Playlist Search Reinvented: LLMs Behind the Curtain

Geetha S. Aluri
Amazon Music Search
California, USA
aluriga@amazon.com

Siddharth Sharma
Amazon Music Search
California, USA
eshasidd@amazon.com

Tarun Sharma
Amazon Music Search
California, USA
tarunsh@amazon.com

Joaquin Delgado
Amazon Music Search
California, USA
dejoaqui@amazon.com

ABSTRACT

Improving search functionality poses challenges such as data scarcity for model training, metadata enrichment for comprehensive document indexing, and the labor-intensive manual annotation for evaluation. Traditionally, iterative methods relying on human annotators and customer feedback have been used. However, recent advancements in Large Language Models (LLMs) offer new solutions. This paper focuses on applying LLMs to playlist search. Leveraging LLMs' contextual understanding and generative capabilities automates metadata enrichment, reducing manual efforts and expediting training. LLMs also address data scarcity by generating synthetic training data and serve as scalable judges for evaluation, enhancing search performance assessment. We demonstrate how these innovations enhance playlist search, overcoming traditional limitations to improve search result accuracy and relevance.

KEYWORDS

Retrieval, Semantic Search, MLOps, LLM augmentation

ACM Reference Format:
Geetha S. Aluri, Siddharth Sharma, Tarun Sharma, and Joaquin Delgado. 2024. Playlist Search Reinvented: LLMs Behind the Curtain. In *18th ACM Conference on Recommender Systems (RecSys '24), October 14–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3640457.3688047

1 INTRODUCTION

As a music service, we aim to improve user experience by enhancing retrieval mechanisms. Traditional lexical search methods use bag-of-words indexing for relevance scoring, but struggle with natural language queries. To address this limitation, we recently demonstrated successful utilization of transformer based bi-encoder models for podcast semantic search [2], embedding and indexing document sentences using an Approximate Nearest Neighbor (ANN) search algorithm [3]. This method enables more accurate retrieval by leveraging semantic similarity between queries and document content.

RecSys '24, October 14–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0505-2/24/10.
https://doi.org/10.1145/3640457.3688047

Our music service features Editorial Playlists (EPLs) curated by experts and Community Playlists (CPLs) created by users. EPLs offer polished, genre-spanning music, while CPLs cover niche interests and trends. While EPLs provide rich data in the form of expert curated descriptions for representation, CPLs lack metadata, posing challenges for semantic search model development. Developing semantic search models for playlists relies on query-clicked playlist pairs mined from search behavioral logs. However, due to the lexical matching nature of current search systems, these logs may lack pairs suitable for successful semantic search. Additionally, manual human annotation for model evaluation is challenging because annotators must sift through the entire catalog to find all relevant matches, making the process labor-intensive and time-consuming.

To overcome these limitations, we've integrated LLMs into our ML pipeline for model development, training, and evaluation. In upcoming sections, we review the ML model development landscape and detail our recent architectural enhancements leveraging LLMs. Through a fusion of advanced retrieval methods and LLM support, our goal is to enhance playlist search accuracy and relevance, facilitating seamless discovery of audio content tailored to customers' preferences and interests.

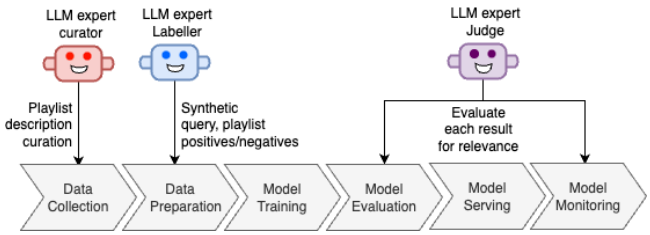

Figure 1: Enhanced ML Pipeline with LLM Assistance

2 APPROACH

Figure 1 illustrates our approach. Recent research underscores the effectiveness of Large Language Models (LLMs) for content enrichment[1, 5]. In the context of Community Playlists (CPLs), which often lack the rich metadata needed for embeddings, leveraging LLMs offline offers an optimal solution to this challenge. By employing the LLM expert curator to craft detailed descriptions for CPLs, we provide the first 15 tracks to capture themes, genres, activities, eras, and artists, resulting in highly effective descriptions. Our approach integrates off-the-shelf models with over 100 billion

parameters and fine-tuned Flan-T5-XL models with 3 billion parameters, fine-tuned using both CPL and editorial playlist tracks and description data. Surprisingly, even smaller fine-tuned models perform exceptionally well, suggesting further potential for improvement. Consequently, CPLs now offer richer descriptions and user-provided titles, significantly enhancing their representation and usability.

Transitioning from metadata enrichment to training data for semantic models, our focus shifts towards refining the datasets used for training. Recent research has highlighted the effectiveness of LLMs in synthetic data generation[7, 8]. Drawing from past experiences with semantic search models, we recognize the paramount importance of high-quality and diverse data sources. Fine-tuning bi-encoder models with such datasets yields substantial improvements compared to selecting state-of-the-art models with simplistic training datasets. Our approach encompasses a diverse array of data sources: <query, clicked-playlist> pairs mined from search behavioral logs provide invaluable insights despite varying quality. Additionally, synthetic pairs generated from metadata offer controlled datasets, leveraging the rich titles and descriptions of both EPLs and CPLs to bolster model robustness. Completing this trio is synthetic data generated from advanced generative models, forming a comprehensive training framework. This approach involves synthesizing queries from all playlist metadata, paired with playlists to create training data. Subsequently, smaller models and Amazon Titan text embeddings are utilized to generate <query, playlist> pairs from synthetic and low-performing queries from search logs. These pairs are scored for relevance using the LLM expert labeller, effectively creating a mix of positive and negative examples. This multi-faceted approach to data collection ensures our embedding models are trained on a wide variety of rich and high-quality data, significantly enhancing their ability to handle real-world scenarios.

To fine-tune semantic models effectively, we employ Parameter-Efficient Fine-Tuning (PEFT) techniques[9], building upon the training data sources outlined earlier. Techniques like Low-Rank Adaptation (LoRA)[6] are instrumental, introducing trainable low-rank matrices to transformer layers, significantly reducing parameter counts compared to full fine-tuning. This approach preserves the original pre-trained model weights, mitigating catastrophic forgetting, while allowing for task specialization through targeted fine-tuning. The compact parameter footprint of LoRA leads to substantial computational savings, enabling rapid experimentation cycles for iterative model refinement without extensive retraining overhead. These fine-tuned models form the foundational basis for generating embeddings for both query and playlist metadata.

As we move from fine-tuning semantic models to the crucial phase of evaluation, one of the prominent methods involves utilizing part of the training datasets. However, with the emergence of new traffic patterns and evolving query structures, traditional evaluation approaches may fall short, necessitating a more dynamic solution. In response to this challenge, Large Language Models (LLMs) have emerged as effective tools for evaluation, capable of closely approximating human judgment[4, 10] when provided with appropriate prompts. By bootstrapping the LLM expert judge with a smaller sample of human annotation and prompt engineering, we can harness their judgment capabilities to streamline the evaluation process. This not only facilitates model iteration but also serves as a valuable asset for daily monitoring of production models, enabling us to promptly identify any signs of performance degradation.

3 RESULTS

Our models are evaluated on three unique datasets, each containing ground truth query and playlist pairs: the benchmark dataset, an internally manually annotated set; the SEO dataset, sourced from the Search Engine Optimization (SEO) team to drive search engine traffic; and a paraphrasing dataset, generated using an LLM for given playlist titles and descriptions. The Table 1 illustrates the performance enhancement achieved by incorporating LLMs into the machine learning pipeline.

Table 1: Recall@K Improvement on Evaluation Datasets

Dataset	Recall@1	Recall@3	Recall@10
Benchmark Dataset	+17.1%	+16.6%	+12.9%
SEO Dataset	+27.1%	+21.1%	+16.0%
Paraphrasing Dataset	+32.1%	+26.0%	+18.2%

4 CHALLENGES

Despite the advancements in LLMs, ensuring the quality and diversity of the data used for content enrichment remains a challenge. Since this data is used solely for generating representations and isn't customer-facing, the risk is low. Nevertheless, human oversight is still necessary to maintain data quality. Scaling up LLM-based approaches for content enrichment and synthetic data generation to handle large datasets efficiently is also difficult. It requires robust infrastructure and optimization strategies to process vast amounts of data while maintaining model performance. Assessing the effectiveness and performance of LLM-based methods, especially when used as judges, presents challenges in terms of evaluation metrics and interpretability. Developing reliable evaluation frameworks that capture the nuances of LLM-generated content and training data, as well as interpreting the decisions made by LLM-based judges, remains an ongoing challenge in the field.

5 CONCLUSION

In conclusion, the integration of LLMs into the retrieval pipeline presents a significant advancement in enhancing search functionality and user experience. By leveraging LLMs for content enrichment, synthesizing training data, and serving as judges for evaluation, we have addressed key challenges such as data scarcity, metadata enrichment, and evaluation complexity. However, challenges persist in ensuring data quality, scalability, and interpretability of LLM-based methods. Moving forward, continued research and development efforts are crucial to refine LLM-based techniques, optimize infrastructure, and establish robust evaluation frameworks. Despite these challenges, the innovative application of LLMs holds promise for further elevating the accuracy, relevance, and efficiency of playlist search results, ultimately enhancing the music discovery experience for users.

BIO

Geetha S. Aluri is an Applied Scientist at Amazon Music where she develops at-scale machine learning models for Music Search. With prior experience as a Data Scientist at Walmart Labs, she has a proven track record in developing robust Trust & Safety ML solutions for catalog content. She earned her Ph.D. in Electrical and Computer Engineering from George Mason University.

Siddharth Sharma is a Machine Learning Engineer at Amazon Music. He earned his Master's degree from North Carolina State University, Raleigh, NC. Siddharth's work focuses on search, retrieval and ranking models.

Tarun Sharma is an Applied Science Manager leading Search Relevance for Amazon Music. He has more than ten years of experience building and leading science and engineering teams to launch internet scale machine learning solutions and high performant systems in Ads and Recommendation domains at Amazon and previously at Microsoft. He holds a Masters in Language Technologies from Carnegie Mellon University, Pittsburgh, USA and a Bachelors in Computer Science from Indian Institute of Technology, Varanasi, India.

Joaquin Delgado is currently serving as a Sr. Software Development Manager at Amazon, leading Amazon Music Search. Previously he held Director positions at Groupon, Verizon and Intel and was the CTO of AdBrite and Co-founder/CTO of LendingClub. He also worked at Yahoo! and Oracle. His expertise lies in distributed systems, advertising technology, machine learning, recommender systems, and information retrieval. He holds a Ph.D. in Computer Science and Artificial Intelligence from Nagoya Institute of Technology, Japan and a Computer Engineering degree from Universidad Simon Bolivar, Venezuela.

REFERENCES

[1] Saurabh Agrawal, John Trenkle, and Jaya Kawale. 2023. Beyond Labels: Leveraging Deep Learning and LLMs for Content Metadata. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore, Singapore) *(RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 1. https://doi.org/10.1145/3604915.3608883

[2] Geetha Sai Aluri, Paul Greyson, and Joaquin Delgado. 2023. Optimizing Podcast Discovery: Unveiling Amazon Music's Retrieval and Ranking Framework. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore, Singapore) *(RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 1036–1038. https://doi.org/10.1145/3604915.3610240

[3] Sunil Arya, David M Mount, Nathan S Netanyahu, Ruth Silverman, and Angela Y Wu. 1998. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)* 45, 6 (1998), 891–923.

[4] Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 15607–15631. https://doi.org/10.18653/v1/2023.acl-long.870

[5] SeungHeon Doh, Minhee Lee, Dasaem Jeong, and Juhan Nam. 2024. Enriching Music Descriptions with A Finetuned-LLM and Metadata for Text-to-Music Retrieval. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 826–830. https://doi.org/10.1109/ICASSP48485.2024.10446380

[6] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL]

[7] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations. arXiv:2310.07849 [cs.CL]

[8] Parth Patwa, Simone Filice, Zhiyu Chen, Giuseppe Castellucci, Oleg Rokhlenko, and Shervin Malmasi. 2024. Enhancing Low-Resource LLMs Classification with PEFT and Synthetic Data. arXiv:2404.02422 [cs.CL]

[9] Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment. arXiv:2312.12148 [cs.CL]

[10] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL]