Verifying LLM-Driven Music Recommendation Systems

Orpaz Goldstein
goldsteo@Amazon.com
Amazon Music
Culver City, CA, USA

Roman Nazarov
rnazarov@Amazon.com
Amazon Music
San Fransisco, CA, USA

ABSTRACT

We propose a framework to evaluate Large Language Model (LLM)-driven music recommendations by aligning their outputs with structured embedding spaces that combine text and acoustic representations. Our approach introduces intrinsic evaluation tests targeting analogy reasoning, genre consistency, thematic relevance, and attribute adherence. We present empirical results across multiple embedding models, demonstrating the framework's effectiveness in capturing nuanced musical information. Through extensive experimentation we show that our method shows significant promise in recommendation relevance evaluation. Our framework offers a systematic, interpretable pathway for verifying LLM-based recommendation systems.

KEYWORDS

LLM, Generative AI, Amazon, Music, Recommendations

ACM Reference Format:
Orpaz Goldstein and Roman Nazarov. 2025. Verifying LLM-Driven Music Recommendation Systems. In *Proceedings of LLM4ECommerce Workshop at the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (LLM4ECommerce Workshop at KDD '25).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recommender systems enhanced by LLMs leverage rich semantic knowledge to personalize suggestions and provide contextual explanations. In music recommendation, while LLMs can handle nuanced queries, verifying their recommendations' relevance and grounding remains challenging. Conventional evaluation metrics fall short in capturing musical semantics, particularly when assessing contextual and thematic alignment.

Our primary objective is evaluating recommendation quality within the constraints of an LLM's music understanding capabilities. We aim to enhance user experience in music streaming services by providing on-demand recommendations that adhere to user-specified themes. For instance, when processing a query for "90's grunge," the system should generate a cohesive group of tracks that cluster together both acoustically and semantically. Furthermore, the prompt should accurately represent the contextual information corresponding to the recommended tracks.

LLM4ECommerce Workshop at KDD '25, August 4, 2025, Toronto, ON, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

To address these challenges, we propose an integrated framework that combines LLM recommendation generation with structured embedding-based verification. Our approach utilizes both textual and acoustic spaces to assess recommendation quality intrinsically. We demonstrate the framework's effectiveness through comprehensive testing of musical knowledge baselines and present a generalizable definition of "quality" that enables custom measurement integration while maintaining framework compatibility.

Motivation for E-commerce. In a subscription environment such as Amazon Music, *early positive interactions* are the strongest predictor of long-term retention; users who are exposed and engage with music recommendations in their first week churn less on average.[1] High-fidelity recommendations also open incremental revenue streams: *playlist-linked merch*, ticket up-sell, smart-speaker cross-sell, and ad-supported storefront dwell time all improve when a user stays inside a music session longer. Therefore, a verification framework that can *prove* an LLM understands nuanced musical requests is not merely academic—it directly supports satisfaction, retention, and downstream monetization.

2 RELATED WORK

2.1 Music Representation Learning

Transformer-based models like MuLan [3], CLAP [2], and MusiLingo [1] align music audio and text embeddings for retrieval tasks. MARBLE [8] provides comprehensive benchmarks for music representations across various tasks. ChatMusician [7] demonstrates successful adaptation of LLaMA2 for symbolic music, achieving superior performance compared to GPT-3.5 in music theory tasks.

2.2 LLM Evaluation in Recommendation Systems

Recent work in ontology-based triplet evaluations [6] and lyric-based keyword extraction with KeYric [4] highlights specific challenges in LLM music knowledge representation. Our framework builds upon these foundations while extending their capabilities. Furthermore, systematic behavioral testing approaches like CheckList [5] inform our evaluation design.

2.3 Embedding Space Analysis

Previous research in embedding space analysis for recommendation systems has focused primarily on single-modality representations. Our work extends this by introducing a multi-modal approach that combines acoustic and semantic features in a unified embedding space, enabling more nuanced recommendation evaluation.

[1] Internal aggregated metric, Jan–Dec 2024.

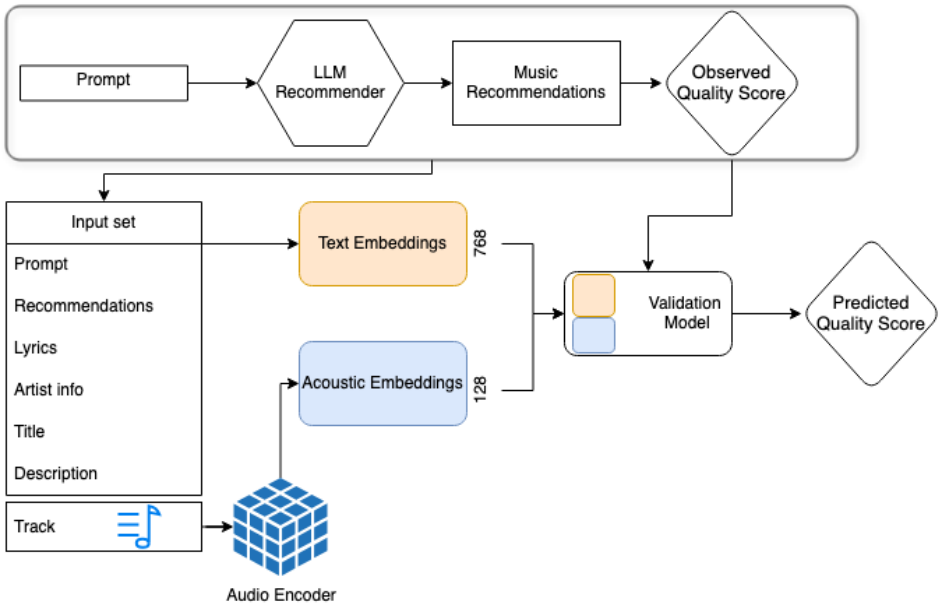


Figure 1: Architectural diagram of the music recommendation evaluation framework. The framework evaluates LLM-generated music recommendations by encoding both prompt and track-related metadata (lyrics, artist info, title, description) into a 768-dimensional text embedding, and tracks into 128-dimensional acoustic embeddings via an audio encoder. These embeddings are concatenated and fed into a learned validation model that predicts recommendation quality. Observed user behavior (e.g., playbacks, saves) provides a ground truth quality score, enabling supervised training and alignment between semantic/music-based representations and real-world engagement.

3 PROPOSED FRAMEWORK

3.1 Architecture Overview

Our framework evaluates LLM-generated recommendations via joint text-acoustic embedding spaces. User prompts and recommendations are encoded into this space using a novel multi-modal transformer architecture, enabling direct semantic comparisons. Figure 1 illustrates the system architecture.

3.2 Embedding Space Construction

We construct our embedding space using a combination of:

- Acoustic features: 128-dimensional MP3 encoded vectors
- Textual features: 768-dimensional contextual embeddings from large language models
- Metadata features: Genre hierarchies, release information, and artist relationships

4 BASELINE EVALUATION STRATEGY

4.1 Intrinsic Testing Framework

We implement comprehensive intrinsic tests targeting four key dimensions:

4.1.1 Movement/Era Score. Measures the embeddings' capability to capture historical and stylistic consistency of music across distinct eras and movements. For example, a query such as *1980s synth-pop* should closely match embeddings of artists like Depeche Mode or New Order. We utilize cosine similarity metrics across over 100 carefully selected era-specific examples, aggregating similarity scores to evaluate how accurately embeddings reflect historical musical characteristics. We quantify historical and stylistic consistency using:

$$E_{score} = \frac{1}{N} \sum_{i=1}^{N} \cos(v_i, c_e) \cdot w_e \tag{1}$$

where v_i is the embedding vector of track i, c_e is the era centroid, and w_e is an era-specific weight. N is the number of samples.

4.1.2 Analogical Reasoning. Probes relational knowledge through genre, era, and artist analogies. A successful analogy test might look like *Nirvana : Grunge :: Tupac : Hip-hop*, confirming the embedding's relational coherence. This test is executed on an extensive set of 100 analogies, computing accuracy as the proportion of correctly preserved analogical relationships in the embedding space.

We evaluate relational knowledge through:

$$A_{score} = \frac{\sum_{i=1}^{M} \mathbb{1}(||v_b - v_a + v_c - v_d|| < \epsilon)}{M} \tag{2}$$

where v_a, v_b, v_c, v_d are embedding vectors in the analogy $(a : b :: c : d)$.

4.1.3 Subgenre Clustering. Assesses hierarchical consistency via clustering metrics. Tracks within subgenres such as *progressive rock* and *alternative rock* should distinctly cluster yet remain close within broader genres. We compute silhouette scores using 100+ labeled examples per subgenre, quantifying embedding effectiveness in capturing nuanced genre hierarchies. We employ hierarchical

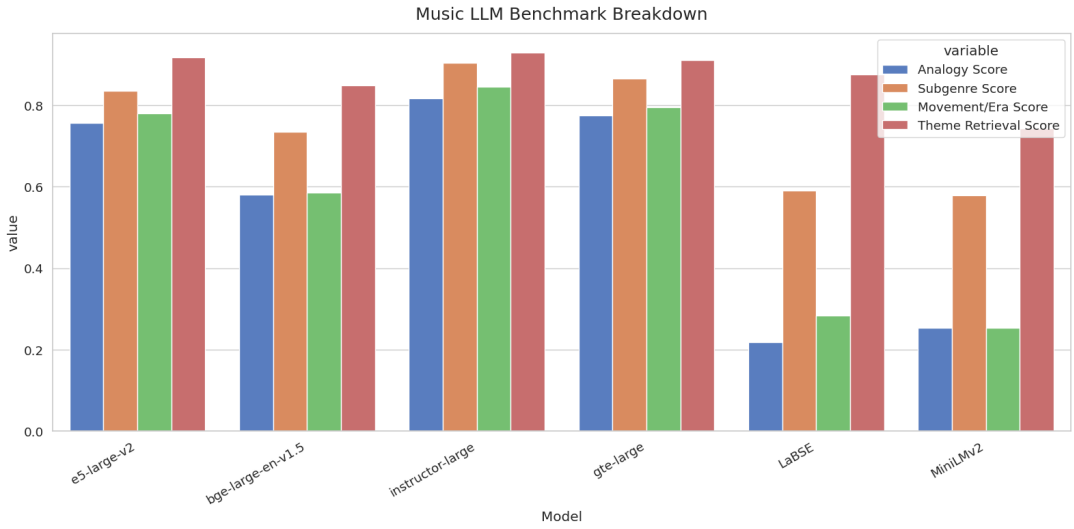
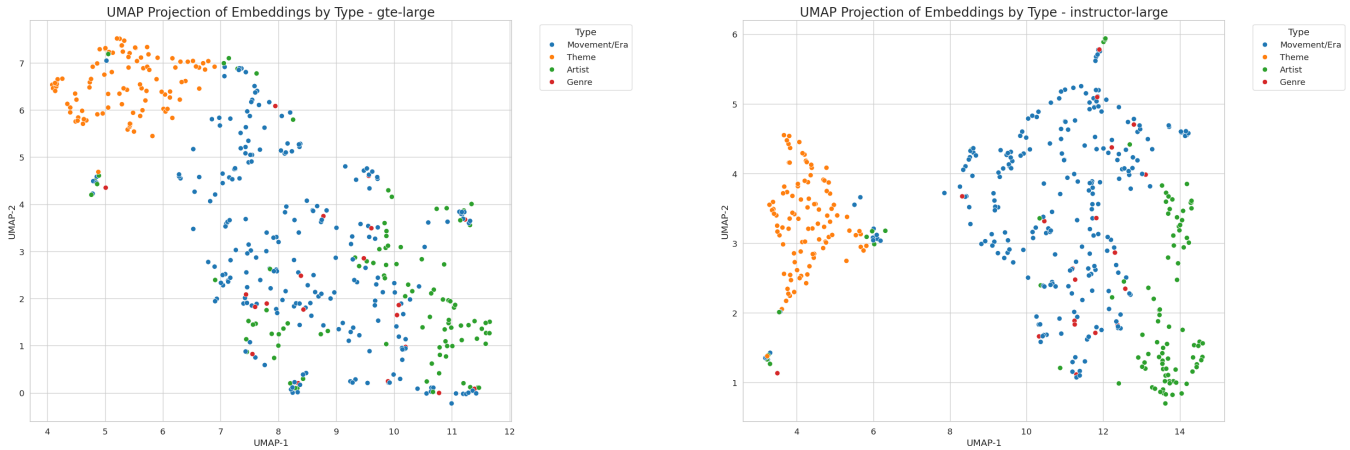


Figure 2: Comparison of each model across our musical knowledge tests. This figure benchmarks six embedding models on four intrinsic music understanding tasks: analogical reasoning, subgenre clustering, movement/era alignment, and theme retrieval. Instructor-large achieves the highest overall performance, followed closely by GTE-large. These results inform our selection of embedding backbones for downstream evaluation.



(a) gte-large **(b) Instructor-large**

Figure 3: Cluster visualizations across the two leading models in embedding music tests. UMAP projections show labeled clusters for Instructor-large and GTE-large embeddings. Instructor-large forms tighter, more distinct clusters, reflecting stronger genre separation. These visualizations support the quantitative differences shown in 2

consistency metrics:

$$E_{score} = \frac{1}{N}\sum_{i=1}^{N}\cos(s_i, g_e) \tag{3}$$

where s_i is the embedding vector of subgenre i, g_e is the genre centroid. N is the number of samples.

4.1.4 Theme Retrieval Score. Verifies embeddings' capability in capturing abstract musical concepts and thematic constraints provided by users. For instance, prompts like *songs about social justice* should retrieve tracks from artists like Bob Dylan or Public Enemy, demonstrating theme-based retrieval accuracy. Over 100 thematic examples, we evaluate retrieval success using cosine similarity and binary adherence to user-defined constraints, providing comprehensive verification of thematic and constraint-based embedding performance. We verify thematic alignment using:

$$T_{score} = \lambda \cdot \cos(v_t, c_t) + (1-\lambda)\cdot \text{constraint}_{\text{adherence}} \tag{4}$$

where v_t is the track embedding and c_t is the theme centroid. Evaluation across across 100 distinct themes. $\text{constraint}_{\text{adherence}}$ is a

continuous score for fulfilling explicit constraints. This measures how well the recommendation respects explicit user constraints, such as: "No vocals" → only instrumental tracks should be included. "Only female vocals" → all recommendations must have female vocalists. λ is between [0,1] and balances the importance of semantic alignment vs. constraint adherence. $\lambda = 1$ prioritize semantic similarity only. $\lambda = 0$ prioritize strict constraint adherence only.

4.2 Model Benchmarking

We evaluated six state-of-the-art embedding models:

- E5-large (768d)
- GTE-large (1024d)
- bge-large-en (1024d)
- Instructor-large (768d)
- MiniLMv2 (384d)
- LaBSE (768d)

Performance metrics are shown in Figure 2, with GTE-large and Instructor-large demonstrating superior performance across all evaluation dimensions.

4.3 Quantitative Results

Table 1 reports the mean metric values plus the Overall score for six embedding models.

Table 1: Intrinsic benchmark results ($N = 100$ per metric). Bold = best.

Model	Analogy	Subgen.	Era	Theme	Overall
Instructor-large	**0.82**	**0.90**	**0.85**	**0.93**	**0.87**
GTE-large	0.77	0.87	0.79	0.91	0.84
E5-large-v2	0.76	0.84	0.78	0.92	0.82
BGE-large-en-v1.5	0.58	0.74	0.59	0.85	0.69
LaBSE	0.22	0.59	0.28	0.88	0.49
MiniLMv2	0.25	0.58	0.25	0.74	0.46

Discussion. Instructor-large leads on all four metrics, highlighting its strong grasp of both lyrical semantics and acoustic nuance. Its superior Theme score (+0.02 over GTE) matches qualitative findings: prompts like *counter-culture protest songs* return Bob Dylan and Public Enemy without violating instrumentation filters. UMAP plots (Figure 3) show Instructor's embeddings separating *shoegaze* from *stoner rock* more cleanly than GTE, corroborating the silhouette gain.

4.4 LLM Inference Pipeline

The production recommender uses an internal instruction-tuned LLM. A template inserts the user request into a system message:

> `{SYS}`: You are a playlist assistant. Return 10 track titles with explicit artist names.
> `{USER}`: {prompt}

Decoding uses temperature 0.7, top-p 0.9, and max 128 tokens. Post-processing deduplicates results, verifies catalog availability, and canonicalises artist strings before evaluation.

5 LLM RECOMMENDATION EVALUATION STRATEGY

Each prompt–recommendation set is evaluated in production via user engagement signals. Let R denote the set of tracks surfaced for prompt P. We record main-screen plays, playlist saves, and subsequent plays from the saved set. A flexible engagement score is

$$\text{QualityScore}(P,R) = \alpha\frac{\text{Plays}(R)}{|R|} + \beta\frac{\text{Saves}(R)}{|R|} + \gamma\frac{\text{SavedPlays}(R)}{\max(1,\text{Saves}(R))},$$

where α, β, γ are tunable (currently left symbolic). This modular design lets teams swap in any engagement-based objective without altering the intrinsic-metric pipeline described above.

6 PRELIMINARY INSIGHTS

The best-performing embedding (Instructor-large) aligns clusters closely with known genre taxonomies in UMAP space, and its higher Theme score correlates with qualitatively better LLM outputs for prompts such as *counter-culture protest songs*. Failure cases—e.g., prompts requesting *instrumental jazz without drums*—highlight that acoustic constraints are occasionally violated if lyric semantics dominate similarity.

7 CONCLUSION AND FUTURE WORK

We introduced a multimodal, intrinsically grounded framework for verifying LLM-driven music recommendations. Beyond the present study, we identify three complementary research axes:

- **Music-specific text embeddings.** We plan to pre-train or domain-adapt an Instructor-style model on a large lyric–metadata corpus, yielding embeddings optimised for musical semantics rather than general language alone.
- **Closed-loop optimisation.** The predicted quality scores produced by our validation model can be fed back into the recommendation pipeline—either as a reward signal for RL-based LLM fine-tuning or as a feature in a learn-to-rank layer—enabling automatic improvement of both the LLM and the downstream ranking model.
- **Human-in-the-loop curation.** We will utilize professional music curators to label edge cases (e.g. niche genres, cross-cultural themes). These expert annotations will calibrate quality-score thresholds and guide active-learning updates to the evaluation model.

Together, these extensions aim to tighten the feedback loop between domain knowledge, model training, and user-centred evaluation, while paving the way for application to adjacent verticals such as podcasts and audiobooks.

REFERENCES

[1] Zihao Deng, Yinghao Ma, Yudong Liu, Rongchen Guo, Ge Zhang, Wenhu Chen, Wenhao Huang, and Emmanouil Benetos. 2023. Musilingo: Bridging music and text with pre-trained language models for music captioning and query response. *arXiv preprint arXiv:2309.08730* (2023).

[2] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[3] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. 2022. Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415* (2022).
[4] Xichu Ma, Varun Sharma, Min-Yen Kan, Wee Sun Lee, and Ye Wang. 2024. KeYric: Unsupervised Keywords Extraction and Expansion from Music for Coherent Lyrics Generation. *ACM Transactions on Multimedia Computing, Communications and Applications* 21, 1 (2024), 1–28.
[5] M.T. Ribeiro et al. 2020. Beyond Accuracy: Behavioral Testing of NLP models with CheckList. In *ACL*.
[6] Yannis Vasilakis, Rachel Bittner, and Johan Pauwels. 2024. Evaluation of pretrained language models on music understanding. *arXiv preprint arXiv:2409.11449* (2024).
[7] Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, et al. 2024. Chatmusician: Understanding and generating music intrinsically with llm. *arXiv preprint arXiv:2402.16153* (2024).
[8] Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Deng, et al. 2023. Marble: Music audio representation benchmark for universal evaluation. *Advances in Neural Information Processing Systems* 36 (2023), 39626–39647.

Received 15 March 2025