Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Personalization and recency in retrospective music recommendation

Samer Abaddi

eng.sameralabbadi@gmail.com

IDentity Research https://orcid.org/0000-0003-4612-7384

Research Article

Keywords: Music recommendation, Personalization, Recency weighting, Item-based collaborative filtering, Chronological evaluation, Reproducibility

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11743861/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Personalization and recency in retrospective music recommendation

Samer Abbadi

Research Department, IDentity Research, Amman, Jordan

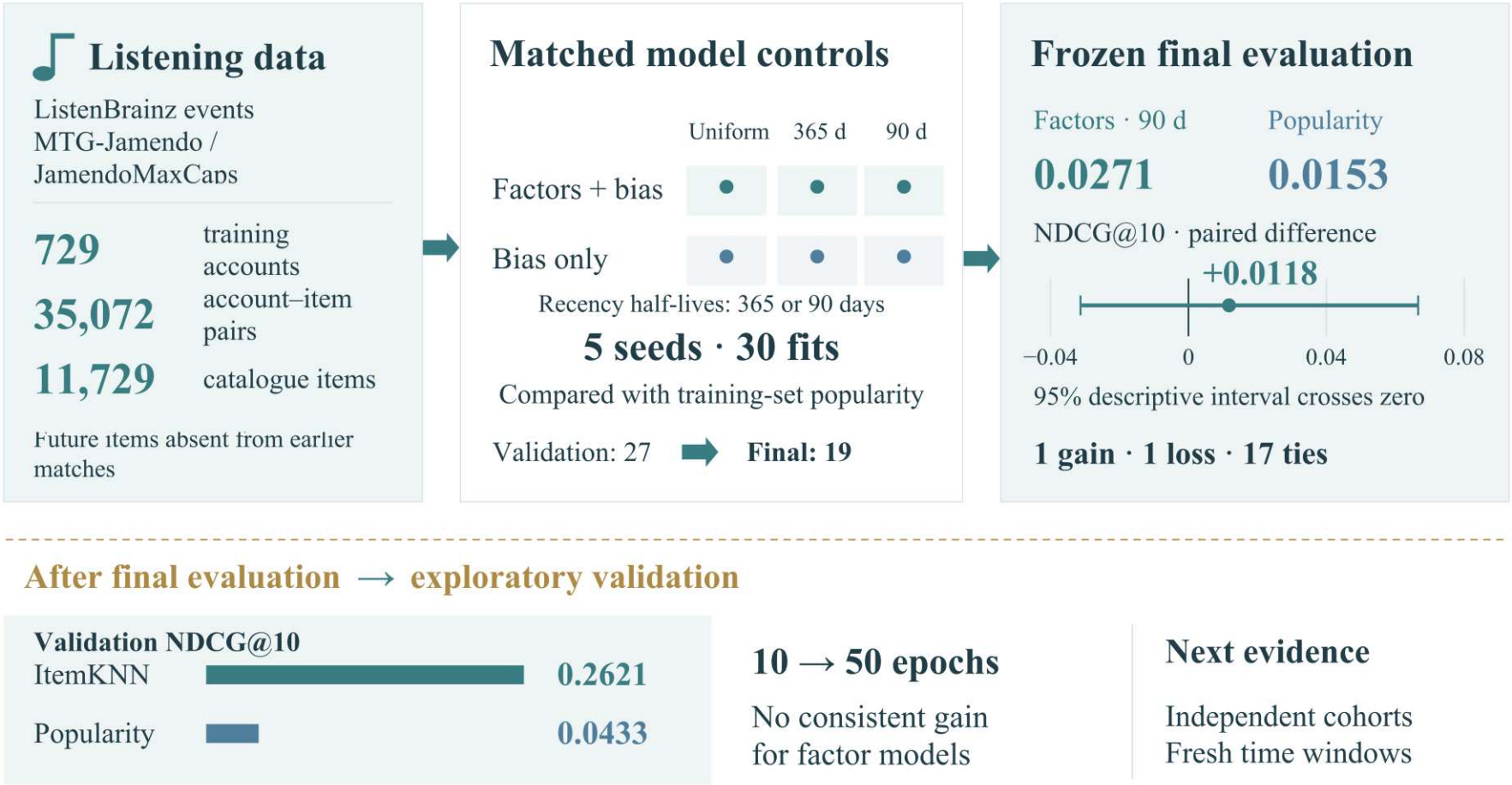
ORCID: 0000-0003-4612-7384

Correspondence: eng.sameralabbadi@gmail.com

Abstract

A personalized music ranking need not contain a strong personal signal. We examined the contributions of account-specific factors and recency to retrospective recommendation of catalogue items absent from earlier matched histories. The training data comprised 35,072 account–item pairs from 729 accounts and 11,729 items. Two model families, three weighting schemes and five seeds produced 30 fits, compared with popularity under chronological validation and final evaluation involving 27 and 19 accounts, respectively. Rankings covered the full eligible catalogue; targets outside the training catalogue remained in metric denominators. Popularity was selected in validation. The personalized challenger achieved final NDCG@10 of 0.0271 versus 0.0153 for popularity, a paired difference of 0.0118 with a descriptive 95% account-bootstrap interval of [-0.0313, 0.0666]. One account benefited, one lost and 17 tied; factor and bias-only recommendation sets were closely aligned. Recency contrasts did not establish a factor-specific benefit. A later exploratory validation analysis found substantially higher NDCG@10 for ItemKNN (0.2621) than popularity (0.0433), while longer training did not consistently improve factor models. These results do not establish a reliable personalization advantage in this small final cohort. The study contributes an auditable comparison that distinguishes account-specific effects, temporal emphasis and catalogue constraints, and identifies item neighbourhoods as a priority for independent chronological evaluation.

Keywords: Music recommendation; Personalization; Recency weighting; Item-based collaborative filtering; Chronological evaluation; Reproducibility.



Graphical abstract

1 Introduction

Music discovery depends as much on what a recommender makes visible as on what a catalogue contains (Schedl et al., 2018). This distinction matters in an industry increasingly organized around streaming. Global recorded music revenue reached US$31.7 billion in 2025, streaming represented 69.6% of that revenue, and paid streaming subscription accounts served 837 million users (International Federation of the Phonographic Industry [IFPI], 2026). At this scale, evidence about ranking decisions matters to listeners seeking relevant music, artists seeking exposure and services choosing how to organize their catalogues.

Personalization connects listeners with music inferred from their interests, although recorded interactions can reflect different listening needs: familiar tracks may accompany routines, unfamiliar tracks may be encountered incidentally. Music recommendation consequently involves contextual needs and discovery goals that extend beyond predicting an observed interaction (Schedl et al., 2018). A recent review of 55 content-driven recommendation studies identifies continuing challenges involving novelty, explanation, context, sequencing, efficiency and cold start (Deldjoo et al., 2024). Together, these challenges connect the choice of recommendation task to the listening experience it is intended to support.

A central methodological question is whether an apparently personalized ranking adds information beyond the popularity already present in the data. Popular items can dominate both observed histories and recommended lists, while the consequences of that concentration depend on the setting and evaluation method (Klimashevskaia et al., 2024). In an influential reproducibility analysis, conceptually simple methods could outperform 11 of 12 reproducible neural approaches under the examined conditions (Ferrari Dacrema et al., 2021). In a separate environmental prediction study, foundation embeddings improved regional transfer accuracy but worsened temporal prediction (Abaddi, 2026a). These contrasting findings reinforce the need to test added model information against appropriate controls.

Ji et al. (2023) showed that ignoring the global interaction timeline can change recommendation accuracy and the relative ordering of algorithms when training includes information unavailable at the intended prediction date. Chronological separation therefore matters, but so does the set of items a model can actually recommend. A later observed item outside the training catalogue cannot receive a recommendation from a model restricted to that catalogue. Removing such targets from evaluation changes the question being answered and can obscure an important practical constraint.

These concerns define a focused empirical gap for studies built from archived, catalogue-matched listening records: how much of the observed ranking performance can be attributed to account-specific information, how much to recency or shared item preference, and how much is constrained by the construction of the evaluation data? A recent survey of recommender-system reproducibility research finds improved documentation alongside continued dependence on a narrow range of datasets and evaluation practices (Said & Bellogín, 2026). Transparent experiments on less standardized records can therefore be useful when their measurement boundaries remain explicit.

The present study examines retrospective recommendation of catalogue items absent from an account's earlier matched history. It uses 35,072 distinct training account–item pairs from 729 accounts and a fixed catalogue of 11,729 items, with 27 accounts contributing validation targets and 19 contributing final-test targets. Novelty is defined relative to recorded assignments. This task captures whether a model can recover later items beyond those already associated with the account, while keeping the available recommendation catalogue fixed. The unequal training and evaluation cohorts also motivate reporting coverage alongside ranking accuracy.

The study aims to evaluate whether an account-balanced personalized pairwise-ranking model improves future-item ranking over training-account popularity, and to examine the contributions of personalization and recency through matched item-bias controls and repeated

random seeds. It also assesses catalogue coverage, recommendation concentration and sensitivity to individual accounts, and compares item-neighbour methods, globally time-decayed popularity and fixed training durations in exploratory validation analyses. Normalized discounted cumulative gain and recall summarize ranking performance, while account-level comparisons and seed variation describe different sources of uncertainty. Exploratory comparisons are reported separately from the frozen final evaluation.

The contribution is an auditable empirical assessment of when these modelling choices provide useful ranking information within a bounded music catalogue. Scientifically, the controlled comparisons help distinguish account-specific effects from shared item preference and temporal weighting. Methodologically, explicit target accounting makes unavailable items and the small effective evaluation sample visible, while preserved configurations, checkpoints and outputs connect reported claims to executed analyses. Practically, comparison with inexpensive baselines helps researchers judge whether additional training effort is justified for similar data.

The manuscript proceeds through related work, methodology, results, discussion, implications and limitations, ending with conclusions and priorities for further research.

2 Literature review and study framework

2.1 Listening evidence and temporal context

Deldjoo et al. (2024) distinguish content-driven recommendation, which uses acoustic or semantic descriptions, from approaches built around listening behaviour. Their review identifies unresolved questions about novelty, context and efficient use of content. These themes connect music representation to the kind of recommendation being made: relistening draws on familiar preferences, whereas recommending an item absent from recorded history requires extending beyond those associations. For the latter task, the relevant question is whether the available listening evidence supports useful rankings within the catalogue.

Kowald et al. (2024) model music preference through frequency, recency and semantic context within a human-memory framework. Their relistening setting provides a rationale for temporal signals and suggests that the age of an interaction can affect its relevance. Two distinct mechanisms follow: emphasizing recent listening within an account and favouring items with recent activity across accounts. Comparing these mechanisms helps distinguish an individual temporal signal from a shared trend. Their combination with account-specific factors also tests whether recent behaviour carries information beyond item popularity.

2.2 Baselines and evaluation boundaries

Ferrari Dacrema et al. (2021) demonstrate how competitive conventional methods can overturn apparent neural improvements, motivating explicit controls for shared item preference. In music, Yambda combines chronological benchmarking, organic-interaction indicators and audio embeddings with comparisons spanning ItemKNN and sequential models (Ploshkin et al., 2025). Its richer feedback illustrates how data design expands the questions a benchmark can address. Item neighbourhoods offer a complementary explanation of recommendation performance: items may be useful because they occur in related listening histories, even without a learned account representation. A road-image study found no established superiority of regional failure prediction over simpler presence cues under a fixed

review budget, offering a methodological parallel for testing incremental information (Abaddi, 2026b).

Gusak et al. (2025) show that splitting and target-selection strategies affect sequential-recommender rankings. Their findings make the evaluation population and target definition part of the scientific question. Retaining future targets outside the training catalogue reveals the opportunity available to a restricted recommender. Popularity concentration deserves separate attention because aggregate accuracy can coexist with narrow exposure (Klimashevskaia et al., 2024). Table 1 relates these findings to the present study. Said and Bellogín (2026) identify continuing limitations in dataset and evaluation diversity, supporting analyses that make their comparison settings and evidence accessible.

Table 1. Selected literature and the present study's focus

Study	Relevant evidence	Present study focus
(Kowald et al., 2024)	Memory-based music relistening	Items absent from prior matched history
(Ploshkin et al., 2025)	Temporal music benchmark; richer feedback	Controlled analysis of saved catalogue assignments
(Gusak et al., 2025)	Rankings depend on splitting and targets	Fixed chronology and all-target accounting
(Klimashevskaia et al., 2024)	Popularity bias requires contextual interpretation	Concentration alongside accuracy

2.3 Research gap and study framework

The research gap addressed here concerns the relative contributions of account-specific affinity, shared item preference and temporal weighting in archived catalogue-matched listening records. Recent literature develops these components in different settings, but their combined behaviour depends on the available histories, candidate catalogue and future targets. A controlled comparison can establish where these signals agree, where they produce different rankings and which accounts drive the aggregate result. This focus connects model

behaviour to the information present in the records without requiring a new recommendation architecture.

Fig 1 presents a preference–opportunity framework linking listening evidence to ranking signals and recommendation outcomes. Listening histories supply associations, temporal context distinguishes recent from earlier activity, and catalogue boundaries determine the available opportunity. Personalization, popularity, recency and item neighbourhoods provide competing explanations of ranking performance. Three outcome dimensions connect them: recovery of future items, concentration of recommended exposure and coverage of eligible targets. Examining these dimensions together clarifies whether a gain reflects more useful account information, shared trends or a favourable candidate set. The framework guides the study questions; the methodology specifies their implementation.

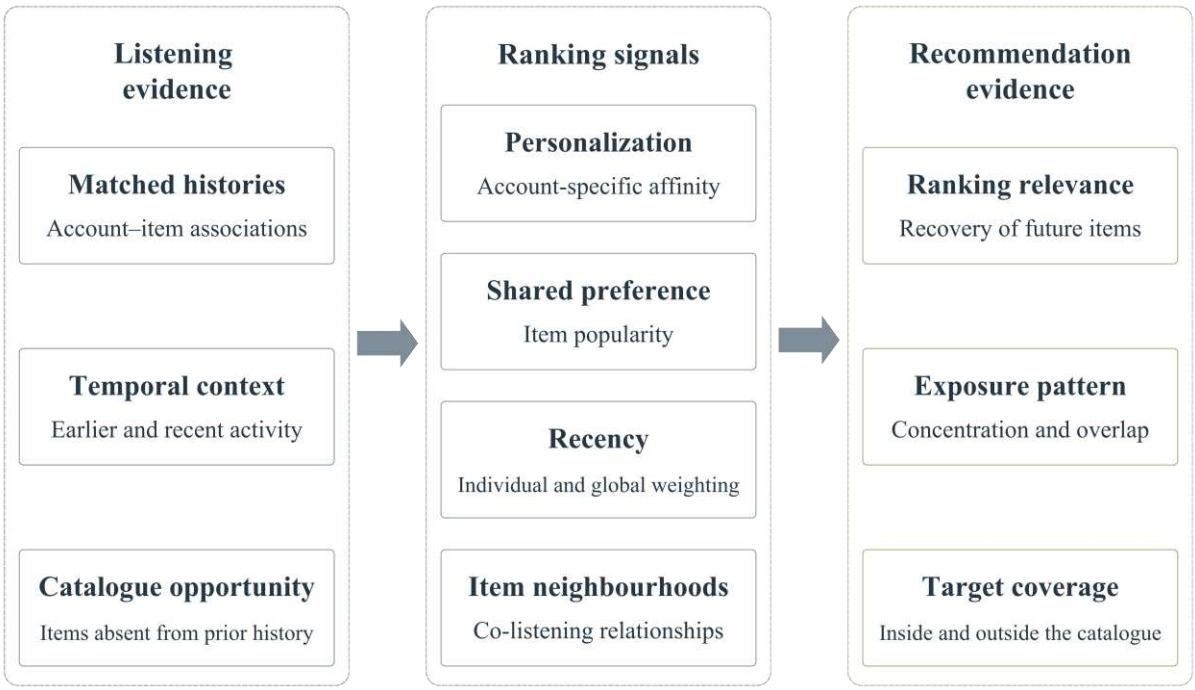


Fig 1. Preference–opportunity framework for retrospective music recommendation

3 Methodology

3.1 Study design and data sources

We conducted a retrospective, account-level recommendation study to distinguish personalization from shared item preference and recent listening. The endpoint was a future catalogue item absent from an account's earlier matched history. Archived ListenBrainz records supplied the listening events (MetaBrainz Foundation, n.d.); the saved catalogue combined MTG-Jamendo metadata (Bogdanov et al., 2019) with original metadata distributed through JamendoMaxCaps (Roy et al., 2025). Models used account–item assignments and timestamps, without acoustic features or generated captions.

The compact source contained 347,287 event keys, 6,316 account hashes and 18,305 catalogue identifiers. Each record retained its timestamp, assigned item, normalized song key, matching-evidence tier and eligibility flag. Frozen exclusion registers covered unresolved track, artist, version and event cases. Analyses concern account-level catalogue assignments. Fig 2 links the data sources, cohort construction and analysis stages.

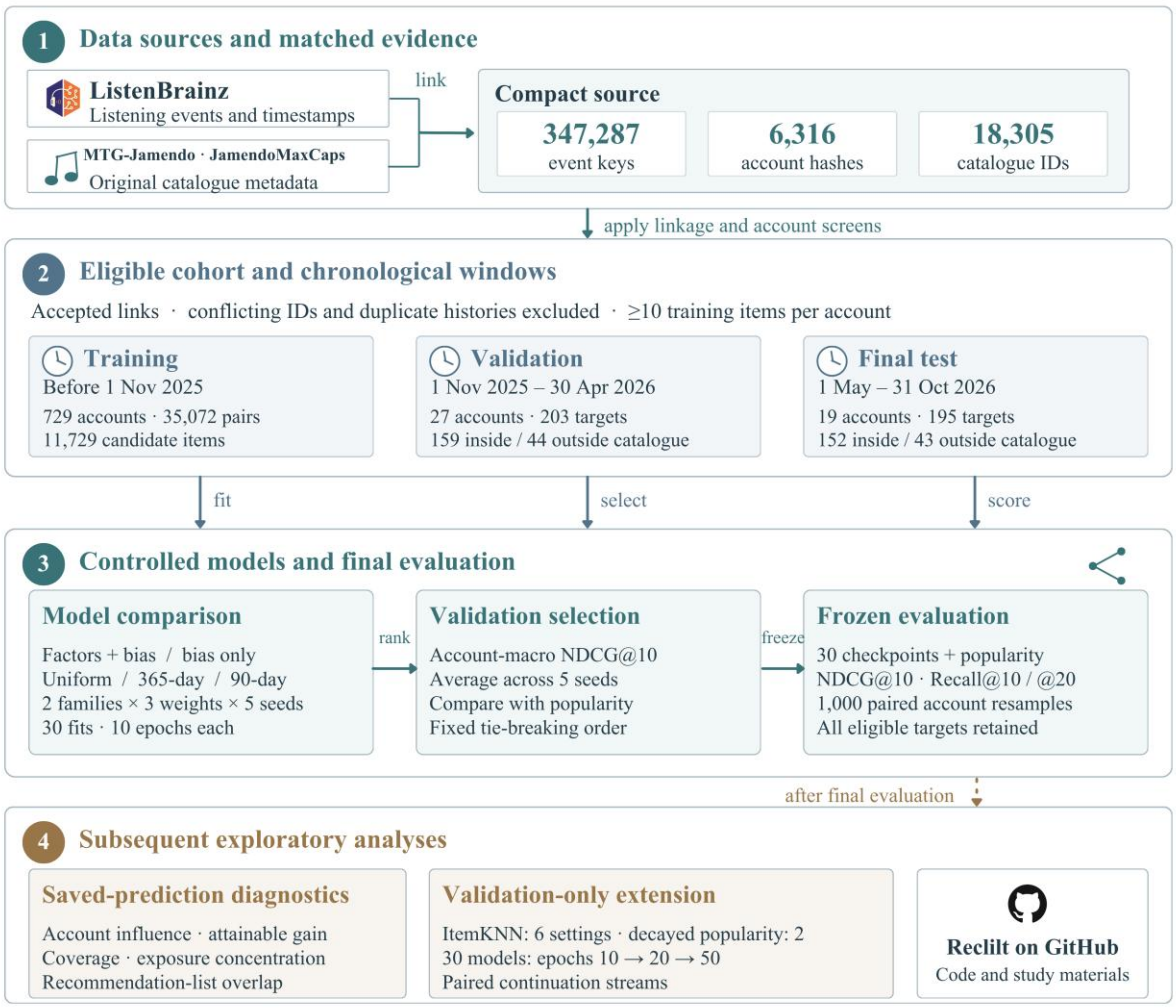


Fig 2. Data sources, cohort construction and chronological analysis architecture. Validation and test accounts are subsets of the training cohort; the exploratory analyses follow final evaluation. ListenBrainz logo © MetaBrainz Foundation, CC BY-SA 4.0; GitHub mark © GitHub.

3.2 Eligibility and chronological partitioning

Eligible records required the core-eligibility flag plus one of three links: source-identifier consistency, prior provisional recording-identifier consistency, or an exact artist–title pair supported by duration or album evidence. All competing assignments at the same account–timestamp and all identifiers in ambiguous normalized-song groups were quarantined. Known account exclusions were retained, and every member of an identical accepted pre-cutoff history group was excluded. Accounts needed at least 10 distinct accepted training items; eligibility did not require future activity.

Training ended immediately before 1 November 2025 UTC. Validation covered 1 November 2025–30 April 2026; final evaluation covered 1 May–31 October 2026. Global cutoffs constrain temporal leakage (Ji et al., 2023). Model parameters and account factors remained fitted to training data; only exclusion masks advanced to the evaluation cutoff. Every earlier assigned item, including weak or held assignments, entered the mask. Within each future window, repeated assignments became one binary account–item target.

The fixed candidate universe comprised 11,729 training items, including 10,571 from MTG-Jamendo and 1,158 from JamendoMaxCaps original metadata. Rankings used the entire eligible catalogue after masking, without sampled evaluation negatives. Otherwise eligible targets outside that catalogue remained in metric denominators. Table 2 defines the training and evaluation populations. Accounts without targets had undefined ranking metrics; accounts with exclusively out-of-catalogue targets remained evaluable with zero attainable gain.

Table 2. Training and evaluation populations

Quantity	Training	Validation	Final test
Accounts contributing to each stage	729	27	19
Accepted training event keys	160,250	—	—
Distinct account–item pairs / targets	35,072	203	195
Targets inside the fixed catalogue	—	159	152
Targets outside the fixed catalogue	—	44	43
Training accounts without eligible future targets	—	702	710

3.3 Ranking models and controlled comparisons

The personalized model combined item bias with a 16-dimensional account–item interaction in Eq. (1), following the pairwise formulation of Rendle et al. (2009):

$$s_{ui} = b_i + {p_u}^T q_i \quad (1)$$

Here, b is an item bias and p and q are account and item vectors. An item-bias-only model removed the interaction and was retrained under the same objective. Unique-account

popularity provided a deterministic comparator, consistent with the need for competitive conventional baselines (Ferrari Dacrema et al., 2021). Each positive training item was compared with 4 distinct unobserved catalogue items per epoch, excluding all prior assignments. These sampled references were not interpreted as dislikes.

For account u, pairwise comparisons P were optimized using the normalized weighted objective in Eq. (2):

$$L_u = \frac{\sum_{(i,j)\in P_u} w_{ui}\left[ln\left(1+e^{-\left(s_{ui}-s_{uj}\right)}\right)+\frac{\lambda R_{uij}}{2}\right]}{\sum_{(i,j)\in P_u} w_{ui}} \quad (2)$$

R sums the squared norms of the account vector, both item vectors and their biases; the bias-only control retains only bias terms. Gradients were accumulated across an account's comparisons before one update per epoch. Uniform weights and the exponential weights in Eq. (3), with half-lives of 90 or 365 days, separated personalization from temporal emphasis:

$$w_{ui} = \frac{|I_u| 2^{\frac{-(T_0-t_{ui})}{h}}}{\sum_{\ell\in I_u} 2^{\frac{-(T_0-t_{u\ell})}{h}}} \quad (3)$$

I is the account's distinct accepted training-item set, t is the latest accepted timestamp for that pair, and T_0 is the training cutoff, with elapsed time expressed in days. Normalization gives mean-one weights within each account; the same weights apply to loss and regularization. Learning rate was 0.1 and regularization strength 0.001. Factors began as zero-mean Gaussian values with standard deviation 0.05; biases began at zero.

The developmental comparison crossed 2 model families with 3 weighting schemes and seeds 11, 23, 37, 53 and 71, producing 30 ten-epoch fits. One completed fit was reused; 29 were new. New paired runs shared seed-controlled initialization, account ordering and reference draws. Paired contrasts assessed the factor contribution, recency effects and their difference-in-differences interaction. Video 1 documents the comparison. Settings were fixed for the batch; the study was not preregistered.

3.4 Selection, final evaluation and uncertainty

Selection used mean validation account-macro normalized discounted cumulative gain at rank 10 (NDCG@10), averaged over the five seeds and compared with popularity. A fixed simpler-model ordering resolved numerical ties. For target set T_u and ranked item i_{ur} at position r, binary-relevance NDCG and recall were computed using Eqs. (4) and (5), respectively (Järvelin & Kekäläinen, 2002):

$$NDCG_u@k = \frac{\sum_{r=1}^{k} \frac{\mathbb{1}(i_{ur} \in T_u)}{log_2(r+1)}}{\sum_{r=1}^{min(k,|T_u|)} \frac{1}{log_2(r+1)}} \quad (4)$$

$$Recall_u@k = \frac{\sum_{r=1}^{k} \mathbb{1}(i_{ur} \in T_u)}{|T_u|} \quad (5)$$

Membership indicators identify target hits. The ideal-gain denominator includes all eligible targets, including unavailable catalogue items. NDCG@10 was primary; Recall@10 and Recall@20 were secondary. Score ties were resolved by ascending catalogue identifier. The frozen final stage evaluated the 30 existing checkpoints and popularity without refitting or reselection. The primary estimate concerned the overall validation-selected method; the primary contrast compared the selected personalized configuration with popularity. Aggregate split counts were available before final scoring.

Each account's metric was first averaged across seeds, then equally across evaluable accounts. Paired account bootstrap intervals used 1,000 resamples, seed 20261114 and the empirical 2.5th and 97.5th percentiles. Seed standard deviations were reported separately. Intervals describe variation across the observed accounts conditional on the fitted seeds; they are not adjusted for selection or multiple contrasts.

3.5 Subsequent exploratory analyses

Post-test diagnostics used saved predictions to examine account-omission influence, attainable gain, catalogue coverage, list overlap and exposure concentration. Omission did

not refit models or redefine the cohort. A separate extension examined item neighbours, global recency and training duration using validation data only.

ItemKNN used binary co-listening cosine similarity multiplied by $\frac{n}{n+\alpha}$, where n is the number of shared training accounts. Each candidate retained its strongest 20, 50 or 100 neighbours, with shrinkage α equal to 0 or 10; scores summed similarities to items in the account's training history. Global popularity summed exponentially decayed last-pair timestamps using 90- or 365-day half-lives, without within-account normalization. These settings supplied 8 baseline configurations.

All 30 models continued from epoch 10 to 50, with validation scoring at epochs 20 and 50 and unchanged architecture and optimization settings. Continuation used fresh random streams seeded with 770000 plus the original seed, paired across configurations. Video 2 documents this validation-only extension.

3.6 Implementation and research materials

Local runs used Python 3.13.7 standard-library implementations on an Apple M3 Pro MacBook Pro with 18 GiB memory and macOS 15.6, one CPU worker and a monitored 2 GiB memory ceiling. Complete-epoch checkpoints, configurations, input hashes, predictions and logs were retained, with verified external-storage copies.

The private Reclilt repository, commit be929083a95a, preserves code, pseudonymized inputs, configurations, checkpoints and outputs (Video 3). SHA-256 manifests bind these materials; reproduction requires documented path rebinding.

4 Results

4.1 Evaluation coverage and model selection

The training cohort comprised 729 accounts and 35,072 distinct account–item pairs. Evaluation was possible for 27 accounts in validation and 19 in the final window (Table 2). Of 203 validation targets, 159 were represented in the training catalogue; final evaluation contained 195 targets, of which 152 were represented. Thus, final-window catalogue coverage was 77.95% at the pooled target level. The account whose targets were entirely unavailable remained in the final metric denominator with zero gain; the other 710 training accounts had no eligible final-window targets and were not assigned ranking scores.

Popularity achieved the highest development NDCG@10, 0.043350, and was selected by the fixed rule (Table 3). Learned configurations ranged from 0.032721 to 0.039362. Among personalized models, the 90-day configuration ranked highest at 0.037574 and became the prespecified challenger for the final comparison. Its seed standard deviation was 0.000110, whereas the uniform factor model varied more across seeds, with a standard deviation of 0.005864. Recent weighting therefore improved the personalized model's development ranking, but no learned configuration displaced popularity. Recall@20 differed little across these configurations, ranging from 0.100490 to 0.104605.

Table 3. Development performance across 27 validation accounts

Configuration	NDCG@10	Seed SD	Recall@10	Recall@20
Popularity	0.043350	—	0.058985	0.104252
Bias only uniform	0.039145	0.000000	0.054870	0.104252
Bias only 365 days	0.039362	0.005864	0.072977	0.104605
Bias only 90 days	0.037593	0.000000	0.050754	0.100490
Factors uniform	0.032721	0.005864	0.032647	0.104252
Factors 365 days	0.035580	0.004702	0.058232	0.104605
Factors 90 days	0.037574	0.000110	0.050754	0.100490

* Five-seed means; SD is seed variation; — indicates deterministic popularity.

4.2 Final evaluation and controlled effects

In the frozen final evaluation, the validation-selected popularity baseline attained NDCG@10 of 0.015326. The selected personalized challenger attained 0.027101, with Recall@10 of 0.059649 and Recall@20 of 0.079031 (Table 4). Its paired NDCG@10 difference from popularity was 0.011775, with a descriptive 95% account-bootstrap interval from -0.031292 to 0.066618. The interval spans both deterioration and improvement; these observations do not establish a reliable personalized-model advantage.

Table 4. Frozen final performance across 19 evaluation accounts

Configuration	NDCG@10	Seed SD	Recall@10	Recall@20
Popularity	0.015326	—	0.023392	0.042607
Bias only uniform	0.007875	0.000165	0.011696	0.054637
Bias only 365 days	0.027661	0.000000	0.064327	0.076023
Bias only 90 days	0.026791	0.000000	0.058480	0.079031
Factors uniform	0.008474	0.001652	0.012865	0.054637
Factors 365 days	0.027661	0.000000	0.064327	0.079031
Factors 90 days	0.027101	0.002025	0.059649	0.079031

* Five-seed means; SD is seed variation; — indicates deterministic popularity.

The controlled comparisons further limited attribution to personalization. Under 365-day weighting, factor and bias-only models had the same mean NDCG@10, 0.027661, and Recall@10, 0.064327. Under 90-day weighting, adding factors changed mean NDCG@10 by only 0.000310, with a descriptive interval from -0.001384 to 0.002315. Uniform weighting produced a difference of 0.000600, with an interval from 0.000000 to 0.001799. Recent weighting yielded higher final scores for both model families, while the incremental factor contribution remained small. The numerically higher score of the 365-day configuration did not trigger reselection.

4.3 Account influence and recommendation concentration

The primary NDCG@10 difference arose from two accounts: one favored the personalized challenger, one favored popularity, and the remaining 17 had zero gain under both methods. Removing the benefiting account from the descriptive calculation changed the mean difference to -0.011010; removing the disadvantaged account changed it to 0.023440 (Fig 3A). All 19 accounts remained in the primary analysis. The direction of the aggregate contrast was therefore sensitive to a single account, despite modest variation across fitted seeds.

Unavailable targets imposed a ceiling, but did not account for the full gap to an ideal ranking. With the original target denominators and candidate masks, the attainable account-macro ceilings were 0.904842 for NDCG@10, 0.816667 for Recall@10 and 0.847368 for Recall@20. These are analytical upper bounds, not model results. Restricting the descriptive calculation to covered targets in 18 accounts gave NDCG@10 of 0.028607 for the challenger and 0.016178 for popularity; this alternative denominator did not materially change their low absolute scores.

Recommendation sets overlapped strongly between factor and bias-only models (Fig 3B). Across matched seeds and accounts, mean top-10 set Jaccard similarity was 0.954 under uniform weighting, 0.989 under 365-day weighting and 0.979 under 90-day weighting. Exact ordered-list agreement was lower: 5.3%, 63.2% and 65.3%, respectively. For the 90-day factor model, each seed's 190 recommendation slots contained 20 distinct items on average, compared with 22 for popularity. The training catalogue's 100 most popular items occupied 88.5% of its slots, and its ten most frequently recommended items occupied 81.4%. The lists thus remained concentrated even when their ordering changed.

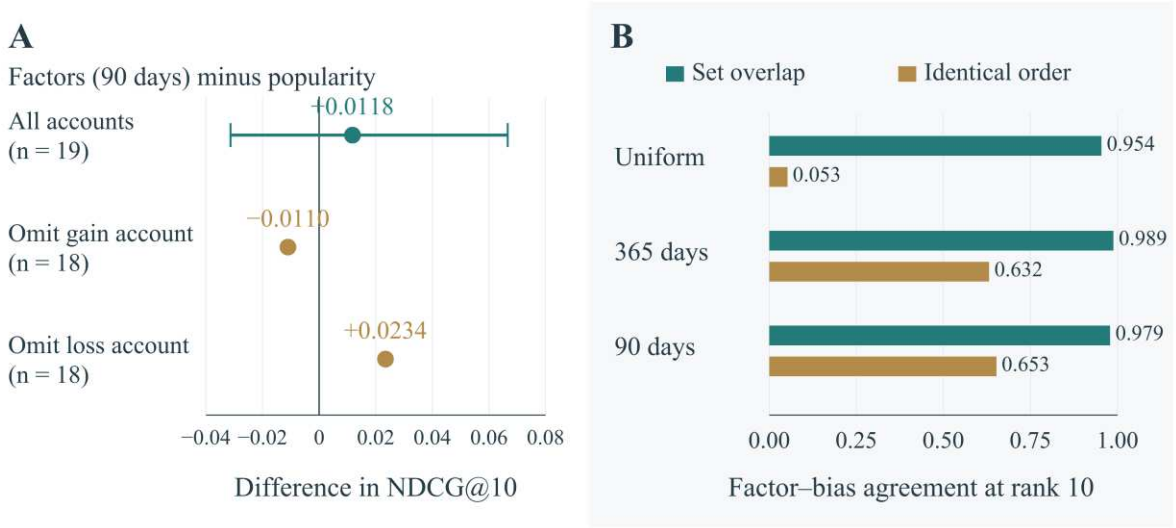


Fig 3. Final-evaluation sensitivity and list agreement. (A) The 90-day factor model minus popularity, with the primary 95% descriptive interval and account-omission means. (B) Top-10 set overlap and exact ordered-list agreement between factor and bias-only models, averaged across accounts and five seeds.

4.4 Exploratory baselines and training duration

The subsequent validation-only extension identified a stronger conventional baseline. All six ItemKNN settings exceeded popularity in validation NDCG@10, ranging from 0.193160 to 0.262058 (Fig 4A). The highest value used 100 neighbours and shrinkage of 10, with Recall@10 of 0.330952 and Recall@20 of 0.459926. Its NDCG@10 difference from popularity was 0.218708, with a descriptive interval from 0.107359 to 0.334729. This comparison was selected within the exploratory validation exercise; the interval is not adjusted for that selection and provides no new final-test estimate. Global time-decayed popularity performed less well, scoring 0.030268 with a 365-day half-life and 0.011368 with a 90-day half-life.

All 30 learned runs still had decreasing sampled training objectives at epoch 10, motivating the duration comparison. Continued optimization did not produce a consistent validation improvement (Fig 4B). From epochs 10 to 50, factor-model NDCG@10 changed from 0.032721 to 0.032763 under uniform weighting, from 0.035580 to 0.037352 under 365-day weighting, and from 0.037574 to 0.036950 under 90-day weighting. The highest

learned score across the evaluated durations was 0.041532 for the 365-day bias-only model at epochs 20 and 50, still below popularity. Supplementary Table S1 reports all 27 configuration–duration groups, including recall and seed variability. The original selection and completed final evaluation remained unchanged.

On the local single-worker setup, the 29 new development fits took 302.1 s and peaked at 222.5 MiB. Final scoring took 5.0 s and 134.1 MiB. The extension completed 30 continuations and 68 new validation scorings in 16.32 min, peaking at 346.2 MiB.

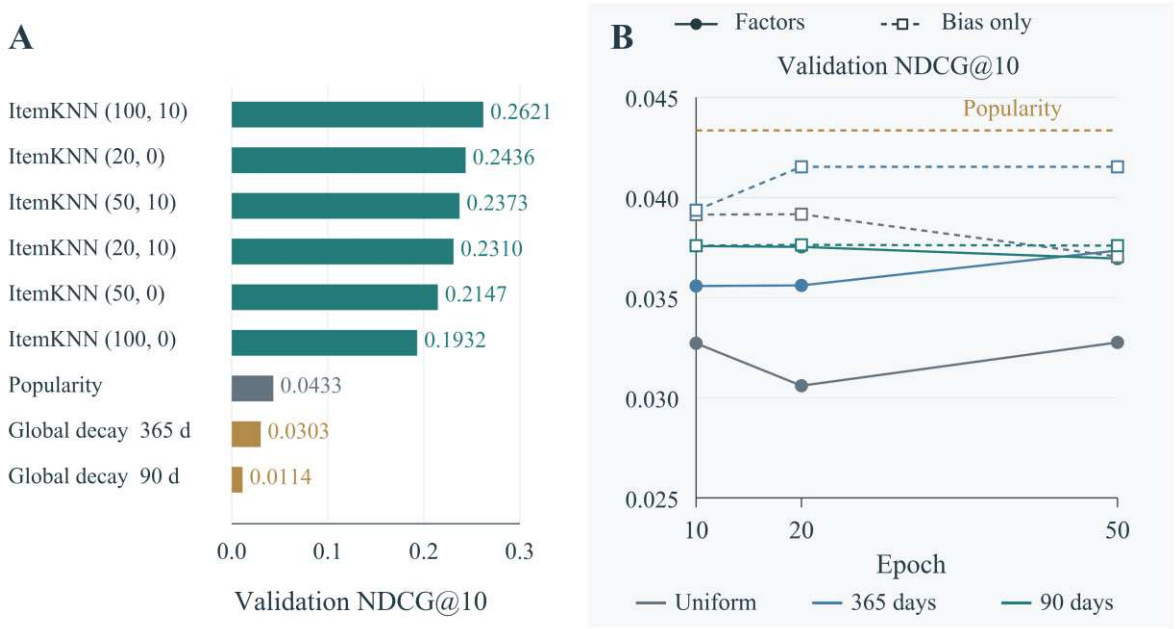


Fig 4. Exploratory validation. (A) Conventional baselines; ItemKNN labels give (neighbours, shrinkage). (B) Five-seed NDCG@10 over training duration.

* The extension used validation data after final evaluation.

5 Discussion

5.1 Interpreting personalization and recency

The contribution of personalization becomes clearest when shared item preference and temporal weighting are tested alongside account-specific factors. In this study, popularity remained the validation-selected method, while the selected factor model's higher final score was accompanied by an interval spanning deterioration and improvement. The controlled comparisons explain why this aggregate difference offers limited evidence for the added value of factors: the 365-day factor and bias-only models had identical mean NDCG@10, and their 90-day counterparts differed by only 0.000310. Strong agreement between their recommendation sets further suggests that shared item preferences shaped much of the observed ranking.

Kowald et al. (2024) combine frequency, recency and semantic context to model music relistening through human memory. Here, targets were absent from earlier matched histories, and temporal emphasis improved numerical final scores even in the shared-bias model. Fig 5 separates the factor contribution, recency contrast and their interaction. The interaction estimates provide no clear evidence that temporal weighting specifically strengthened account factors. Moreover, global decayed popularity performed poorly in validation, showing that weighting observations within accounts and weighting activity across the whole catalogue are not interchangeable operations. Supplementary Table S2 provides all saved final-evaluation NDCG@10 contrasts.

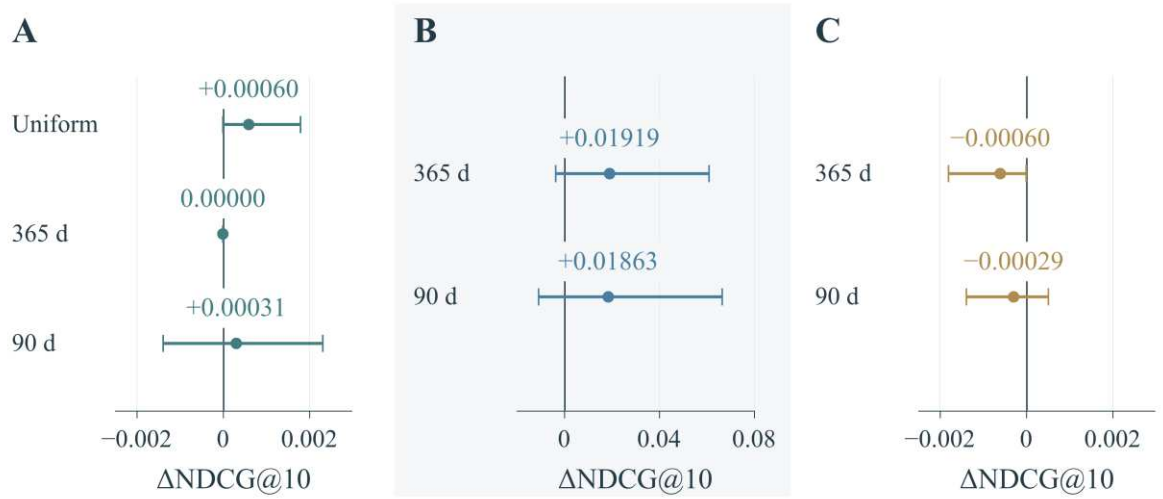


Fig 5. Controlled final-evaluation contrasts in NDCG@10. (A) Factors minus bias only under each weighting scheme. (B) Recency minus uniform weighting within factor models. (C) The recency contrast in factor models minus the corresponding bias-only contrast. Points show means; bars show 95% account-bootstrap intervals.

* d = days; axes differ; intervals are descriptive and unadjusted.

5.2 Strong baselines and the stability of apparent gains

Ferrari Dacrema et al. (2021) found that conventional methods could outperform many reproducible neural recommenders, emphasizing the consequences of baseline choice. The exploratory ItemKNN result is consistent with that methodological lesson: co-listening neighbourhoods recovered considerably more validation targets than either popularity or the tested factor models. ItemKNN also personalizes rankings through each account's history. Its promise therefore concerns how personal information is represented, rather than establishing that personalization itself is unnecessary. The result identifies a credible comparator for further evaluation; because this extension followed final-test disclosure, its validation advantage cannot serve as a new held-out performance claim.

Variation across accounts was more consequential than variation across fitted seeds. Only two accounts contributed nonzero NDCG@10 to the primary comparison, and omitting the benefiting account reversed its direction. Small seed variation therefore describes numerical stability within this cohort, not reliability across a wider listener population. Reporting the mean alone would obscure this distinction. The paired contrasts and account-omission display in Fig 3 make this dependency visible.

Longer training did not consistently improve validation ranking, despite continuing reductions in the sampled objective. This weakens the explanation that the initial ranking pattern was solely a consequence of stopping at ten epochs, while leaving broader optimization choices unresolved. A related caution appears in Abaddi (2026a), since adding foundation embeddings improved regional transfer relative to the transferred conventional model but worsened temporal prediction. Its relevance is methodological: additional representation capacity must be assessed against the comparator and prediction setting in which it will be used. The present evidence similarly supports conditional, component-specific claims.

5.3 Implications for recommendation and reproducible evaluation

The fixed catalogue could not recover the 43 out-of-catalogue final targets, regardless of model quality. Yet the attainable NDCG@10 ceiling of 0.904842 was far above the observed 0.027101 for the personalized challenger. Missing catalogue coverage therefore cannot explain most of the absolute ranking deficit. Retaining unavailable targets in denominators exposes that limitation; expanding the catalogue would instead require a new candidate policy with verified temporal availability.

The catalogue's most popular 100 items occupied 88.5% of the challenger's top-10 slots, so rearranging those items offers limited evidence of broader discovery. Klimashevskaia et al. (2024) argue that popularity bias must be interpreted through its consequences for stakeholders. Our concentration measures identify an exposure pattern but cannot establish listener dissatisfaction, artist disadvantage or unfairness. Yambda's separation of organic and recommendation-driven interactions illustrates the additional evidence needed to investigate such mechanisms (Ploshkin et al., 2025). Comparable exposure information and listener feedback would be required before recommending diversity interventions for this cohort.

The local execution demonstrates that this bounded comparison can be conducted with modest hardware: the exploratory extension finished within 16.32 minutes on one CPU worker. That observation supports practical repeatability at the present scale, not a production throughput claim. Retained configurations, predictions, checkpoints and hashes make individual findings traceable, while the documented exploratory chronology identifies which

conclusions require fresh validation. This addresses the emphasis on transparent experimental evidence in Said and Bellogín (2026). Independent execution remains necessary to establish reproducibility beyond the originating environment.

5.4 Limitations and priorities for further validation

The endpoint measures future catalogue assignments, not verified first encounters with recordings or artists. Retrospective matching can merge versions or miss valid links; exclusions reduce known ambiguity without quantifying residual error. Account hashes are not verified independent people, and identical-history exclusions cannot detect every shared or automated account. Requiring sufficient accepted training history and eligible future targets also creates a selected evaluation population. The 710 training accounts without eligible final targets may reflect inactivity, unmatched activity or exclusions; they should not be interpreted as confirmed listener attrition.

Global temporal cutoffs reduce interaction leakage of the kind identified by Ji et al. (2023), but they do not establish that every retrospective catalogue association was available in the same form at the historical prediction date. The six-month target window also differs from next-item prediction. Gusak et al. (2025) show that splitting and target selection can change model rankings, so comparisons with their sequential benchmarks concern evaluation design rather than transferable score levels. In our setting, model parameters and account factors remained fixed while exclusion masks advanced; an adaptively updated service could behave differently.

The final cohort comprised only 19 accounts, with sparse gains and substantial account influence. The bootstrap intervals are descriptive, conditional on the fitted seeds and not adjusted for the many contrasts. Repeated seeds do not increase the number of independent accounts. The controlled design also covers one factor dimension, fixed optimization settings and a restricted set of temporal weights; extending training duration is not equivalent to tuning the model family. No acoustic, contextual or sequential representation was evaluated, and offline target recovery does not establish satisfaction or causal recommendation effects.

Further validation should prioritize a larger, independently assessed cohort and a fresh chronological evaluation window, with the ItemKNN settings and factor comparators

selected before scoring. Where historical time coverage permits, additional earlier development windows could assess whether the observed rankings persist across periods. Catalogue-link uncertainty, available targets and account dependence should remain explicit in that design. These steps would test whether the strong neighbourhood result generalizes and whether recent listening contributes information beyond shared item preference, without reusing the completed final test for model selection.

6 Conclusion

Useful personalization depends on the information it adds beyond shared listening patterns. This study separated account-specific factors, shared item preference and recency in chronological recommendation of catalogue items absent from earlier matched histories. The personalized challenger achieved a higher final NDCG@10 than popularity, but the paired interval crossed zero and the direction depended on individual accounts. Closely aligned factor and bias-only rankings further limited the evidence for an incremental factor contribution. Recency changed scores without establishing a factor-specific benefit. The frozen evaluation therefore does not establish a reliable personalization advantage in this cohort.

The contribution is an auditable comparison that keeps model effects, catalogue coverage and account influence distinguishable. The later exploratory validation analysis identified item neighbourhoods as a promising alternative, whereas extending training did not consistently improve factor ranking. Larger independently assessed cohorts, stronger catalogue linkage and fresh chronological evaluation windows are the next priorities. Selecting ItemKNN and factor comparators before scoring would test whether richer listening structure improves recommendation beyond shared preference while preserving a clear boundary between development and final evidence.

Declarations

Competing interests

The author declares no financial or non-financial competing interests.

Funding

This study received no external funding.

Data availability

Source listening records and catalogue metadata are available through ListenBrainz, MTG-Jamendo and JamendoMaxCaps under their respective terms. Derived pseudonymous inputs and numerical results are retained in the private study repository for authorized review. Source-account mappings, raw history archives and audio are excluded from that repository; access to permitted review materials may be requested from the corresponding author.

Code availability

Code, configurations, checkpoints, predictions and reproduction notes are preserved in the private Reclilt repository, snapshot version 0.1, commit be929083a95a. The code and permitted supporting materials will be made public after acceptance, following the documented rights and privacy review. Third-party data and software retain their respective terms.

Supplementary videos

Three videos support this submission.

Video 1, Personalization and Recency Comparison, documents the controlled development comparison. YouTube: https://youtube.com/shorts/6n6w5Gvw6z0

Video 2, Exploratory Baselines and Training Duration, documents the subsequent validation-only exploratory extension. YouTube: https://youtu.be/EEtWNhDNtGY

Video 3, Reclilt Repository and Research Materials, documents the repository organization and research materials. YouTube: https://youtu.be/A-5RUqdKLZM

Ethics approval and human participation

Not applicable to the reported retrospective computational analysis. No participants were recruited, and no intervention was performed for the recommendation experiments.

Consent for publication

Not applicable. The manuscript reports aggregate findings without identifiable participant information.

Author contributions

Samer Abbadi: Conceptualization, methodology, investigation, data curation, software, validation, formal analysis, visualization, writing—original draft, and writing—review and editing.

Use of artificial intelligence

AI tools (Grammarly AI and Claude.ai) were used for proofreading and improving the structure and flow of the manuscript. The final approval of the scientific content and results rests with the author.

References

Abaddi, S. (2026a). *AlphaEarth and evapotranspiration prediction in Jordan* [Preprint]. Research Square. https://doi.org/10.21203/rs.3.rs-11417788/v1

Abaddi, S. (2026b). *Failure prediction and presence cues for budgeted review of Saudi road images* [Preprint]. engrXiv. https://doi.org/10.31224/8654

Bogdanov, D., Won, M., Tovstogan, P., Porter, A., & Serra, X. (2019). *The MTG-Jamendo dataset for automatic music tagging* [Workshop paper]. ICML 2019 Workshop on Machine Learning for Music Discovery. https://repositori.upf.edu/bitstream/handle/10230/42015/bogdanov_ICML2019__Jamendo.pdf

Deldjoo, Y., Schedl, M., & Knees, P. (2024). Content-driven music recommendation: Evolution, state of the art, and challenges. *Computer Science Review, 51*, Article 100618. https://doi.org/10.1016/j.cosrev.2024.100618

Ferrari Dacrema, M., Boglio, S., Cremonesi, P., & Jannach, D. (2021). A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems, 39*(2), Article 20. https://doi.org/10.1145/3434185

Gusak, D., Volodkevich, A., Klenitskiy, A., Vasilev, A., & Frolov, E. (2025). Time to split: Exploring data splitting strategies for offline evaluation of sequential recommenders. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems* (pp. 874–883). Association for Computing Machinery. https://doi.org/10.1145/3705328.3748164

International Federation of the Phonographic Industry. (2026). *Global music report 2026: State of the industry*. https://www.ifpi.org/wp-content/uploads/2026/03/GMR2026_SOTI.pdf

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems, 20*(4), 422–446. https://doi.org/10.1145/582415.582418

Ji, Y., Sun, A., Zhang, J., & Li, C. (2023). A critical study on data leakage in recommender system offline evaluation. *ACM Transactions on Information Systems, 41*(3), Article 75. https://doi.org/10.1145/3569930

Klimashevskaia, A., Jannach, D., Elahi, M., & Trattner, C. (2024). A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction, 34*(5), 1777–1834. https://doi.org/10.1007/s11257-024-09406-0

Kowald, D., Reiter-Haas, M., Kopeinik, S., Schedl, M., & Lex, E. (2024). Transparent music preference modeling and recommendation with a model of human memory theory. In B. Ferwerda, M. Graus, P. Germanakos, & M. Tkalčič (Eds.), *A human-centered perspective of intelligent personalized environments and systems* (pp. 113–136). Springer. https://doi.org/10.1007/978-3-031-55109-3_4

MetaBrainz Foundation. (n.d.). *Data downloads*. ListenBrainz. Retrieved November 14, 2026, from https://listenbrainz.org/data/

Ploshkin, A., Tytskiy, V., Pismenny, A., Baikalov, V., Taychinov, E., Permiakov, A., Burlakov, D., & Krofto, E. (2025). Yambda-5B—A large-scale multi-modal dataset for ranking and retrieval. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems* (pp. 894–901). Association for Computing Machinery. https://doi.org/10.1145/3705328.3748163

Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2009). BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (pp. 452–461). AUAI Press. https://www.ismll.uni-hildesheim.de/pub/pdfs/Rendle_et_al2009-Bayesian_Personalized_Ranking.pdf

Roy, A., Liu, R., Lu, T., & Herremans, D. (2025). *JamendoMaxCaps: A large scale music-caption dataset with imputed metadata* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2502.07461

Said, A., & Bellogín, A. (2026). Reproducibility in recommender systems: A survey. *ACM Transactions on Recommender Systems*. Advance online publication. https://doi.org/10.1145/3831686

Schedl, M., Zamani, H., Chen, C.-W., Deldjoo, Y., & Elahi, M. (2018). Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval, 7*(2), 95–116. https://doi.org/10.1007/s13735-018-0154-2

Supplementary materials

Table S1. Complete exploratory validation results across 27 accounts

Configuration	Epoch	NDCG@10	Seed SD	Recall@10	Recall@20
Popularity	—	0.043350	—	0.058985	0.104252
Global decay 365 days	—	0.030268	—	0.023398	0.031981
Global decay 90 days	—	0.011368	—	0.009132	0.025122
ItemKNN (20, 0)	—	0.243613	—	0.339859	0.381707
ItemKNN (20, 10)	—	0.231007	—	0.312434	0.413277
ItemKNN (50, 0)	—	0.214737	—	0.300754	0.377239
ItemKNN (50, 10)	—	0.237251	—	0.287390	0.459573
ItemKNN (100, 0)	—	0.193160	—	0.256310	0.321683
ItemKNN (100, 10)	—	0.262058	—	0.330952	0.459926
Bias only uniform	10	0.039145	0.000000	0.054870	0.104252
Bias only uniform	20	0.039165	0.000047	0.054870	0.104252
Bias only uniform	50	0.037045	0.004753	0.047462	0.104252
Bias only 365 days	10	0.039362	0.005864	0.072977	0.104605
Bias only 365 days	20	0.041532	0.004805	0.080384	0.104605
Bias only 365 days	50	0.041532	0.004723	0.080384	0.104605
Bias only 90 days	10	0.037593	0.000000	0.050754	0.100490
Bias only 90 days	20	0.037635	0.000057	0.050754	0.100490
Bias only 90 days	50	0.037595	0.000124	0.050754	0.100490
Factors uniform	10	0.032721	0.005864	0.032647	0.104252
Factors uniform	20	0.030601	0.004776	0.025240	0.104252
Factors uniform	50	0.032763	0.005826	0.032647	0.104252
Factors 365 days	10	0.035580	0.004702	0.058232	0.104605
Factors 365 days	20	0.035609	0.004682	0.058232	0.104605
Factors 365 days	50	0.037352	0.005944	0.065569	0.104605
Factors 90 days	10	0.037574	0.000110	0.050754	0.100490
Factors 90 days	20	0.037534	0.000089	0.050754	0.100490
Factors 90 days	50	0.036950	0.001088	0.050754	0.100490

* SD describes seed variation.

Table S2. Final-evaluation NDCG@10 contrasts and account-level directions

Contrast	Mean difference	95% interval	+ / − / =
Factors − bias, uniform	0.000600	[0.000000, 0.001799]	1 / 0 / 18
Factors − bias, 365 days	0.000000	[0.000000, 0.000000]	0 / 0 / 19
Factors − bias, 90 days	0.000310	[-0.001384, 0.002315]	1 / 1 / 17
Recency in factors, 365 days	0.019186	[-0.003523, 0.061082]	1 / 1 / 17
Recency in factors, 90 days	0.018627	[-0.010737, 0.066618]	1 / 1 / 17
Recency interaction, 365 days	-0.000600	[-0.001799, 0.000000]	0 / 1 / 18
Recency interaction, 90 days	-0.000289	[-0.001384, 0.000517]	1 / 1 / 17
Bias uniform − popularity	-0.007451	[-0.022354, 0.000000]	0 / 1 / 18
Bias 365 days − popularity	0.012335	[-0.024078, 0.061082]	1 / 1 / 17
Bias 90 days − popularity	0.011465	[-0.033608, 0.068002]	1 / 1 / 17
Factors uniform − popularity	-0.006852	[-0.020555, 0.000000]	0 / 1 / 18
Factors 365 days − popularity	0.012335	[-0.024078, 0.061082]	1 / 1 / 17
Factors 90 days − popularity	0.011775	[-0.031292, 0.066618]	1 / 1 / 17

* +/−/= denote positive/negative/tied account differences.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Video1PersonalizationandRecencyComparison.mp4
- Video2ExploratoryBaselinesandTrainingDuration.mp4
- Video3ReclIltRepositoryandResearchMaterials.mov