# Generalized User Representations for Large-Scale Recommendations and Downstream Tasks

Ghazal Fazelnia, Sanket Gupta, Claire Keum, Mark Koh, Timothy Heath, Guillermo Carrasco Hernández, Stephen Xie, Nandini Singh, Ian Anderson, Mounia Lalmas, Maya Hristakeva, Petter Pehrson Skidén
{ghazalf,sanketg,ckeum,markkoh,theath,guillermoc,stephenxnandinis,iananderson,mounial,mayah,peppe}@spotify.com
Spotify

## ABSTRACT

Accurately capturing diverse user preferences at scale is a core challenge for large-scale recommender systems like Spotify's, given the complexity and variability of user behavior. To address this, we propose a two-stage framework that combines representation learning and transfer learning to produce generalized user embeddings. In the first stage, an autoencoder compresses rich user features into a compact latent space. In the second, task-specific models consume these embeddings via transfer learning, removing the need for manual feature engineering.

This approach enhances flexibility by allowing dynamic updates to input features, enabling near-real-time responsiveness to user behavior. The framework has been deployed in production at Spotify with an efficient infrastructure that allows downstream models to operate independently. Extensive online experiments in a live setting show significant improvements in metrics such as consumption share, content discovery, and search success. Additionally, our method achieves these gains while substantially reducing infrastructure costs.

## KEYWORDS

user model, embeddings, recommender systems, cold-start models

**ACM Reference Format:**
Ghazal Fazelnia, Sanket Gupta, Claire Keum, Mark Koh, Timothy Heath, Guillermo Carrasco Hernández, Stephen Xie, Nandini Singh, Ian Anderson, Mounia Lalmas, Maya Hristakeva, Petter Pehrson Skidén. 2025. Generalized User Representations for Large-Scale Recommendations and Downstream Tasks. In *18th ACM Conference on Recommender Systems (RecSys '25), September 22–26, 2025, Prague, Czech Republic.* ACM, New York, NY, USA, 4 pages. https://doi.org/XX.XXX/XXXXXXX.XXXX

## 1 INTRODUCTION

Online music streaming services have grown increasingly popular in recent years. Music recommender systems face different challenges than traditional recommender systems. The tracks are often short, and several tracks can be played in a single session without any feedback [11]. Users also have conflicting desires: revisiting favorite tracks while discovering new music to diversify their experiences [10, 15]. Furthermore, relying on interaction and consumption as primary signals for training recommender systems makes it particularly challenging to provide recommendations for new users [12, 13]. Despite recent advancements in user modeling, accurately capturing and representing user interests in large-scale music streaming services remains a challenging task [9]. Treating each task independently (e.g., retrieval from long-term taste, bandits for new users) can increase system complexity and lead to disjointed early user experiences due to uncoordinated model interactions [2, 7].

In this paper, we detail how our model builds an effective user representation, address the challenges associated with this framework, and demonstrate how it captures both holistic user interests and current preferences. Our contributions aim to answer the following three questions: **[RQ1]** How can we efficiently design a model to capture rich user representation in a large-scale recommender system that encompasses core user interests while being adaptable to various downstream tasks? **[RQ2]** How can we design a user representation model that works for cold-start users and improves as users become more established on the platform? and **[RQ3]** How can we devise effective evaluation strategies to measure the efficacy of the vector embedding space to ensure value across diverse downstream tasks?

We address these questions through a two-stage process. In the first stage, we develop an autoencoder model that takes in various user features such as listening history, demographics, and contextual information, to learn rich generalized user representations. The second stage applies these learned representations in a real-world transfer learning paradigm, adapting them for a range of applications. Finally, we demonstrate the advantages of our method compared to baseline approaches. Our approach supports diverse downstream tasks through a shared, adaptable user embedding. Using Spotify's interbal data, we demonstrate strong performance and highlight each component's contribution. This system is now deployed in large-scale production.

## 2 APPROACH

Figure 1 presents an overview of the architecture of our generalized user representation model. Initially, modality encoders preprocess the track features and represent each track with 80-dimensional vectors derived from co-occurrence statistics and audio features. This embedding space has been shown to perform well for music recommendation tasks [1, 4]. These track embeddings, combined with other user data such as demographics, onboarding signals, and

*RecSys '25, September 22–26, 2025, Prague, Czech Republic*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XX.XXX/XXXXXXX.XXXX

contextual information, form the input features used in the main model training to generate the user representations. The resulting learned representations can then be used as feature inputs for a variety of recommendation tasks. This preprocessing steps enables faster training, higher efficiency and low-latency for training and serving the resulting representations.

## 3 TRANSFER LEARNING

Our approach supports large-scale recommendation tasks such as retrieval, ranking, and generation,using a generalized user representation that transfers well across systems.

*Responsiveness:* Fast inference alone does not guarantee responsiveness to changes in user preferences, especially when relying on stale batch features. To address this, we employ Near-Real-Time (NRT) inference triggered by user activity events. These events are pre-processed, passed through the model, and stored in an online feature store within minutes. BY combining NRT with batch inference, we support both active and inactive users effectively.

*Cold-Start Awareness:* During onboarding, users may select preferred artists or languages. These signals are encoded and passed through the same model used for existing users, enabling immediate personalization. We combine onboarding features with demographic information and gradually transition to behavior-based inputs as more data becomes available.

*Stability:* For effective transfer learning, user representations must evolve within stable vector spaces that preserve semantic meaning over time. Although periodic retraining is necessary to combat model drift [8], unsynchronized updates can disrupt downstream dependencies. To mitigate this, we implement Batch Management, a coordinated retraining strategy that assigns a unique batch ID to each update. Downstream models are retrained in sync, ensuring that embeddings are only compared within the same batch. Both training and inference operate within a consistent batch context, while production systems continue running on a "legacy" batch during updates. This approach guarantees synchronization, consistent comparisons, and uninterrupted service, preserving stability across the entire model pipeline.

## 4 EMPIRICAL STUDIES

We evaluate our framework across a range of tasks and baselines, targeting both cold-start and established users. The training dataset is constructed using audio streaming data drawn from Spotify's global user base. For evaluation, we sample over five million users at random, ensuring a representative distribution by considering attributes such as country of registration, subscription status (free or paid), and account age. Online performance is assessed through A/B tests conducted on production traffic.

To measure the effectiveness of our generalized user representations, we first examine their performance in future track prediction tasks. For established users, we use a 7-day prediction window and compare our approach to several baseline models, including matrix factorization techniques (NMF [14], PMF [5]), deep learning models (LightFM [2], DLRM [6])), VAE-CF [3], and embeddings averaged from users' listening histories. All models are aligned in terms of embedding dimensionality. As reported in Table 1, our model consistently outperforms these baselines, validating its predictive power and supporting our first research question (RQ1).

**Table 1: Performance for predicting listening within next seven days for established users using generalized user representation.**

| Comparison Vs | Accuracy | AUC |
|---|---|---|
| NMF | +15.2% | +18.6% |
| PMF | +10.1% | +12.7% |
| LightFM | +2.3% | +3.9% |
| DLRM | +1.5% | +2.8% |
| VAE-CF | +4.0% | +5.7% |
| average embeddings | +1.8% | +1.6% |

**Table 2: Performance for predicting listening within four hours following user registration.**

| Comparison Vs | Onboarding Status | Accuracy | AUC |
|---|---|---|---|
| popularity heuristic | Completed | +26.2% | +27.0% |
| popularity heuristic | Not Completed | +24.6% | +24.7% |
| average embeddings | Completed | +5.0% | +5.1% |

**Table 3: Performance at finding similar clusters for user representation vs. average embeddings.**

| Cluster Heuristic | nDCG@50 |
|---|---|
| Same favorite artists | +2.9% |
| Same country of most listened artists | +5.5% |
| Same new user onboarding | +26.2% |

For cold-start scenarios, we evaluate performance over a 4-hour window. Here, our method is compared to baselines that average onboarding artist embeddings or use popular tracks inferred from demographic features. Results in Table 2 show that our unified representation provides clear improvements, effectively capturing early user intent even in the absence of complete onboarding information. This supports RQ2 and demonstrates the model's robustness in sparse data settings.

We also conduct a clustering-based evaluation to assess the intrinsic quality of our embeddings. By comparing nDCG@50 values obtained through nearest-neighbor lookups, we find that our representation outperforms those derived from averaged item embeddings across diverse user clusters. These findings, presented in Table 3, lend support to RQ3 and highlight the model's ability to group users in behaviorally meaningful ways.

Finally, we test the applicability of our user representations in real-world scenarios by integrating them into downstream production models. Our offline experiments and online A/B tests demonstrate that the learned embeddings provide meaningful transferability and measurable impact when deployed in live systems.

Table 4 provides a summary of the key results achieved across various applications of the generalized user model.

*Candidate generation.* The model powers the generation of album and playlist candidates in Home page shelves. It led to significant increases in album discoveries and impression-to-stream (i2s) rates,

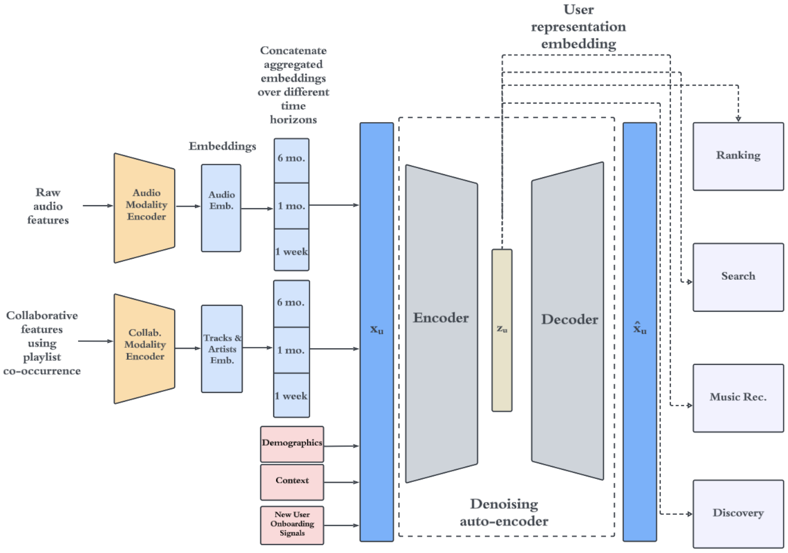

**Figure 1: Generalized User Representation Model Architecture. Catalog interactions are embedded using pre-trained modality encoders, which then serve as input to the autoencoder model. The resulting representation serves as the User Representation for various downstream recommendation tasks.**

reinforcing the value of using a generalized user representation for early-stage retrieval.

*Search.* Used to re-rank search results to personalize the search results, the model yielded a +0.06% overall improvement and a +0.76% increase in podcast search success, which highlights effectiveness across modalities. These gains, particularly in a highly optimized system, underscore the model's ability to effectively capture cross-modal user behaviors through its embeddings.

*Home Ranking Model.* Applied to rank item shelves on the Home page, the model enhanced music discovery and increased the share of Home content consumption. The results show a successful shift in user engagement from familiar content toward new discoveries, reflecting the model's ability to encourage broader exploration.

*Artist preference modeling.* This captures user-artist affinity and is leveraged across several recommendation and ranking scenarios. We achieved a 50% reduction in infrastructure and feature-related costs without affecting top-line metrics, resulting in simplified models, reduced technical debt, and faster deployment cycles.

To investigate the relative importance of features, we performed an ablation study on the input features. We removed specific features to assess their impact.

When we exclude new user onboarding signals, we observe a 13.8% decrease in nDCG@50 for clusters formed using the same new user onboarding data. This indicates that without onboarding signals, the model's ability to identify new user clusters is diminished. Excluding modality encoder embeddings results in a 4.2% decrease in AUC for the future listening evaluation over 7 days. Additionally, we see a 37.1% drop in nDCG@50 for clusters formed based on favorite artists. This highlights the importance of modality encoder embeddings to the system's success. Finally, omitting user based static features, such as the country of registration, leads to a 12.1% decrease in nDCG@50 for clusters formed based on the country of most listened artists.

**Table 4: Downstream performance using generalized user embeddings: Online results from candidate generation and search models, as well as end-to-end performance in the Home Ranking model, demonstrate the effectiveness of transfer learning.**

| **Downstream Model** | **Metric 1** | **Metric 2** |
|---|---|---|
| Candidate Generation Model | **Discoveries**<br>**+2.9%** | **Shelf level i2s**<br>**+13%** |
| Search Model | **Overall Search Success**<br>**+0.06%** | **Podcast Success**<br>**+0.76%** |
| Home Ranking Model | **Music Discovery Success**<br>**+0.20%** | **Consumption Share**<br>**+.05%** |

## 5 CONCLUSION

After extensive offline and online validation, our approach for learning generalized user representations has been successfully deployed at Spotify. While developed for music, the framework generalizes well to broader audio domains and other verticals like news and e-commerce. Future work could extend the model by incorporating additional modalities, such as lyrics, playlists, and album titles, and enriching representations with embeddings from Large Language Models.

# REFERENCES

[1] Ghazal Fazelnia, Eric Simon, Ian Anderson, Benjamin Carterette, and Mounia Lalmas. 2022. Variational user modeling with slow and fast features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 271–279.

[2] Maciej Kula. 2015. Metadata Embeddings for User and Item Cold-start Recommendations. In *Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender Systems co-located with 9th ACM Conference on Recommender Systems (RecSys 2015), Vienna, Austria, September 16-20, 2015. (CEUR Workshop Proceedings, Vol. 1448)*, Toine Bogers and Marijn Koolen (Eds.). CEUR-WS.org, 14–21. http://ceur-ws.org/Vol-1448/paper4.pdf

[3] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*. 689–698.

[4] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th acm international conference on information and knowledge management*. 2243–2251.

[5] Andriy Mnih and Russ R Salakhutdinov. 2008. Probabilistic matrix factorization. In *Advances in neural information processing systems*. 1257–1264.

[6] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G. Azzolini, Dmytro Dzhulgakov, Andrey Mallevich, Ilia Cherniavskii, Yinghai Lu, Raghuraman Krishnamoorthi, Ansha Yu, Volodymyr Kondratenko, Stephanie Pereira, Xianjie Chen, Wenlin Chen, Vijay Rao, Bill Jia, Liang Xiong, and Misha Smelyanskiy. 2019. Deep Learning Recommendation Model for Personalization and Recommendation Systems. *CoRR* abs/1906.00091 (2019). https://arxiv.org/abs/1906.00091

[7] Francesco Sanna Passino, Lucas Maystre, Dmitrii Moor, Ashton Anderson, and Mounia Lalmas. 2021. Where To Next? A Dynamic Model of User Preferences. *WWW* 21 (2021), 19–23.

[8] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. 2017. Data Management Challenges in Production Machine Learning. In *Proceedings of the 2017 ACM International Conference on Management of Data* (Chicago, Illinois, USA) *(SIGMOD '17)*. Association for Computing Machinery, New York, NY, USA, 1723–1726. https://doi.org/10.1145/3035918.3054782

[9] Deepjyoti Roy and Mala Dutta. 2022. A systematic review and research perspective on recommender systems. *Journal of Big Data* 9, 1 (2022), 59.

[10] Markus Schedl and David Hauger. 2015. Tailoring music recommendations to users by considering diversity, mainstreaminess, and novelty. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval*. 947–950.

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

[12] Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. 2021. Contrastive learning for cold-start recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5382–5390.

[13] Ali Yürekli, Cihan Kaleli, and Alper Bilge. 2021. Alleviating the cold-start playlist continuation in music recommendation using latent semantic indexing. *International Journal of Multimedia Information Retrieval* 10, 3 (2021), 185–198.

[14] Sheng Zhang, Weihong Wang, James Ford, and Fillia Makedon. 2006. Learning from incomplete ratings using non-negative matrix factorization. In *Proceedings of the 2006 SIAM international conference on data mining*. SIAM, 549–553.

[15] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. 2012. Auralist: introducing serendipity into music recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining*. 13–22.

# AUTHOR BIOGRAPHIES

**Ghazal Fazelnia** is a Senior Research Scientist at Spotify. She obtained her Ph.D. in Electrical Engineering from Columbia University in 2019. Her research focuses on machine learning and recommender systems, and she has published over 30 papers at venues such as ICML, NeurIPS, WSDM, The Web Conference, and CIKM.

**Sanket Gupta** is a Senior Machine Learning Engineer at Spotify. He obtained his M.Sc. in Electrical Engineering from Columbia University in 2015.

**Claire Keum** is a Senior Software Engineer at Spotify where she build recommender systems in music space.

**Mark Koh** is a Senior Machine Learning Engineer at Spotify. He obtained his B.Sc. in Computer Science and Software Engineering from Drexel University in 2016.

**Dr. Timothy Heath** is a Staff Machine Learning Engineer at Spotify. He obtained a PhD in Number Theory from Columbia University in 2015.

**Guillermo Carrasco Hernández** is a Senior Machine Learning Engineer at Spotify. He completed his Computer Science Engineering studies in 2013 at the Polytechnical University of Catalonia, Spain.

**Stephen Xie** is a Senior Engineer at Spotify. He obtained his B.Sc. in Mathematics from University of Waterloo.

**Nandini Singh** is a Senior Data Scientist at Spotify. She completed her Bachelor of Technology in Information Technology at SRM University in 2014.

**Ian Anderson** was a Staff Machine Learning Engineer at Spotify. He obtained his Ph.D. in Physics from Johns Hopkins University in 2015.

**Mounia Lalmas** is a Senior Director of Research and Head of Tech Research in Personalization at Spotify. Mounia regularly serves on senior program committees for major conferences like WSDM, KDD, and SIGIR. She has co-chaired SIGIR 2015, WWW 2018, WSDM 2020, and CIKM 2023, and has authored over 250 papers.

**Petter Pehrson Skidén** is a Senior Product Manager at Spotify. He completed his Master of Science in Applied Physics and Electrical Engineering in 2013 at the University of Linköping, Sweden.

**Maya Hristakeva** is a Senior Machine Learning Engineering Manager at Spotify. She completed her Master of Science in Computer Science in 2009 from University of California, Santa Cruz.