Beyond Collaborative Filtering: Using Transformers for Personalized Music Recommendation

Timothy Greer
Amazon Music
greertim@amazon.com

Nicholas Capel*

Emanuele Coviello
Amazon Music
emacov@amazon.com

Amina Shabbeer
Amazon Music
ashabb@amazon.com

Abstract

Music recommendation systems face the dual challenge of capturing both immediate context and long-term preferences in users' listening patterns. We adapt a generalized sequential model architecture for music recommendation, introducing modifications that acknowledge how music preferences combine temporal patterns and stable tastes. By removing causal masking constraints typically used in sequential models, we better capture users' overall preferences rather than strictly sequential patterns. This technique achieves approximately 28% improvement in F1 scores compared to a neural item-item baseline. Through ablation studies, we show that using positional encoding and removing the causal mask during training results in the best personalized recommendations. Our findings demonstrate that transformer-based architectures can effectively model music preferences while being computationally efficient for large-scale deployment.

1 Introduction

Music recommendation is a critical component of modern streaming systems, with ranking systems playing a pivotal role in surfacing relevant tracks to customers in a context-aware manner that matches their longer-term tastes. While traditional collaborative filtering approaches have proven effective for general recommendation tasks, the *sequential* nature of music listening presents unique challenges and opportunities. Users' music preferences exhibit both long-term patterns reflecting stable tastes and short-term dynamics influenced by mood and context: both should be leveraged for candidate generation and recommendation.

Recent advances in transformer architectures offer a promising direction for capturing these complex patterns. In this paper, we present a transformer decoder architecture that builds on the Generalized Self-Attentive Sequential Recommendation (gSASRec) framework to create personalized music recommendations. Our approach adapts this architecture for music sequence modeling while remaining computationally efficient for large-scale deployment. Experimental results show significant improvements over existing methods in both immediate contextual relevance and longer-term preference modeling.

Music recommendation systems must balance immediate contextual relevance with long-term preference modeling [12]. While collaborative filtering approaches have shown success [5], recent work demonstrates the potential of transformer architectures for capturing both sequential patterns and stable preferences [1]. However, existing transformer-based music recommenders [8] may overemphasize strict sequential dependencies at the expense of modeling stable user preferences.

*work done while at Amazon Music

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

2 Related Work

2.1 Sequential and General Recommendation

Recommendation systems traditionally balance two key approaches: modeling sequential patterns in user behavior and capturing general user-item affinities. While early work focused on Markov Chain models for sequential patterns [11], the field has evolved to include sophisticated deep learning approaches like GRU4Rec [4] and Caser [14] for modeling temporal dependencies.

The introduction of transformer architectures, particularly SASRec [7], demonstrated how self-attention mechanisms could theoretically capture both sequential patterns and general preferences. The gSASRec model [9] improved this approach by addressing the overconfidence problem in prediction through a *generalized* binary cross-entropy loss. Recent findings suggest that strict adherence to temporal ordering may be less critical than previously thought, as demonstrated by improved recommendation quality when using bidirectional temporal patterns rather than purely sequential approaches [6].

2.2 Music Recommendation

Music recommendation poses unique challenges for creating personalized experiences. Traditional collaborative filtering approaches [5] excel at capturing user-item affinities, while session-based approaches [10] can model contextual preferences. Recent work has explored transformer architectures for music recommendation [2, 3], primarily focusing on content-based features. Our work investigates using transformer-based models like gSASRec for music recommendation, exploring methods that use both sequence information and do not sequence information.

Other studies have investigated the relative importance of sequential patterns in music consumption. Vall et al. [15] found that while song context improves playlist generation, strict song order appears less crucial than previously thought. Schweiger et al. [13] investigated sequential patterns in user-generated playlists, finding less variance between sequential tracks than random pairs. These findings suggest that while strict track ordering may be flexible, local context and transitions remain important considerations in modeling music consumption patterns.

3 Methods

Our model builds on the gSASRec architecture [9], which combines self-attention (Attention$(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V$) with position-wise feed-forward networks (FFN$(x) = \text{ReLU}(xW^{(1)} + b^{(1)})W^{(2)} + b^{(2)}$). While traditional sequential recommendation frames the problem as predicting the next item given a sequence, we recognize that music listening patterns often reflect stable user preferences more than strict sequential dependencies. We therefore using positional encodings P and causal mask M as hyperparameters. Typically, we compute H_0 and apply a mask M_{ij} such that:

$$H_0 = E + P; \quad M_{ij} = 0 \text{ if } i \geq j \text{ else } -\infty$$

We train a model without these so it treats user listening as an unordered set of preferences rather than a strict sequence:

$$H_0 = E; \quad \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \text{ (without masking)}$$

To address the overconfidence problem common in sequential recommendation, we employ a generalized binary cross-entropy loss:

$$\mathcal{L}\text{gBCE}^{\beta} = -\frac{1}{|I_k^-| + 1}\left(\log(\sigma^{\beta}(s_{i^+})) + \sum_{i \in I_k^-} \log(1 - \sigma(s_i))\right)$$

where β controls the calibration of predicted probabilities and I_k^- represents the set of sampled negative items. We set β to .9 based on hyperparameter optimization.

Table 1: F1 scores at different cutoffs k. We sampled 1 million customers randomly, and the 95% confidence intervals between the sequential model variants and the Baseline (neural item-item) model are non-overlapping.

k	Baseline	Proposed (mask, no PE)	Proposed w/o causal masking	Proposed w/ PE	Proposed w/ PE, w/o causal masking
10	0.0178	0.0196	0.0175	0.0200	**0.0228**
50	0.0148	0.0164	0.0167	0.0168	**0.0190**
100	0.0120	0.0133	0.0136	0.0136	**0.0154**
200	0.0087	0.0097	0.0099	0.0100	**0.0112**

For models with causal masking, we use teacher forcing during training where each prediction is conditioned on the true previous items. For variants without causal masking, all items in the input are considered simultaneously without sequential constraints. All models employ dropout with a rate of .0625 and layer normalization, parameters chosen through hyperparameter tuning. At inference time, we generate recommendations by computing scores for all candidate tracks and selecting the top k highest-scoring items. Training was performed using 8 NVIDIA A100 GPUs.

We evaluate our approach on a proprietary dataset from Amazon Music containing 14 million users and 50 million tracks collected over 60 days in 2024. While public datasets offer reproducibility benefits, our production dataset better reflects real-world music consumption patterns with detailed playthrough signals that enable more nuanced evaluation of user preferences versus sequential effects.

4 Evaluation Methodology

4.1 Profile-Based Evaluation Dataset Construction

Traditional sequential recommendation evaluation focuses on next-item prediction accuracy. However, in music streaming, a user's next track choice can depend more on their overall preferences than previous playbacks, particularly if the next item is part of a new listening session. We therefore developed an evaluation protocol that emphasizes stable user preferences over sequential patterns.

We constructed our evaluation dataset using seven days of held-out listening history data. For each user profile, we classified track playbacks as positive ($\geq$ 30 seconds of listening) or negative ($<$ 30 seconds) and computed Wilson scores for track playthrough probabilities to account for observation uncertainty [16]. We then ranked these tracks in descending order by their Wilson scores on playthrough rate (positive playback divided by all playbacks). To ensure data quality, we filtered out profiles with fewer than three playbacks and removed tracks with only one playback per profile.

4.2 Evaluation Metrics

We evaluate models based on their ability to rank tracks according to users' estimated true preferences, as measured by Wilson score-adjusted playthrough rates. This method reduces the impact of presentation bias caused by the existing recommendation system. By evaluating against stable listening patterns over a weeklong period rather than individual track playbacks, we can better assess whether models are capturing enduring user preferences rather than just short-term sequential effects.

5 Results

5.1 Profile-to-Track Alignment

We held out seven days of customer listening data and trained both our model's and a neural item-item baseline model's track and profile embeddings. We evaluated the models' ability to predict tracks that were listened to for more than 30 seconds for each profile, as described in Section 4.2.

Table 1 shows that our proposed model consistently outperforms the baseline across all evaluation cutoffs, with the strongest performance from the variant using positional encoding without causal masking. The improvements are most pronounced at lower k values, suggesting particularly effective identification of users' highest-probability tracks.

Table 2: F1 scores by history length cohort at different cutoffs k. For each cohort, we sampled 250,000 million customers randomly, and the 95% confidence intervals between the sequential model variants and the Baseline (neural item-item) model are non-overlapping.

Hist. Length	k	Baseline	Proposed	Proposed w/o masking	Proposed w/ PE	Proposed w/ PE, w/o masking
(0, 50]	10	0.0165	0.0170	0.0173	0.0181	**0.0185**
	50	0.0124	0.0128	0.0129	0.0136	**0.0144**
	100	0.0098	0.0101	0.0102	0.0108	**0.0113**
	200	0.0070	0.0073	0.0074	0.0078	**0.0082**
(50, 100]	10	0.0171	0.0193	0.0195	0.0203	**0.0214**
	50	0.0145	0.0151	0.0153	0.0158	**0.0167**
	100	0.0115	0.0120	0.0121	0.0126	**0.0133**
	200	0.0082	0.0086	0.0087	0.0091	**0.0096**
(100, 150]	10	0.0181	0.0190	0.0192	0.0202	**0.0213**
	50	0.0145	0.0154	0.0156	0.0161	**0.0173**
	100	0.0118	0.0123	0.0125	0.0129	**0.0137**
	200	0.0085	0.0089	0.0090	0.0094	**0.0099**
>150	10	0.0190	0.0211	**0.0213**	0.0212	0.0212
	50	0.0161	0.0171	**0.0173**	0.0172	0.0172
	100	0.0128	0.0138	**0.0140**	0.0139	0.0139
	200	0.0094	0.0101	0.0102	**0.0103**	**0.0103**

5.2 Cohort Analysis

We analyzed performance across different user cohorts, categorized by listening volume (in tracks) in the sixty days *before* the evaluation data start date. Table 2 shows F1 scores across cohorts, where performance improvements of the sequential models are consistent across all user segments.

The breakdown by cohort reveals several important patterns in model performance. For users with limited history (0-50 tracks), adding positional encoding (PE) yields a substantial improvement, suggesting that for light users, some temporal information helps bootstrap recommendations despite limited data. For high-volume listeners (>150 tracks), the benefits of our modifications diminish. This suggests that for users with extensive listening history, simple preference modeling becomes increasingly effective, and complex temporal dynamics play a reduced role. This aligns with previous findings that heavy listeners often have more stable, well-defined preferences [13]. The strong performance of variants without causal masking suggests that music preferences combine both sequential and stable components. While traditional transformer architectures use strict masking to model pure sequential dependencies [7], our results indicate this may be overly restrictive for music recommendation. Removing the causal mask allows the model to consider a user's full listening history when making predictions, better capturing their overall music taste profile rather than just recent listening context. This aligns with recent findings that music consumption patterns exhibit both temporal and preference-based structure [1].

6 Conclusion

We presented a transformer-based architecture for music recommendation that significantly outperforms neural item-item baselines by better balancing immediate context and long-term preferences. Our evaluation demonstrates that removing causal masking while maintaining positional encoding leads to the best performance, particularly for users with rich listening histories. This suggests that while temporal context matters in music recommendation, strict sequential ordering may be less critical than previously assumed. Future work will explore hybrid approaches combining sequential and similarity-based models while maintaining computational efficiency.

Acknowledgments and Disclosure of Funding

This work would not have been possible without the support and code contributions to a common codebase at Amazon Music from Giuseppe Di Benedetto, Yannik Stein, Kuntal Ganguly, and others.

References

[1] Walid Bendada, Théo Bontempelli, Mathieu Morlon, Benjamin Chapus, Thibault Cador, Thomas Bouabça, and Guillaume Salha-Galvan. Track mix generation on music streaming services using transformers. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 112–115, 2023.

[2] Shuo Chen, Josh L Moore, Douglas Turnbull, and Thorsten Joachims. Playlist prediction via metric embedding. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 714–722, 2012.

[3] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. Transfer learning for music classification and regression tasks. *arXiv preprint arXiv:1703.09179*, 2017.

[4] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*, 2015.

[5] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining*, pages 263–272. Ieee, 2008.

[6] Juyong Jiang, Peiyan Zhang, Yingtao Luo, Chaozhuo Li, Jae Boum Kim, Kai Zhang, Senzhang Wang, Sunghun Kim, and Philip S Yu. Improving sequential recommendations via bidirectional temporal data augmentation with pre-training. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[7] Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pages 197–206. IEEE, 2018.

[8] Shimiao Liu and Alexander Lerch. Enhancing video music recommendation with transformer-driven audio-visual embeddings. In *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, pages 1–6. IEEE, 2024.

[9] Aleksandr Vladimirovich Petrov and Craig Macdonald. gsasrec: Reducing overconfidence in sequential recommendation trained with negative sampling. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 116–128, 2023.

[10] Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. Sequence-aware recommender systems. *ACM computing surveys (CSUR)*, 51(4):1–36, 2018.

[11] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 811–820, 2010.

[12] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, 7(2):95–116, 2018.

[13] Harald Victor Schweiger, Emilia Parada-Cabaleiro, and Markus Schedl. Does track sequence in user-generated playlists matter?. In *ISMIR*, pages 618–625, 2021.

[14] Jiaxi Tang and Ke Wang. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 565–573, 2018.

[15] Andreu Vall, Massimo Quadrana, Markus Schedl, and Gerhard Widmer. The importance of song context and song order in automated music playlist generation. *arXiv preprint arXiv:1807.04690*, 2018.

[16] Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.