexical diversity (type-token ratio, MTLD, average word length, lexical density) and syntactic structure (average sentence length, number of sentences, verb variety, adjective density). We found a small but significant difference only for lexical density (Cohen's $d = -0.186$), with rap profiles being slightly more lexically dense than metal ones. All other metrics showed no reliable difference. Qualitative inspection of the highest-genre-ratio profiles reveals three systematic differences. First, metal profiles consistently foreground the genre name unambiguously, while rap profiles reframe the genre through geographic or cultural lenses even when rap dominates the listening history (e.g., "blend of Latin and Caribbean rhythms" for a user with 87% rap content). Second, French rap, which is prominent across multiple rap-heavy users, is consistently described as niche or regional (e.g., "underground French scene", "world music influences") rather than as a mainstream genre, potentially alienating users who identify primarily as rap listeners. Third, metal profiles anchor genre identity through universally recognizable landmark artists (Metallica, Slipknot, Iron Maiden), while rap profiles either reference niche French underground artists or resort to generic subgenre labels without artist anchors. Together, these patterns suggest that the performance gap is driven by representational choices in the LLM rather than text quality, and that it is particularly pronounced for listeners of non-English-language rap.

4 Downstream Evaluation

In this section, we evaluate the generated profiles in a downstream recommendation task. The code for implementation is made public in our repository[4]. The ultimate purpose of automatically generating natural language user taste profiles is to leverage them for downstream recommendation. In the original study [26], the evaluation of recommendations based on these profiles used a test set constructed by uniformly sampling consumed items (albeit not used for profile generation) over the entire time period considered. While this provides a controlled benchmark, it does not reflect the real-world usage scenario where profiles are generated at a given point in time and then used to predict *future* user consumption.

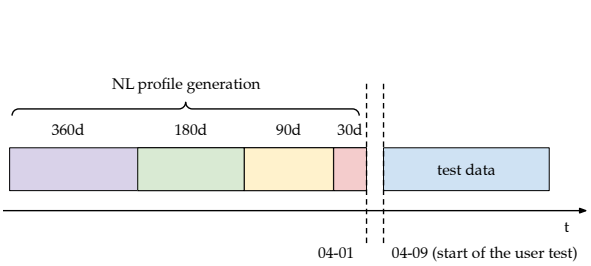
4.1 Test Set Definition

In practical settings, once an NL profile is created from a user's past history, the system's task is to anticipate what the user will want to consume going forward. To better approximate this temporal prediction setting, we instead construct our test sets from items that each user consumed after the start of the evaluation period (April 9, 2025), ensuring that recommendations are evaluated on genuinely future interactions rather than randomly sampled past ones (see Figure 7a).

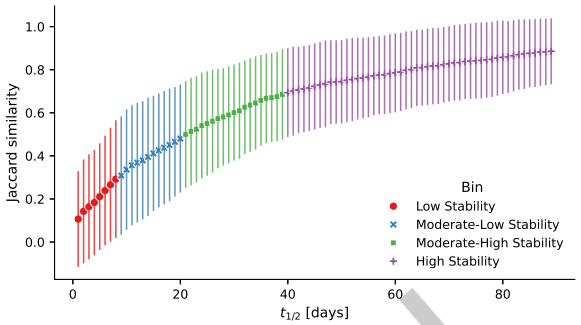
For building the test set, we aim to select items that best reflect a user's preferences around a given reference time, namely the start of the test period. However, listening behavior is inherently noisy [28]: individual plays may reflect momentary curiosity or contextual factors rather than stable preferences. Consequently, single interactions are often unreliable indicators of long-term interest. Rather than relying on the first interaction after the reference point, we focus on tracks that users have consumed multiple times, as repeated interactions are more indicative of sustained engagement and stable preferences [27]. At the same time, musical tastes evolve over time, and interactions that occur far from the reference point become progressively less informative of a user's preferences at test time. Extending the considered consumption window too far may therefore capture a taste profile that deviates from the user's preferences at the reference time.

To balance interaction frequency and temporal relevance, we assign each interaction between user u and item v a time-decayed relevance weight $r(u, v, t) = exp(-t/\tau)$, where t denotes the elapsed time (in days) since the

[4]https://github.com/deezer/TORS_llm_biases



(a) Test item sampling scheme. Recommendations are evaluated by predicting items consumed after the start of the evaluation period.



(b) Stability of sampled items under different half-life values.

Fig. 7. Test set construction and stability analysis. Left: sampling scheme. Right: Jaccard similarity of top-3 items sampled with different half-lives compared with non-decaying sampling.

interaction relative to the reference time, and $\tau = t_{1/2}/ln(2)$, with $t_{1/2}$ being the half-life of the item's relevance. For example, a half-life of two days indicates that two days after the reference point, that item's relevance is halved from the start. The half-life parameter provides an interpretable way to control how quickly interactions lose influence. We adopt an exponential decay with a half-life parameter because it offers a simple and principled trade-off between emphasizing behavior close to a reference time and retaining information from repeated interactions. By summing these relevance weights and selecting the top-K items per user, we aim to construct a test set that is robust to behavioral noise, such as short-term exploration, while remaining representative of users' preferences at the reference time.

We collect user streaming data from April 9, 2025, to June 30, 2025. To select representative half-life values, we vary $t_{1/2}$ and, for each setting, compare the top-3 items per user with the top-3 items obtained without temporal decay (i.e., equal weights). Figure 7b reports the mean Jaccard similarity and corresponding std between each half-life setting and the non-decayed baseline.

We interpret this similarity as a measure of preference stability. High similarity indicates that the selected items are largely insensitive to temporal decay, reflecting stable long-term preferences. Low similarity suggests that recent interactions substantially alter the selected items, indicating more volatile or rapidly evolving tastes. Intermediate values correspond to moderate stability, where both long-term and recent behaviors contribute meaningfully. To capture these different stability regimes, we discretize the Jaccard similarities into four linear bins (Figure 7b) and take the floor of the mean half-life within each bin to obtain four representative values: 4, 14, 30, and 64 days. For each half-life value, we sample 10 positive items per user and 1K negative items. To account for sampling variability, we repeat this procedure using five different random seeds, resulting in a total of 20 test sets (4 half-life values × 5 seeds).

4.2 Encoding user and item Representation from NL Descriptions

Following prior work, we adopt a recommendation approach by embedding NL profiles and item metadata into a shared latent space [12, 22]. Profiles are used as-is, while each track is associated with a textual metadata representation, as used in the prompts. Importantly, not all items have the same amount of available metadata, and the handling of missing information can substantially affect the resulting representations. A portion of tracks lack key attributes, such as the main genre (about 10%) or the artist name (9.8%). This occurs because metadata collection is often imperfect and subject to systematic biases and degraded quality for the long tail; for instance,

among tracks originating from Ukraine, 12% are missing genre information, compared to a median of 2% across other countries.

Filling missing fields with placeholder values like "unknown" may introduce spurious signals, as encoder models can learn to associate these specific tokens with patterns of data scarcity, thereby biasing the resulting vector representations. To try and mitigate this, we generate textual descriptions using only the available metadata, omitting missing fields entirely. Consequently, items are represented with varying degrees of granularity, ranging from complete descriptions (e.g., "title: Smash, artist: The Offspring, origin country: United States of America, album release date: 1994, main genre: Rock music, secondary genre: Grunge") to more sparse, attribute-reduced sets.

Unlike the original study [26], where we fine-tuned a single text encoder, in this work we instead evaluate four pretrained text encoders without additional fine-tuning. This decision is motivated by several considerations. First, fine-tuning large text encoders—particularly those based on LLM architectures, typically requires substantially more data than the 1200 user profiles available as positive examples in our setting (100 users × 3 LLMs × 4 time windows; not all users completed the study). Second, we aim to disentangle recommendation performance from the intrinsic representation capacity of the encoders, allowing us to assess how well generated user profiles align with encoded item representations without further optimization. Third, the original approach relied on a teacher–student paradigm [10], where a cross-encoder fine-tuned on music genres and tags acted as the teacher. In practice, we observe that many modern pretrained encoders already capture genre distinctions and tag semantics effectively, a result we further analyze in Section 4.2.2.

The primary criterion for selecting the encoders was that they be open-source and achieve strong retrieval performance within their size category on the MTEB leaderboard.[5]. In addition, we aimed to cover diverse architectures, training objectives, and datasets. Specifically, `embeddinggemma-300m` represents an LLM-based embedding approach built on a T5 encoder–decoder architecture with Gemma3 as its backbone decoder [32]. `gte-multilingual-base` is a mid-sized, encoder-only model optimized for multilingual retrieval and representation learning [36]. Finally, `all-MiniLM-L6-v2` and `msmarco-bert-base-dot-v5` are widely used sentence-transformer baselines [24], compact models based on BERT [8] and MiniLM [34], respectively, and fine-tuned on different retrieval datasets. We did not include larger LLM-based embedding models due to their higher computation cost, which would make deployment in a recommender system less practical. Overall, this selection enables us to evaluate the robustness of our approach across diverse embedding families, rather than relying on a single representation paradigm.

4.2.1 User Representation. Investigating the structure of the encoded profiles provides insight into the characteristic patterns induced by each LLM. Figure 9 presents a two-dimensional t-SNE projection of the embeddings of a sample of generated profiles, with each point annotated according to the LLM used for profile generation. This visualization suggests that different encoders promote distinct structural organizations in the embedding space. For example, profiles encoded with `msmarco-bert-base-dot-v5` exhibit pronounced clustering with respect to the generating LLM.

To further examine the latent space structure, we conduct a local neighborhood analysis that mirrors the retrieval mechanism employed in our recommendation task. For each encoder, we randomly sample 500 profiles and retrieve their 11 nearest neighbors based on cosine similarity, corresponding to the fact that each user is associated with 12 profiles (3 LLMs × 4 time windows). Within each neighborhood, we compute the average consistency with respect to both user id and generating model. Since all profiles associated with a given user are derived from the same listening history, we expect well-structured embeddings to retrieve profiles belonging to the same user. Additionally, by measuring consistency with respect to the profile-generating LLM, we assess the extent to which profiles produced by the same model are encoded close to each other in the latent space.

[5]https://huggingface.co/spaces/mteb/leaderboard

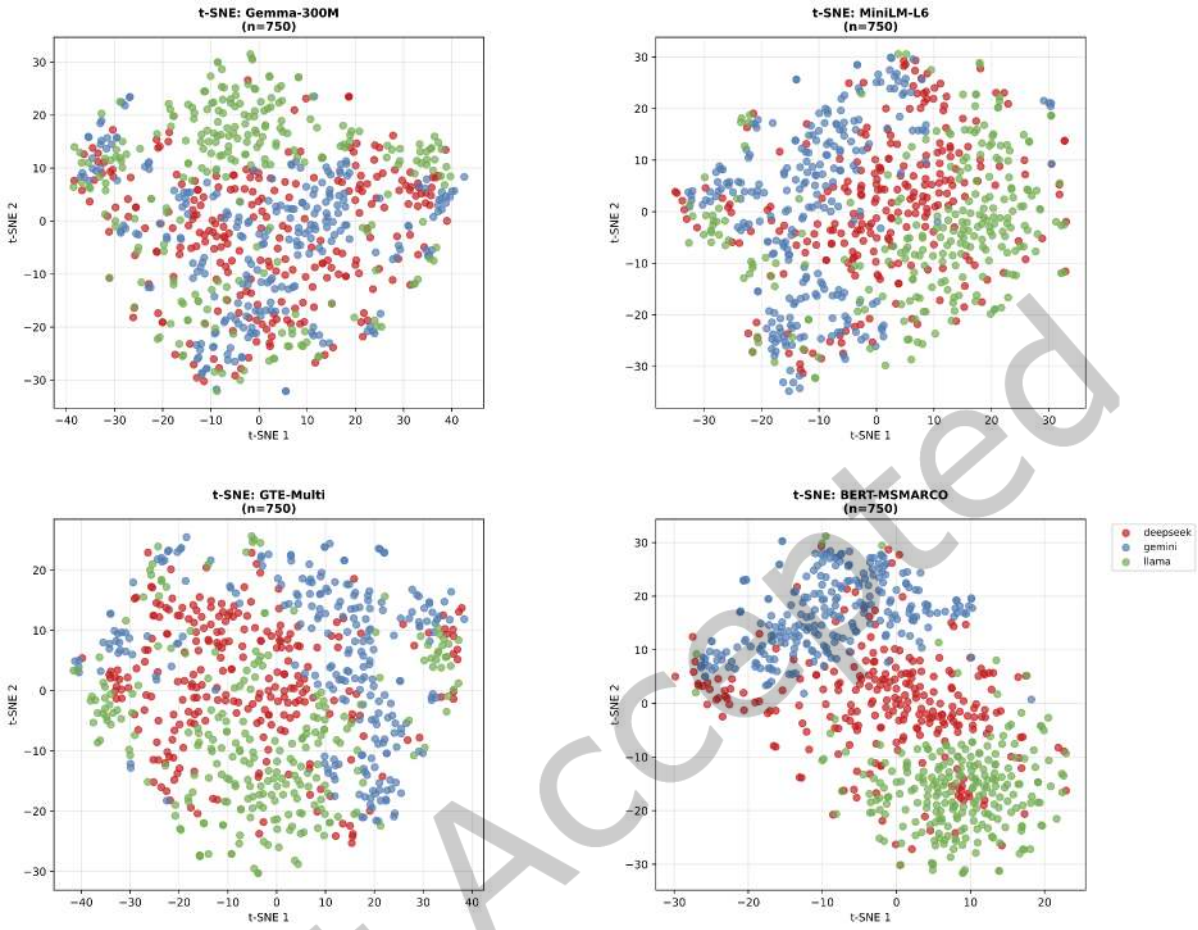


Fig. 8. t-SNE projection of the generated profiles encoded representations.

From Table 1, we observe that, for all encoders except `embeddinggemma-300m`, profiles generated with `Gemini` achieve the highest user-match values. This indicates that, in the embedding space, profiles belonging to the same user tend to lie closer to each other when generated by `Gemini`. At the same time, `Gemini`-generated profiles also exhibit consistently higher model-match values across encoders, meaning they are more likely to have nearest neighbors produced by the same LLM. This suggests stronger clustering by model, or the presence of description patterns that are more distinctive than those produced by the other LLMs. While such consistency may indicate more stable user representations, which is desirable, it may also reflect reduced stylistic diversity, which could limit the expressiveness of the resulting profiles. We recall that `Gemini` generated profiles tended to be shorter than the others, in number of characters: `Gemini`: 396.43 ± 70.62; `DeepSeek`: 543.00 ± 153.06; and `Llama`: 638.16 ± 117.54.

4.2.2 Item Representation. To provide a qualitative view of the information captured by the encoded item NL information, Figure 9 shows the 2D t-SNE projection of a sample of 10K NL descriptions of music tracks from our dataset. For clarity, we display only the ten most frequent genres, together with an "unknown" category for

Encoder	LLM	User Match	Model Match
embeddinggemma-300m	DeepSeek	25.2% ± 18.5%	46.8% ± 18.7%
embeddinggemma-300m	Gemini	26.5% ± 15.9%	83.5% ± 14.9%
embeddinggemma-300m	Llama	30.1% ± 21.2%	57.9% ± 24.9%
all-all-MiniLM-L6-v2-v2	DeepSeek	16.1% ± 16.5%	53.4% ± 22.2%
all-all-MiniLM-L6-v2-v2	Gemini	21.3% ± 14.2%	78.0% ± 16.9%
all-all-MiniLM-L6-v2-v2	Llama	16.1% ± 14.0%	68.1% ± 24.1%
gte-multilingual-base	DeepSeek	14.4% ± 14.3%	70.5% ± 20.7%
gte-multilingual-base	Gemini	23.5% ± 15.6%	80.0% ± 19.2%
gte-multilingual-base	Llama	21.9% ± 16.1%	65.4% ± 21.0%
msmarco-bert-base-dot-v5	DeepSeek	11.9% ± 12.7%	66.3% ± 25.0%
msmarco-bert-base-dot-v5	Gemini	17.0% ± 11.1%	91.9% ± 11.2%
msmarco-bert-base-dot-v5	Llama	10.8% ± 9.4%	76.5% ± 15.9%

Table 1. Profile characterization: local neighborhood analysis.

Encoder	Unknown	Rock	Rap	Electronic	Pop	Soul	World	Metal	Jazz	Chanson
embeddinggemma-300m	0.988	0.989	0.995	0.994	0.983	0.987	0.983	0.991	0.986	0.993
all-all-MiniLM-L6-v2-v2	0.970	0.905	0.929	0.949	0.811	0.902	0.824	0.935	0.945	0.838
gte-multilingual-base	0.984	0.969	0.972	0.978	0.926	0.974	0.938	0.964	0.970	0.974
msmarco-bert-base-dot-v5	0.994	0.954	0.963	0.980	0.909	0.938	0.951	0.959	0.970	0.941
Support	2007	3985	2653	2029	1742	1174	1069	1054	688	495

Table 2. Per-genre F1-scores of logistic regression classifiers trained on encoder embeddings.

tracks without genre information. Although the degree of clustering varies across models, all encoders exhibit some level of genre-structure preservation. We further quantify this property by using the resulting embeddings as input to a logistic regression classifier trained on 74k sampled representations to predict the genre labels (17 most popular) of 18k tracks. As reported in Table 2, all models achieve strong genre classification performance. This is expected, as the natural language descriptions explicitly include genre metadata when available. This experiment serves as a validation of the encoders' ability to preserve and prioritize key categorical signals within the latent space. In particular, `embeddinggemma-300m` achieves near-perfect performance (F1 > 0.98), indicating that its embeddings perfectly linearize these explicit textual attributes. While `gte-multilingual-base` and `msmarco-bert-base-dot-v5` perform similarly well, the slightly lower performance of `all-MiniLM-L6-v2` on broad categories like *World music* suggests a less robust handling of diverse linguistic contexts, even when the tokens are present.

The consistently high predictive performance for the "unknown" category across all encoders is noteworthy. In principle, models could leverage attributes such as artist names or track titles to infer missing genre information. However, our results suggest that, in practice, the representations tend to encode the presence of missing metadata itself, rather than reliably extrapolating the underlying genre.

To gain insight into which metadata attributes most strongly structure the latent space, we perform a local neighborhood analysis similar to the one above. For each encoder, we sample 1K random tracks (excluding those with missing attributes) and retrieve their 10 nearest neighbors based on cosine similarity. Within each

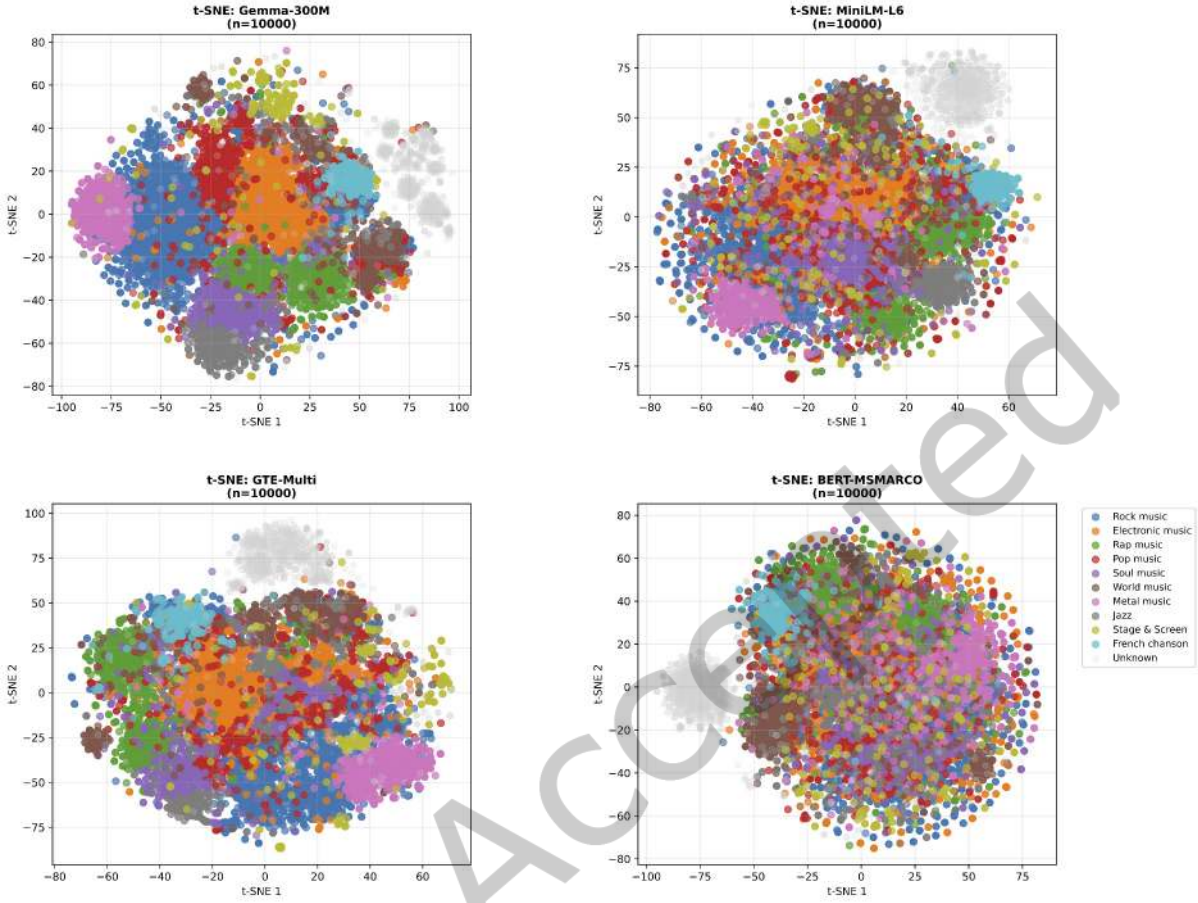


Fig. 9. t-SNE projection of item encoded representations.

neighborhood, we compute the average consistency (match rate) for genre, country of origin, and release year. These statistics provide an empirical measure of the extent to which similarity relationships are driven by specific metadata: higher match rates suggest that an attribute is a more prominent axis of organization within the representation.

Table 3 shows that all encoders exhibit substantial alignment with both genre and country of origin, with `embeddinggemma300m` and `gtemultilingualbase` achieving the highest consistency. This suggests that topological proximity in the latent space is strongly associated with these two categorical attributes. In contrast, release year is captured to varying degrees: while `gte-multilingual-base` shows substantial sensitivity (65.3%), `all-MiniLM-L6-v2` and `msmarco-bert-base-dot-v5` exhibit relatively weak sensitivity. This lower consistency for release year is expected, as year metadata is often noisy due to re-issues, compilations, or remastered versions. Finally, the relatively large standard deviations indicate that the embedding space is heterogeneous, with different regions emphasizing different metadata attributes. Taken together, these results show that the choice of encoder meaningfully affects the semantics encoded in item representations.

Encoder	Genre Match	Country Match	Year Match
embeddinggemma-300m	85.9% ± 20.7%	84.2% ± 23.9%	40.4% ± 26.2%
all-MiniLM-L6-v2	71.8% ± 29.1%	75.8% ± 29.4%	18.4% ± 19.2%
gte-multilingual-base	79.7% ± 22.6%	80.7% ± 27.4%	65.3% ± 27.7%
msmarco-bert-base-dot-v5	62.3% ± 31.7%	69.0% ± 31.2%	14.7% ± 17.1%

Table 3. Item characterization: local neighborhood analysis.

Encoder	Recall@10	NDCG@10
embeddinggemma-300m	6.61% ± 1.75%	6.7% ± 1.93%
all-MiniLM-L6-v2	5.01% ± 0.93%	4.97% ± 0.99%
gte-multilingual-base	3.92% ± 0.93%	4.03% ± 1.03%
msmarco-bert-base-dot-v5	2.85% ± 0.72%	3.04% ± 0.88%

Table 4. Overall recommendation performance per encoder, aggregating all LLMs used for profile generation, time windows and test sets.

4.3 Recommendation Task Performance

We now use the derived embeddings to predict the top-10 tracks sampled using the time-decay scheme described in Section 4.1, considering four different $t_{1/2}$: 4, 14, 30, and 64. Following [26], we evaluate performance using Recall@10 and NDCG@10, two standard recommendation metrics. Item relevance is computed as the cosine similarity between user profile and item embeddings. Figure 10 reports the mean performance and 95% confidence intervals for both metrics across half-life values, for each encoder and profile generation model. It is notable that the pattern is not the same for all encoders and LLMs, for instance, the recommendation performance for the pair `embeddinggemma-300m` and `Llama`, that achieve the highest score, follow a slight inverted-U shape, peaking at the test set sampled with 14 days as half-life. On the other hand, `embeddinggemma-300m` and `DeepSeek` show a slight upward trajectory, with performance increasing with the higher half-life value. Note that here we're mixing profiles that were sampled from different time windows, hereafter, we will solely focus on `embeddinggemma-300m` as the encoder, as it achieves the highest overall performance as seen in Table 4.

To investigate the effects of the test set's half-life, time window, and profile-generation model on Recall@10, we conducted a three-way ANOVA. The analysis revealed significant main effects of half-life ($F(3, 192) = 47.12$, $p = 7.24 \times 10^{-23}$), time window ($F(3, 192) = 37.98$, $p = 2.58 \times 10^{-19}$), and profile-generation LLM ($F(2, 192) = 1612.00$, $p = 9.54 \times 10^{-121}$). Among these, the LLM factor exhibited the largest effect, indicating that the choice of LLM and its profile structure is the dominant driver of Recall@10. In addition, all two-way interactions were significant: half-life × time window ($F(9, 192) = 20.55$, $p = 5.20 \times 10^{-24}$), half-life × profile-generation LLM ($F(6, 192) = 95.01$, $p = 8.96 \times 10^{-55}$), and time window × profile-generation LLM ($F(6, 192) = 51.01$, $p = 3.29 \times 10^{-37}$), as well as the three-way interaction half-life × time window × profile-generation LLM ($F(18, 192) = 5.22$, $p = 1.02 \times 10^{-9}$). Overall, these results indicate that the impact of temporal sampling on Recall@10 depends jointly on the evaluation setting and the profile-generation model, implying that no single time-window/half-life configuration is uniformly optimal across profile-generation LLMs.

Following [26], we investigate the relationship between recommendation performance and user self-recognition by fitting bootstrapped linear regression models that predict normalized recommendation metrics from user ratings. The resulting coefficients quantify the average change in recommendation performance associated with a one-unit increase in user self-identification. Figure 11 reports the mean and 95% confidence intervals

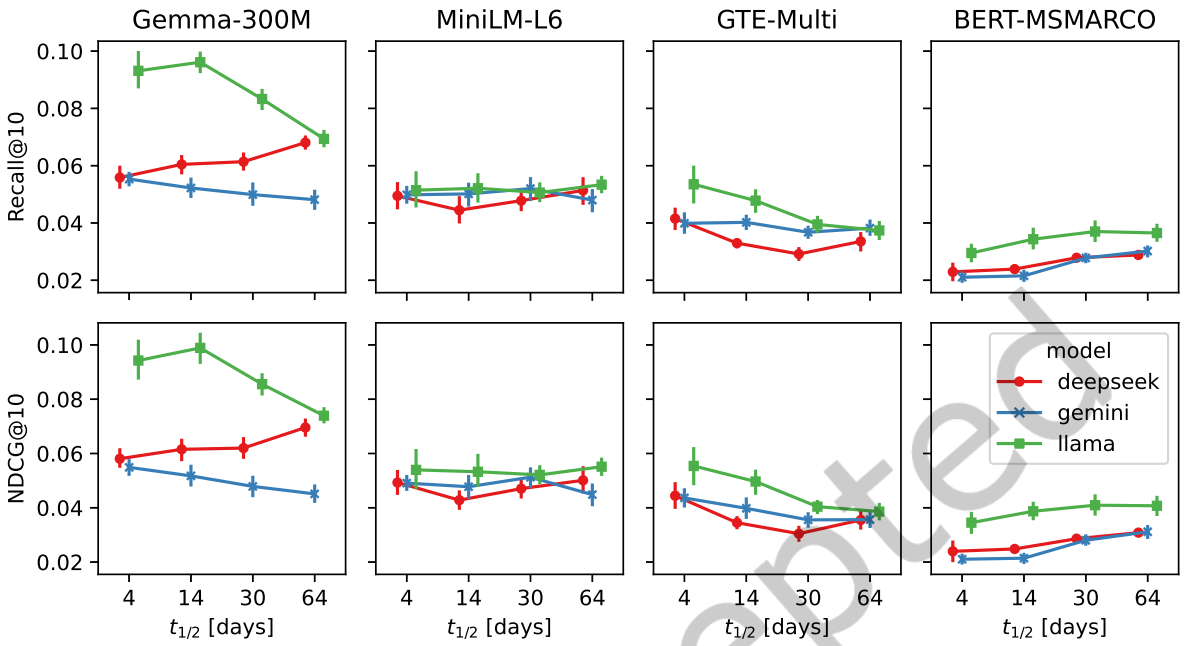


Fig. 10. Recall@10 and NDCG@10 over different test sets obtained by encoding profile and item metadata with different encoders.

obtained from 1K bootstrap resamples for each LLM used to build the profiles and test-set configuration. All metrics and LLMs show a weak positive linear relationship, indicating a misalignment between self-recognition and recommendation performance. Within each LLM, differences across test sets are generally small. With the exception of `DeepSeek` at $t_{1/2} = 30$, whose confidence interval crosses zero, most models exhibit a slight decreasing trend in the regression coefficient as $t_{1/2}$ increases, with the weakest association observed for $t_{1/2} = 64$. Although these effects are modest, they suggest that temporally informed item sampling plays a role in aligning recommendation performance with user perceptions.

The largest differences, however, arise across profile-generation models. In particular, profiles generated by `Llama` consistently exhibit stronger correlations between recommendation quality and users' self-identification ratings, indicating better alignment with subjective user evaluations.

5 Discussion and Conclusion

This paper assessed the quality of LLM-generated NL user profiles along two complementary dimensions: user ratings and downstream recommendation performance. By comparing profiles generated from items sampled over varying time windows of users' listening histories, we found that the length of the sampling window had only a limited impact on user ratings. As long as key preferences were captured, participants generally recognized themselves in the generated profiles. Nevertheless, one participant remarked that their profile was "missing quite a big part of my taste in music" and questioned whether the observation period was sufficiently long to reflect their interest in reggae.

In contrast, the largest variation in user ratings was attributable to the choice of LLM rather than to the underlying data window. This variation appears to reflect systematic stylistic differences in the generated profiles.

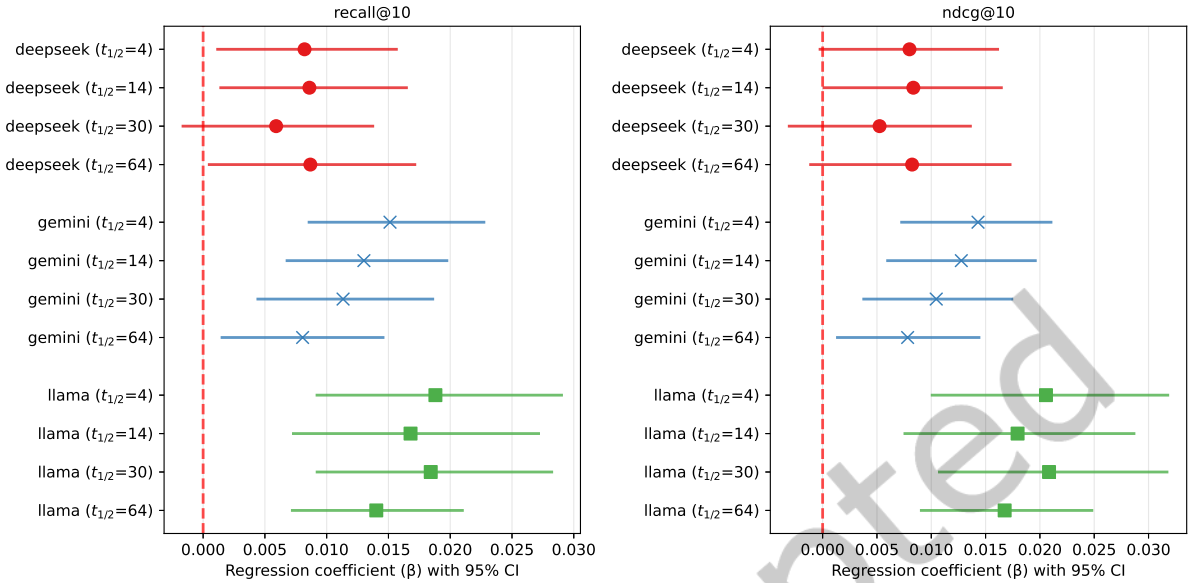


Fig. 11. Bootstrapped linear coefficients predicting normalized recommendation scores from ratings.

For example, `Gemini` tended to produce shorter and more high-level descriptions, whereas `Llama` frequently emphasized specific artists and track names, despite explicit prompt instructions. Some participants expressed a preference for this more concrete style, with one noting that they favored profiles "with the name of artists I listened to." `DeepSeek`, in turn, occasionally foregrounded technical metadata such as "remastered" tags in track titles, which several users perceived as less meaningful for representing their preferences. Overall, these observations suggest that user evaluations are influenced not only by the informational content of the profiles, but also by their stylistic and representational choices. Given the diversity of user expectations and the model-specific biases observed, a more systematic qualitative analysis would be required to identify which profile characteristics most strongly contribute to user satisfaction.

We further examined how user and item characteristics influenced profile quality. Profiles of specialist users, as measured by high GS-scores, tended to receive higher ratings, suggesting that more focused musical tastes may be easier to capture in NL form. Item characteristics, such as genre and country ratios, were also correlated with ratings, both positively and negatively. To account for potential confounding factors such as taste alignment, we employed a DR framework. The resulting average treatment effects indicate that some content types systematically influenced ratings: for example, rap content tended to reduce scores, whereas metal content increased them, within the limits of our sample size. While it remains unclear whether these effects stem primarily from metadata quality or from model behavior, the fact that different LLMs responded differently to identical metadata suggests the presence of 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 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 disproportionately disadvantaged. We note, however, that our study considered only three LLMs, and these conclusions may not generalize to all models. Moreover, our user study recruited participants exclusively from within a music streaming company, which may further limit the generalizability of our findings. Future research should investigate a larger, more diverse population. That said, as discussed

in Section 3.1, participants' musical engagement as measured by the Gold-MSI [19] was below the reported population norm, suggesting they do not interact with music more intensively than the general population despite their professional context.

In the downstream recommendation task, we evaluated the usability of the generated profiles for predicting items consumed after the user study. We encoded both user profiles and item metadata using four different encoders and analyzed the resulting embeddings. Neighborhood analyses revealed that item representations were largely governed by genre and country of origin and were particularly sensitive to missing metadata. This raises important concerns, as recommendations based on such embeddings may suffer from limited diversity, for example by favoring items from the same region or genre. Moreover, long-tail items may be systematically under-recommended if they lack complete metadata and end up projected together in the latent space. Mitigating these effects will require more carefully designed representation and recommendation strategies. Future work should investigate strategies such as prompt engineering with explicit masking tokens or imputation instructions to prevent encoders from clustering items based on data sparsity rather than semantic content. Another promising avenue is to leverage additional sources of information, such as audio signals or collaborative filtering, to enrich item representations and reduce dependence on textual metadata completeness.

Furthermore, user ratings and recommendation performance were only weakly correlated, revealing a misalignment between perceived representativeness and algorithmic utility. This finding has important implications for user trust: if users do not recognize themselves in profiles optimized for recommendation accuracy, their confidence in the system may be undermined. A plausible explanation lies in how users conceptualize their own musical taste, and whether it primarily reflects recent consumption patterns or a more stable, long-term construct. In this study, participants evaluated how well profiles captured their perceived musical taste, rather than how well they matched their current recommendation needs. While these notions are likely correlated, they are not equivalent, and for some users they may not substantially overlap. A key motivation for music recommendation is precisely to facilitate discovery of tracks and genres users have not previously encountered [18], and might not yet recognize themselves in. Moreover, stated preferences may be inconsistent with revealed preferences [1]. One possible direction to address this is to maintain two separate representations: one grounded in established musical taste, and one optimized for predictive ranking performance. In deployment, the relative weight of each could be adjusted per user and context, reflecting different trade-offs between familiarity and exploration. Future work could investigate how to expose and tune this trade-off in an interpretable way, and more carefully disentangle self-perception, recommendation goals, and system optimization.

References

[1] Moshe Ben-Akiva, Takayuki Morikawa, and Fumiaki Shiroishi. 1992. Analysis of the reliability of preference ranking data. *Journal of Business Research* (1992).

[2] 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.

[3] 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).

[4] 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.

[5] 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.

[6] 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.

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

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423

[9] 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).

[10] 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.

[11] 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.

[12] 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).

[13] 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).

[14] 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.

[15] 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.

[16] 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).

[17] 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.

[18] Marta Moscati, Darius Afchar, Markus Schedl, and Bruno Sguerra. 2025. Familiarizing with Music: Discovery Patterns for Different Music Discovery Needs. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*. 63–72.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

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

[26] 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*. 527–532.

[27] Bruno Sguerra, Viet-Anh Tran, and Romain Hennequin. 2023. Ex2vec: Characterizing users and items from the mere exposure effect. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 971–977.

[28] 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*.

[29] 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).

[30] 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.

[31] 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.

[32] Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. 2025. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354* (2025).
[33] 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).
[34] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MINILM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 485, 13 pages.
[35] 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).
[36] Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. 2024. mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 1393–1412.
[37] Joyce Zhou, Yijia Dai, and Thorsten Joachims. 2024. Language-Based User Profiles for Recommendation. *arXiv preprint arXiv:2402.15623* (2024).

Received 13 February 2026; revised 18 April 2026; accepted 21 April 2026