When the Catalog Has No Words: Verbalizing Listener Intent for Metadata-Free Music Retrieval

Anonymous ACL submission

Abstract

When audio catalogs lack metadata such as titles, artists, or genres, LLM-based recommenders lose textual item identity while sequential recommenders keep user state latent. We construct item-level pseudo-metadata from audio-embedding clusters, artist-community structure, and interaction aggregates, then prompt an LLM to verbalize a short next-track intent state from the user's recent descriptor window. To ensure that generated claims are grounded, we introduce a descriptor-consistency audit with a four-type inconsistency taxonomy validated by 600 human annotations. Acoustic-style inconsistency is the dominant error type, and controlled interventions show that unsupported claims degrade retrieval quality. Under the Yambda Global Temporal Split, intent-augmented retrieval yields a 16% relative NDCG@10 improvement over SASRec on Yambda-50M and a 14% improvement on Yambda-500M, with paired bootstrap tests confirming significance. Gains are strongest for cold-start users and remain significant after controlling for descriptor coverage and item popularity. A structured one-shot prompt with organic-weighted history produces the best descriptor alignment across three popular 7-8B LLMs. Complementary evaluation on Music4All-Onion shows consistent improvements when descriptors use native metadata.

1 Introduction

Industrial music retrieval must identify the next useful track from a user's chronological listening sequence. This choice is rarely item similarity alone the next track may preserve acoustic proximity, shift energy, match recent completion behavior, or reflect movement from algorithmic exposure to user-initiated discovery. Transformer recommenders such as SASRec (Kang and McAuley, 2018) and BERT4Rec (Sun et al., 2019) capture sequence dynamics well, but their user states remain latent and do not expose which attributes drive retrieval. LLM recommenders can express such attributes, but usually depend on titles, artists, genres, or item descriptions.

Yambda (Ploshkin et al., 2025) removes that textual scaffold. It provides large-scale listening logs, interaction signals, and audio embeddings, but no readable titles, artist names, or genre labels. We therefore separate item-level pseudo-metadata from user-level intent track descriptors summarize acoustic clusters, artist-community structure, and training-only interaction aggregates, while an LLM verbalizes a short *intent state* from the recent descriptor window. Its sentence-vector form is an *intent embedding*, or verbalized intent embedding (VIE) when used as a retrieval vector; Appendix Table 10 gives terminology details.

This verbalization is useful only if it is descriptor-consistent. We define descriptor consistency as support from the constructed descriptor window, including cluster labels and interaction aggregates, rather than metadata-verified factuality. The audit asks whether generated acoustic-style, energy, artist-community, and playback-pattern claims are supported by that window. Because the generated text is embedded for retrieval, unsupported claims can shift the query toward irrelevant items rather than merely producing poor explanations.

This framing yields three empirical questions: (1) can constructed descriptors support descriptor-consistent intent states, and does prompt design matter; (2) how consistent is generated in-

tent with descriptors, and is inconsistency associated with retrieval degradation; and (3) does intent language improve retrieval beyond collaborative and audio signals? We answer these questions with a metadata-free pipeline for descriptor construction and intent-state generation, descriptor coverage and label-validity analyses, a 600-example human-validated descriptor-consistency audit, and retrieval evidence on Yambda GTS with secondary Music4All-Onion evaluation under its source 60/20/20 split, which is not protocol-matched to Yambda. In the Yambda-50M cold-start analysis, the largest gains occur for low-history users, where collaborative state is weakest.

This work contributes four concrete results:

- a reproducible two-stage pipeline for item-level pseudo-metadata construction and user-level intent-state generation from audio-derived clusters and training-only interaction aggregates;
- a descriptor-consistency audit for music intent language, including a four-type inconsistency taxonomy and human validation;
- evidence that descriptor-inconsistency rate is associated with retrieval degradation while separating rule-based checks, LLM-based checks, and human labels;
- retrieval results on Yambda-50M and Yambda-500M under Yambda GTS, with secondary Music4All-Onion evaluation, showing that intent language complements collaborative and audio signals and helps most for low-history users.

2 Related Work

2.1 Sequential Recommendation

Sequential recommenders have progressed from recurrent session models such as GRU4Rec (Hidasi et al., 2016) to self-attentive SASRec (Kang and McAuley, 2018) and BERT4Rec (Sun et al., 2019). These models capture order and recency, but their learned states are not auditable accounts of which attributes drive retrieval. Text-enhanced variants such as UniSRec (Hou et al., 2022), VQ-Rec (Hou et al., 2023b), and RecFormer (Li et al., 2023) improve transfer and semantic representation using item text or learned codes.

2.2 LLMs for Recommendation

LLMs have been used for reranking, item-identifier generation, and ranking backbones, including LLMRank (Hou et al., 2024), BIGRec (Bao et al., 2024), and LlamaRec (Yue et al., 2023). These systems typically assume recognizable item identifiers, titles, or descriptions, or place the LLM directly in the ranking loop.

2.3 Audio Text Music Representations

Music-language models and resources such as MusicLM/MusicCaps (Agostinelli et al., 2023), MusicGen (Copet et al., 2023), LLark (Gardner et al., 2024), M^2UGen (Hussain et al., 2023), and CLAP (Wu et al., 2023; Elizalde et al., 2023) connect audio and text more directly. They motivate audio-text bridges and future automated descriptor labeling.

2.4 Audio-based Music Recommendation

Content-based music recommenders use audio representations for item similarity and cold-start retrieval (van den Oord et al., 2013; Castellon et al., 2021). These approaches typically use audio signals directly for retrieval or similarity ranking, rather than auditing whether generated recommendation language is grounded in the underlying audio-derived evidence.

2.5 Grounding, Faithfulness, and Probing

Source-grounded generation and factuality auditing have been studied in NLP (Maynez et al., 2020; Min et al., 2023), prompt ablations have been used to study reasoning behavior (Wei et al., 2022; Zhou et al., 2023), and probing methods examine what neural embeddings encode (Conneau et al., 2018; Hewitt and Manning, 2019; Belinkov, 2022). However, these tools remain underexplored for metadata-free music recommendation, where intent language may introduce unsupported claims about acoustic style, energy, playback behavior, or artist-community structure. Appendix B gives the expanded related-work discussion.

3 Approach

3.1 Datasets and Evaluation Settings

Table 1 summarizes the datasets used in our experiments. Yambda provides metadata-free listening logs, interaction signals, and audio embeddings; we evaluate the LISTEN+ target rather than the separate like/dislike task. Music4All-Onion is metadata-

Dataset	Users	Items	Listens	Setting
Yambda-50M	10,000	934,057	46,467,212	metadata-free
Yambda-500M	100,000	3,004,578	466,512,103	metadata-free
Music4All-Onion	119,140	109,269	252,984,396	metadata-rich

Table 1: Dataset statistics used in the experiments (Ploshkin et al., 2025; Moscati et al., 2022). Music4All-Onion retrieval uses the 79,072 items with complete features under the source 60/20/20 split.

Dataset	Item cov.	Inter. cov.	$\geq$1/5	$\geq$3/5	5/5	0/5
Yambda-50M	6.3%	70%	90%	75%	50%	3%
Yambda-500M	5%	72%	91%	76%	52%	3%

Table 2: Descriptor coverage. Item coverage is low because descriptors are built for stable head items, but interaction coverage and recent-window coverage are high under the heavy-tailed listening distribution.

Validation metric	Value
Acoustic clusters validated	20/20
Music-aware annotators	3
Representative items per cluster	20
Mean label-fit rating	4.1/5
Median label-fit rating	4.2/5
Fleiss' κ	0.65
Top-label agreement	80%
Ambiguous clusters	15%

Table 3: External validation of acoustic-cluster labels. Annotators rated whether prompt-facing labels fit representative tracks, artists, or proxy summaries on a 1-5 scale.

rich and serves as a secondary evaluation setting with native genres, tags, and Spotify audio features. Retrieval uses the 79,072-track complete-feature subset under the source paper's random 60/20/20 split with binarized interactions.

3.2 Descriptor Construction

Yambda anonymizes catalog identifiers, so each track descriptor $d(i)$ is built from audio embeddings, artist co-occurrence, and training-only interaction aggregates. We restrict descriptor construction to tracks with enough training-period LISTEN+ events for stable playback statistics, then cluster normalized 128-dimensional audio embeddings with $k = 20$. In Yambda-50M, this produces descriptors for 58,903 items ($\geq$50 training-period LISTEN+ events), 56,102 of which have audio embeddings. Although these items represent 6.3% of the 934,057 unique item IDs, the heavy-tailed interaction distribution means this head set covers most interaction mass and many recent-history windows (Table 2). Tracks without descriptors can still appear as retrieval candidates, so descriptor coverage affects query construction rather than restricting the candidate set. The resulting acoustic clusters are named through author inspection of artist communities and playback patterns (Appendix L). These labels are prompt-facing pseudo-metadata rather than native genre facts; external music-aware annotation validates their interpretability (Table 3). Appendix Table 22 tests sensitivity to semantic labels, neutral cluster IDs, random strings, and acoustic-style removal.

Artist co-occurrence provides a second structural signal. We compute artist-to-cluster affinities and group artists into communities, treating co-listening structure as a proxy for broad scene affinity. Playback descriptors are computed from training interactions only track duration, mean played ratio, and mean `is_organic`, discretized into compact bins so the LLM receives stable fields rather than noisy raw counts.

Descriptors follow a fixed template: *Acoustic style: [cluster label]. Artist group: [community]. Duration: [bucket]. Completion: [bucket]. Discovery: [bucket].* The template separates constructed evidence from user-level inference. Energy is a coarse derived cue grounded mainly in cluster-label vocabulary and recent completion behavior, not a native Yambda feature; energy inconsistency therefore measures descriptor consistency rather than external factual accuracy. For Music4All-Onion (Moscati et al., 2022), descriptors use native genres, tags, and Spotify audio features. Appendix C gives full construction details.

Two label-ablation controls test whether gains are merely wording effects. Replacing semantic labels with neutral cluster IDs keeps NDCG@10 at 0.0850, close to 0.0868 with semantic labels, while removing the acoustic-style field still reaches 0.0820 versus 0.0748 for SASRec (Appendix Table 22). Thus, author-written labels add signal but do not solely explain the retrieval gain.

3.3 Intent Generation and Fusion

Let $S_u = \langle i_1, \ldots, i_n \rangle$ denote the chronological interaction history for user u, and let $d(i_j)$ be the descriptor of item i_j. For a recent window $S_u^{(k)} = \langle i_{n-k+1}, \ldots, i_n \rangle$, an LLM generator g produces an intent state:

$$I_u = g(d(i_{n-k+1}), \ldots, d(i_n)).$$

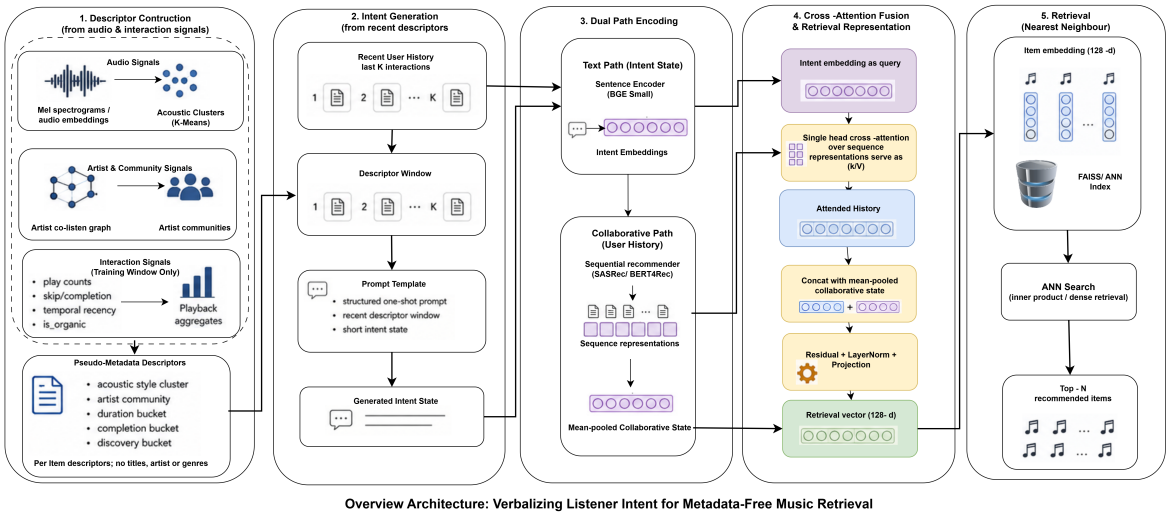

Figure 1: System architecture: descriptor construction from audio and interaction signals, LLM intent generation from recent descriptors, dual-path encoding, and cross-attention fusion for retrieval.

I_u is a short natural-language proxy for likely next-track attributes, not an item identifier. We instantiate g with Qwen3-8B (Qwen Team, 2025), LLaMA-3-8B-Instruct (Grattafiori et al., 2024), and Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), abbreviated Q3-8B, L3-8B, and M7B. The default prompt uses the last five descriptors, one exemplar, a 25-token budget, and organic-weighted history; Appendix D lists templates. We use 7-8B models because intent generation is the dominant per-user cost, roughly 94 ms on our modest A6000 setup (Appendix Table 24), while still being feasible for periodic batch generation.

Prompting varies along five dimensions: output form, exemplar count, history window, token budget, and interaction source. Organic-weighted history retains all interactions but gives higher sampling weight to user-initiated plays identified by `is_organic`. When recent items lack descriptors, the prompt includes only available descriptors, effectively shortening the window. The structured output format, with slots for acoustic style, energy, and listening-completion pattern, is the default because its claims can be audited against descriptor fields.

The intent state is encoded by a frozen sentence transformer to obtain v_I. We evaluate three lightweight encoders: BGE-small-en-v1.5 (Xiao et al., 2024), all-MiniLM-L6-v2 (Wang et al., 2020), and multilingual-e5-small (Wang et al., 2022); BGE-small yields the highest NMI and probe accuracy across all generators (Appendix Table 30) and is therefore the default. A collaborative encoder, usually SASRec, encodes S_u; BERT4Rec is used as a backbone check. Let $H \in \mathbb{R}^{n \times d}$ be the final-layer collaborative hidden states and $v_I \in \mathbb{R}^d$ the projected intent embedding. The cross-attention module forms a query $q = v_I W_Q$, keys $K = HW_K$, and values $V = HW_V$, then computes

$$\tilde{h} = \mathrm{softmax}\left(qK^\top / \sqrt{d_k}\right) V.$$

The attended vector is projected back to the collaborative dimension and combined with the mean-pooled collaborative state $\bar{h}$ using residual LayerNorm before projection to the 128-dimensional retrieval space. This lets verbalized intent selectively weight positions in the user sequence while keeping the LLM outside the ranking loop generation happens once per user context, and retrieval remains dense nearest-neighbor search over item vectors. Audio-direct, descriptor-only, and audio+text controls isolate the contribution of language, acoustic content, and fusion. Training and retrieval settings are in Appendix O.

3.4 Intent Quality Validation

Before descriptor-consistency and retrieval evaluation, we verify that generated intent states are fluent and descriptor-relevant. We evaluate 600 users stratified by history density: short histories with 3-5 unique tracks, medium histories with 6-20, and long histories with more than 20. Automatic metrics use GPT-2-large perplexity, BERTScore F1(Zhang, 2020) against recent descriptors, lexical diversity, and output length. Human valida-

Model	PPL ↓	BS F1 ↑	Flu. ↑	Rel. ↑	Spec. ↑	Hum. Acc. ↑	κ
Q3-8B	16.0	0.790	4.62	4.36	4.21	86.0	0.72
L3-8B	17.4	0.772	4.48	4.18	4.03	82.7	0.68
M7B	19.4	0.749	4.31	3.96	3.82	78.0	0.65

Table 4: Compact linguistic quality summary. PPL and BERTScore F1 are averaged over short, medium, and long history strata; G-Eval and human validation use the 600-example audit sample.

Factor	Best	Worst	ΔNDCG	η^2
Prompt style	P3	P1	+0.0032	0.42
Shot count	1	0	+0.0017	0.14
History window	5	3	+0.0020	0.09
Output budget	25	10	+0.0026	0.07
Interaction src	org-w	org	+0.0014	0.05

Table 5: Prompt ablation summary on Yambda-50M (Qwen3-8B). ΔNDCG is best minus worst within each factor; η^2 is from Appendix Table 16.

tion uses three annotators with binary descriptor-consistency and per-type labels, while a fixed gpt-oss-20B judge scores fluency, relevance, and specificity. Table 4 reports the compact model-level summary; Appendix Tables 11 and 12 give full stratified details.

History density is the strongest quality driver: longer histories yield lower perplexity and higher BERTScore for every generator. Model choice also matters: Qwen3-8B is strongest on automatic, judge-based, and human metrics, while Mistral-7B remains fluent but less specific and less descriptor-consistent. Since outputs stay below the 25-token budget, these differences reflect descriptor-consistent specificity rather than verbosity or grammaticality alone.

4 Prompting and Descriptor Consistency Audit

4.1 Prompt Ablation

We ablate prompt style, exemplar count, history window, output budget, and interaction source on Yambda-50M. The sweep first compares prompt families and shot counts under a fixed history setting, then varies history length, budget, and interaction source for the best prompt family. Table 5 gives the compact Qwen3-8B factor summary; Appendix Tables 15 and 16 report full per-model rows and ANOVA.

Prompt structure is the largest driver. The best configuration is P3 with one exemplar, a 5-track window, 25 tokens, and organic-weighted history. One-shot prompting improves over zero-shot, while three examples recover most but not all of the gain, suggesting that one concrete schema is enough to stabilize the short output format. Organic-weighted history outperforms organic-only history, indicating that recommendation-driven events carry useful context but should not dominate the descriptor window. Across ablation configurations, BERTScore F1 and NDCG@10 are strongly correlated ($r = 0.79$, $p < 0.001$), linking descriptor alignment with retrieval utility; because prompt style may drive both metrics, we treat this association as suggestive. Table 6 bridges the lightweight prompt-ablation setting with the final cross-attention retrieval architecture.

Stage	Configuration	NDCG@10
Baseline	SASRec only	0.0748
Prompt ablation	Best P3, one-shot, lightweight fusion	0.0825
Final fusion	Q3-8B+BGE + cross-attention	0.0865
Final audio+text	Q3-8B+BGE + audio+text	0.0872

Table 6: Bridge between prompt-ablation and final retrieval settings.

4.2 Descriptor-Consistency Audit

We audit descriptor consistency before retrieval: a generated intent is consistent when its acoustic-style, energy, artist-community, and playback-pattern claims are supported by the constructed descriptor window. Although prior NLP work uses *hallucination* for unsupported generated claims, we avoid it here because our evidence consists of constructed proxies rather than ground-truth metadata; *descriptor inconsistency* more precisely captures what we measure. Unsupported claims remain important because they can distort the retrieval vector. Acoustic-style inconsistency occurs when an intent mentions a style absent from recent cluster labels; energy inconsistency conflicts with coarse derived energy cues; artist-group inconsistency introduces an unsupported community; and playback-pattern inconsistency fabricates completion or discovery behavior. Acoustic-style and energy errors are detected by rule matching, while artist-group and playback-pattern claims are checked by a gpt-oss-20B consistency judge. Human annotation validates the combined pipeline, giving precision 0.88, recall 0.90, and $\kappa = 0.73$ against automated labels. Appendix F reports the protocol, shared-model-bias caveat, and retained agreement statistics.

Model	Stratum	Acoustic ↓	Energy ↓	Artist-grp ↓	Playback ↓	Any ↓
Q3-8B	Short	18.4	11.6	5.8	8.7	31.9
Q3-8B	Medium	12.2	7.4	3.9	5.2	22.5
Q3-8B	Long	7.8	4.9	2.1	3.6	13.7
L3-8B	Short	21.6	13.2	7.1	10.4	36.8
L3-8B	Medium	14.7	8.9	4.8	6.6	26.1
L3-8B	Long	9.6	5.8	2.9	4.3	16.8
M7B	Short	25.9	15.4	8.9	12.7	42.3
M7B	Medium	17.8	10.6	5.7	7.8	30.4
M7B	Long	11.9	7.2	3.8	5.5	20.6

Table 7: Descriptor-inconsistency rates (% of intent states with at least one unsupported claim) by type, model, and history stratum.

Input descriptor window (real Yambda-50M user 223300).

T1 Acoustic style: C10; duration: short; engagement: medium; discovery: organic-dominant.

T2 Acoustic style: C10; duration: short; engagement: medium; discovery: organic-dominant.

T3 Acoustic style: C10; duration: short; engagement: medium; discovery: organic-dominant.

T4 Acoustic style: C10; duration: short; engagement: medium; discovery: organic-dominant.

T5 Acoustic style: C10; duration: medium; engagement: low; discovery: organic-dominant.

P3 prompt filled with this window. *Given the five descriptors above, output one intent state using exactly: Acoustic style: [style]. Energy: [low/medium/high]. Engagement pattern: [pattern]. Do not add unsupported attributes.*

Generated intent. *Acoustic style: C10. Energy: high. Engagement pattern: organic-dominant listening with variety seeking.*

Annotation. Supported: C10 acoustic style and organic-dominant listening are grounded in the descriptor window. Inconsistent: high energy is unsupported because the recent engagement evidence is medium/low. "Variety seeking" is vague rather than counted unless it conflicts with a descriptor field.

Descriptor inconsistency is most frequent for sparse histories. Qwen3-8B has the lowest rate in every stratum, and acoustic-style inconsistency is the dominant error type for all generators. Because the annotation export does not retain per-type inter-annotator agreement, individual category rates in Table 7 should be interpreted as approximate; the aggregate inconsistency rate and model-level ordering are better supported by retained human agreement statistics. The pattern is not merely weaker English Mistral-7B remains fluent in Table 4 but introduces more unsupported descriptor claims. Higher descriptor-inconsistency rates are associated with weaker retrieval: the partial Spearman correlation controlling for log unique-track count is $\rho = -0.24$ with NDCG@10 ($p < 0.001$), and within-stratum correlations range from -0.14 to -0.27. Since history density affects both consistency and retrieval, we treat the result as associational and rely on convergence between rule checks, human labels, and retrieval associations rather than on the judge alone.

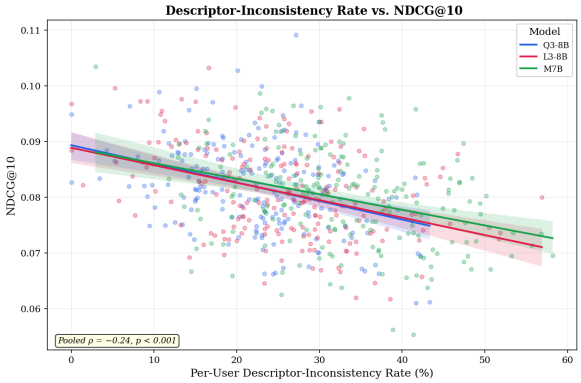

Figure 2: Per-user descriptor-inconsistency rate versus NDCG@10 with model-specific regression lines and 95% confidence bands.

5 Retrieval Experiments

5.1 Experimental Setup

We evaluate Yambda-50M, Yambda-500M (Ploshkin et al., 2025), and Music4All-Onion (Moscati et al., 2022). Yambda provides metadata-free listening events, interaction signals, and audio embeddings, and we use the official Global Temporal Split (GTS) under the LISTEN+ target. This matches our goal of sequential next-track retrieval from playback histories; we do not evaluate the separate LIKE task. Music4All-Onion serves as a secondary metadata-rich setting using the source paper's random 60/20/20 split with binarized interactions on the 79,072 complete-feature items. Because its split, interaction definition, and descriptor source differ from Yambda GTS, Music4All-Onion results are complementary rather than cross-dataset leaderboard comparisons.

The official Yambda GTS leaderboard (Ploshkin et al., 2025) reports results for Random, MostPop, DecayPop, ItemKNN, iALS, BPR, SANSA, and SASRec on LISTEN+; our reproduced SASRec exactly matches the official NDCG@10 (0.0748 on Yambda-50M, 0.0754 on Yambda-500M), confirming protocol alignment. Since ItemKNN (0.0781) and SASRec (0.0748) are the strongest official baselines and all others fall below 0.0407, our comparisons focus on these and on additional methods reported in Table 8. Main comparisons include SASRec, BERT4Rec, descriptor-concat, direct-audio, and intent-fusion variants; expanded baselines are reported in Appendix Table 17. LLMRank is included as a reference for LLM-based recommendation, but its metadata-free performance reflects its assumption of readable item identifiers

Dataset	Method	NDCG@10 ↑	Rec@10 ↑	Cov@10 ↑	Protocol
Yambda-50M					
	SASRec	0.0748 ± 0.0011	0.0325 ± 0.0008	0.0130	GTS
	BERT4Rec	0.0758 ± 0.0013	0.0332 ± 0.0009	0.0125	GTS
	Descriptor-concat	0.0808 ± 0.0012	0.0352 ± 0.0008	0.0175	GTS
	SASRec+audio	0.0805 ± 0.0012	0.0352 ± 0.0008	0.0225	GTS
	SASRec+Q3-8B+BGE	0.0868 ± 0.0009	0.0378 ± 0.0007	0.0145	GTS
	BERT4Rec+Q3-8B+BGE	0.0865 ± 0.0011	0.0374 ± 0.0008	0.0215	GTS
	SASRec+Q3-8B+audio+text	$\mathbf{0.0875 \pm 0.0010}$	$\mathbf{0.0390 \pm 0.0007}$	**0.0412**	GTS
Yambda-500M					
	SASRec	0.0754 ± 0.0010	0.0336 ± 0.0007	0.0347	GTS
	Descriptor-concat	0.0812 ± 0.0011	0.0358 ± 0.0008	0.0415	GTS
	SASRec+audio	0.0810 ± 0.0011	0.0365 ± 0.0008	0.0480	GTS
	SASRec+Q3-8B+BGE	$\mathbf{0.0860 \pm 0.0009}$	$\mathbf{0.0392 \pm 0.0007}$	**0.0554**	GTS
Music4All-Onion					
	SASRec	0.0592 ± 0.0014	0.0268 ± 0.0010	0.0190	60/20/20
	Descriptor-concat	0.0640 ± 0.0013	0.0295 ± 0.0009	0.0260	60/20/20
	SASRec+Q3-8B+BGE	$\mathbf{0.0680 \pm 0.0012}$	$\mathbf{0.0312 \pm 0.0008}$	**0.0365**	60/20/20

Table 8: Compact retrieval summary for matched runs. All intent-augmented rows use SASRec as the collaborative backbone unless otherwise noted. Yambda rows use the official GTS protocol; Music4All-Onion rows use the source paper's random 60/20/20 split with binarized interactions and are not protocol-matched to Yambda. Cross-protocol numbers should not be directly compared. Full metrics, additional baselines, source-paper references, and protocol notes are in Appendix Table 17.

rather than a controlled comparison with our intent-generation approach. Significance tests use 10,000 paired user-level bootstrap resamples over per-user NDCG@10, with paired permutation checks and Holm-Bonferroni correction where applicable. Full baseline, source-reference, hardware, and five-seed details are in Appendix Table 17 and Appendix O.

5.2 Retrieval Results

Under Yambda GTS, intent-augmented retrieval improves over matched baselines. On Yambda-50M, Q3-8B+BGE reaches 0.0868 NDCG@10 versus 0.0748 for SASRec ($\Delta = +0.0120$, 95% CI $[+0.0104, +0.0136]$, $p < 0.001$; Appendix Table 27). Audio+text reaches 0.0875, but its incremental gain over intent-only is small ($\Delta = +0.0007$) and the paired interval crosses zero, so we treat it as suggestive.

The descriptor-concat and SASRec+audio controls both improve over SASRec but remain below intent-augmented retrieval, indicating that the gain is not only due to exposing descriptors or adding acoustic content. Random text with the same fusion machinery stays close to SASRec (0.0753; Appendix Table 25), and neutral cluster IDs retain most of the semantic-label benefit while random strings weaken performance (Appendix Table 22). On Yambda-500M, intent-only reaches 0.0860 ($\Delta = +0.0106$, 95% CI $[+0.0092, +0.0120]$, $p < 0.001$). Applying the pathway to BERT4Rec also improves NDCG@10, confirming gains are not SASRec-specific.

On Music4All-Onion, Q3-8B+BGE improves over SASRec from 0.0592 to 0.0680 under the source 60/20/20 split. Descriptors here come from native metadata, but the result should be read as complementary evidence because the protocol differs from Yambda GTS.

Cold-start gains are strongest for low-history users: with fewer than five unique tracks, SASRec reaches 0.0450 NDCG@10 while Q3-8B+BGE reaches 0.0555 (Appendix Table 18). Gains decrease for longer histories and increase with descriptor coverage while remaining non-negative for long-tail items (Appendix Tables 19 and 20).

LLM	Enc.	NMI ↑	Purity ↑	Δ cos	Probe acc ↑
Q3-8B	BGE-sm	**0.386**	**0.552**	**0.337**	**0.497**
L3-8B	BGE-sm	0.359	0.528	0.314	0.471
M7B	BGE-sm	0.327	0.502	0.285	0.439
Random text	BGE-sm	0.074	0.216	0.005	0.193

Table 9: Compact intent-embedding probing results on Yambda. Majority next-cluster baseline is 0.184.

5.3 Intent Embedding Analysis

To understand why intent embeddings improve retrieval, we test whether they retain musical structure using k-means alignment, nearest-neighbor consistency, and a linear next-cluster probe. These probes test whether generated language places users with similar next-track attributes in nearby sentence-embedding regions.

Qwen3-8B with BGE-small gives the strongest structure, with NMI 0.386 and probe accuracy 0.497 versus a 0.184 majority baseline. The random-text control is far weaker, so the effect is not attributable to the sentence encoder alone. Removing acoustic-style labels reduces but does not eliminate the signal (Appendix Table 31), suggesting that label copying explains part but not all of the representation. Figure 3 visualizes the strongest configuration.

6 Discussion

The central value of verbalization is compression rather than label copying. Recent descriptors mix acoustic clusters, artist communities, completion, duration, and discovery mode; the LLM turns this heterogeneous evidence into a query-like representation aligned with collaborative state. This helps explain why descriptor concatenation and rule templates improve retrieval, but generated intent improves it further.

The audit's circularity deserves explicit treatment. Acoustic-style labels are author assigned,

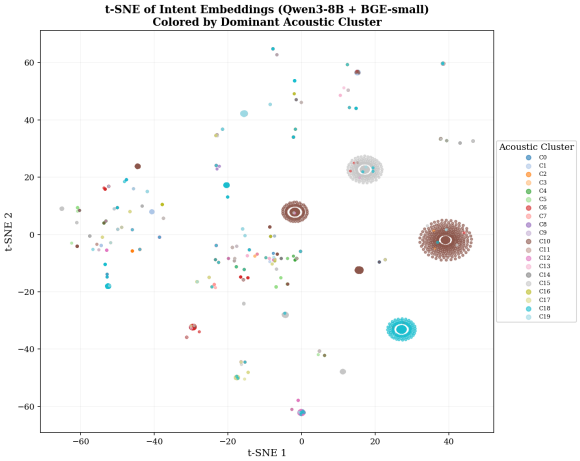

Figure 3: Intent-embedding t-SNE for the Qwen3-8B + BGE-small configuration, colored by dominant acoustic cluster.

so descriptor-consistency partly measures whether the generator reproduces them. External annotation supports the labels as interpretable (mean fit 4.1/5, $\kappa = 0.65$; Table 3), but they remain prompt-facing pseudo-metadata rather than native genre facts. Controls suggest label vocabulary is not the only signal neutral cluster IDs reach 0.0850 NDCG@10 versus 0.0868 with semantic labels, and removing acoustic style still reaches 0.0820 versus 0.0748 for SASRec (Appendix Table 22). Probing without acoustic-style labels reduces NMI from 0.386 to 0.217, still above the random-text baseline of 0.074 (Appendix Table 31), indicating that duration, completion, and discovery cues also contribute. Still, the audit mainly validates the full-label pipeline, and per-type acoustic-style or energy rates may partly reflect label reproduction.

Descriptor inconsistency matters because generated text becomes the retrieval query embedding, not merely an explanation. Unsupported acoustic or energy claims can change query geometry and move the user vector toward irrelevant regions. The observational partial correlation ($\rho = -0.24$, $p < 0.001$) is consistent with this mechanism but remains associational; within-stratum correlations also show that fluency alone is insufficient. Controlled interventions provide stronger mechanism-oriented evidence: style-corrupted intent falls to 0.0810 NDCG@10 and random-corrupted intent to 0.0750, while constrained generation stays near the faithful baseline (Appendix Table 23). We therefore reserve causal language for the controlled intervention.

Within the tested 7-8B range, model quality predicts descriptor-consistency and retrieval. Moving from Mistral-7B to Qwen3-8B reduces medium-history descriptor-inconsistency from 30.4% to 22.5% (Table 7) and raises NDCG@10 from 0.0812 to 0.0827 in the default prompt setting (Appendix Table 15). The monotonic ordering across perplexity, BERTScore, consistency rate, and retrieval suggests stronger deployable generators may help, though scaling beyond 8B remains open.

Operationally, intent generation is semantic preprocessing rather than ordinary feature projection. Sentence encoding and FAISS retrieval add only a few milliseconds per user, whereas Qwen3-8B generation dominates overhead at roughly 94 ms (Appendix Table 24); deployment would therefore require caching, batching, or asynchronous refresh. The design may transfer to e-commerce, video, or podcast recommendation, but each domain needs its own descriptor vocabulary and consistency taxonomy.

7 Conclusion

We introduced descriptor-consistent intent generation for metadata-free music retrieval, where constructed pseudo-metadata lets an LLM verbalize next-track intent without titles, artists, or genres. Structured prompts give the strongest descriptor alignment, acoustic-style inconsistency is the dominant error, and intent embeddings improve retrieval most for low-history users while complementing collaborative and audio signals. User-level paired bootstrap tests confirm gains over SASRec, descriptor-concat, and audio-only controls on Yambda; audio+text is best numerically but only a small suggestive improvement over intent-only. Future work should reduce inconsistency through constrained generation, automate cluster labeling with audio-text models such as CLAP, test intent fusion across more backbones and domains, and evaluate whether the observed consistency-retrieval trend across 7-8B models extends to larger generators.

Limitations

Descriptor construction. Audio-derived cluster labels are proxies for semantic music categories and require manual interpretation. Descriptor construction is restricted to the 58,903 items (of 934,057 total in Yambda-50M) that meet the minimum interaction threshold ($\geq$50 training-period LISTEN+ events) for meaningful playback statis-

tics. The remaining roughly 875K items are effectively long-tail tracks for which playback aggregates would be unreliable. Of the descriptor-covered set, 56,102 (95.2%) have audio embeddings; the 2,801 items without embeddings participate in collaborative signals but not in acoustic-style descriptor fields. We do not characterize whether uncovered long-tail items are missing at random or reflect systematic catalog sparsity, nor do we test whether lowering the interaction threshold with noisier descriptors would extend the method's reach.

Cluster labeling. The acoustic-style labels assigned to k-means clusters are author-assigned verbalizations. Descriptor-consistency is assessed against descriptors containing these labels, creating a circularity: the system's consistency partly measures whether the LLM reproduces the authors' naming choices. The external label-fit study in Table 3 and the cluster-ID ablation in Appendix Table 22 reduce this concern, but they do not make the labels native genre annotations or externally verified music facts.

Evaluation protocol. Yambda and Music4All-Onion are not protocol-identical. Our Music4All-Onion runs use the source paper's random 60/20/20 split with binarized interactions (Moscati et al., 2022), which differs from Yambda GTS in split method, interaction definition, and train/development/test proportions. We also evaluate only the Yambda LISTEN+ task, not the separate LIKE task. This limits claims about explicit preference prediction, but it matches our focus on next-listening retrieval from playback histories. Music4All-Onion source-paper references are protocol-matched within that dataset, but we do not run paired significance tests against source baselines because their per-user outputs are unavailable.

Human annotation. The human evaluation covers 600 intent states with per-type labels, which supports model-level validation but remains limited for fine-grained per-type and per-stratum analysis. The annotation export retained model-level and stratum-level agreement, plus agreement between human labels and the automated pipeline, but did not retain per-type inter-annotator agreement. We therefore do not make claims about per-type human reliability.

Open-source model and compute scope. This study intentionally focuses on open-source LLMs and embedding models that are practical to reproduce. We select 7–8B instruction-tuned generators and widely used Hugging Face sentence encoders, prioritizing models with strong public availability and high download-based adoption at the time of selection. This choice improves reproducibility but limits conclusions about closed-source systems, larger proprietary models, and less popular domain-specific encoders. All experiments run on a Xeon 40-core CPU server with 128GB RAM and a single NVIDIA RTX A6000 GPU with 48GB VRAM, which restricts evaluation to models that fit in memory under float16 inference. Intent generation already dominates per-user latency at roughly 94 ms (Appendix Table 24), so substantially larger models, such as 70B generators, would require distillation, quantization, model parallelism, or asynchronous batch refresh. The monotonic quality improvement from Mistral-7B through Qwen3-8B (Tables 4, 7, and 8) provides a within-band scaling signal, but we cannot confirm whether the trend extends to larger, closed-source, or domain-adapted generators. Similarly, larger embedding models or audio-text encoders such as CLAP may produce stronger intent embeddings, but they remain outside our resource budget.

Causality. The descriptor inconsistency retrieval relationship is correlational except for the controlled descriptor-inconsistency interventions in Appendix Table 23. The `is_organic` flag also gives only a partial view of recommendation exposure, so causal claims about organic versus algorithmic behavior are out of scope.

Generalization. The study focuses on music retrieval. The descriptor construction pipeline, descriptor-consistency taxonomy, and cold-start behavior may change in domains where item semantics, interaction signals, or catalog metadata are structured differently. Although we report five-seed variance and bootstrap confidence intervals for retrieval rows, we do not provide the same uncertainty analysis for every probing configuration.

Backbone scope. Intent augmentation is evaluated on SASRec and BERT4Rec. The cross-attention fusion design may interact differently with other sequential architectures such as FMLP-Rec or GRU4Rec. We do not claim that the observed gains transfer to arbitrary recommender backbones.

Ethical considerations

Our framework does not access personally identifiable information. The Yambda benchmark uses anonymized interaction logs, and intent states are derived from audio-cluster labels and playback-derived bins rather than user demographics. At the same time, descriptor construction can encode cultural assumptions if cluster labels are interpreted too strongly, and Music4All-Onion's user-generated tags may reflect biases in music categorization. We therefore frame generated intent as a retrieval signal rather than a direct statement of user identity or taste.

Open Source AI assistants were used ONLY for grammar polishing, synonym suggestions, and manuscript-style editing. They were not used for ideation, experiment design, coding, analysis, or generation of scientific claims.

References

Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., and Frank, C. 2023. MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Bao, Keqin and Zhang, Jizhi and Wang, Wenjie and Zhang, Yang and Yang, Zhengyi and Luo, Yanchen and Chen, Chong and Feng, Fuli and Tian, Qi 2024. A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems. In *ACM Transactions on Recommender Systems*.

Belinkov, Y. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Zhang* T ,Kishore V ,Wu F ,Q K. Weinberger ,Artzi Y , 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of ICLR*.

Castellon, R., Donahue, C., and Liang, P. 2021. Codified audio language modeling learns useful representations for music information retrieval. In *Proceedings of ISMIR 2021*.

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., and Defossez, A. 2023. Simple and controllable music generation. In *Proceedings of NIPS 2023*.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of ACL 2018*.

Deshpande, M. and Karypis, G. 2004. Item-based top-N recommendation algorithms. *ACM Transactions on Information Systems 2004*, 22(1):143–177.

Elizalde, B., Deshmukh, S., Al Ismail, M., and Wang, H. 2023. CLAP: Learning audio concepts from natural language supervision. In *Proceedings of ICASSP 2023*.

Hewitt, J. and Manning, C. D. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of NAACL 2019*.

Hidasi, B., Karatzoglou, A., Baltrunas, L., and Tikk, D. 2016. Session-based recommendations with recurrent neural networks. In *Proceedings of ICLR 2016*.

Hou, Y., He, Z., McAuley, J., and Zhao, W. X. 2022. Towards Universal Sequence Representation Learning for Recommender Systems. In *Proceedings of ACM SIGKDD 2022*.

Hou, Y., He, Z., McAuley, J., and Zhao, W. X. 2023b. Learning vector-quantized item representation for transferable sequential recommenders. In *Proceedings of The ACM Web Conference 2023*.

Hu, Y., Koren, Y., and Volinsky, C. 2008. Collaborative filtering for implicit feedback datasets. In *Proceedings of ICDM 2008*.

Gardner, J., Durand, S., Stoller, D., and Bittner, R. M. 2023. LLark: A Multimodal Instruction-Following Language Model for Music. *Proceedings of ICML 2024*.

Shansong Liu, Atin Sakkeer Hussain, Qilong Wu, Chenshuo Sun, Ying Shan 2023. M^2UGen: Multimodal music understanding and generation with the power of large language models. *arXiv preprint arXiv:2311.11255*.

Kang, W.-C. and McAuley, J. 2018. Self-attentive sequential recommendation. In *Proceedings of ICDM*.

Li, J., Wang, W., McAuley, J., and Zhao, W. X. 2023. Text Is All You Need: Learning Language Representations for Sequential Recommendation. In *Proceedings of SIGKDD 2023*.

Hou Y, Zhang J ,Lin Z, Lu H, Xie R ,McAuley J,Zhao W. X. 2024. Large Language Models are Zero-Shot Rankers for Recommender Systems. In *Proceedings of ECIR 2024*.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of EMNLP 2023*.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of ACL*.

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-T., Koh, P. W., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long-form text generation. In *Proceedings of EMNLP 2023*.

Moscati, M., Parada-Cabaleiro, E., Deldjoo, Y., Zangerle, E., and Schedl, M. 2022. Music4All-Onion: A large-scale multi-faceted content-centric music recommendation dataset. In *Proceedings of CIKM 2022*.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. OpenAI, August 5, 2025. URL: https://openai.com/index/gpt-oss-model-card.

Ploshkin, K. et al. 2025. Yambda-5B — A Large-Scale Multi-Modal Dataset for Ranking and Retrieval. In *Proceedings of RecSys 2025*.

Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of UAI 2009*, pages 452–461.

Spijkervet, J. and Burgoyne, J. A. 2021. Contrastive learning of musical representations. *arXiv preprint arXiv:2103.09410*.

Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., and Jiang, P. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of CIKM 2019*.

van den Oord, A., Dieleman, S., and Schrauwen, B. 2013. Deep content-based music recommendation. In *Proceedings of NeurIPS 2013*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of NeurIPS 2022*.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proceedings of ICASSP 2023*.

Yue, Z., Rabhi, S., Larson, M., and Hanjalic, A. 2023. LlamaRec: Two-stage recommendation using large language models for ranking. In *Proceedings of CIKM 2023*.

Zhou, K., Yu, H., Zhao, W. X., and Wen, J.-R. 2022. Filter-enhanced MLP is all you need for sequential recommendation. In *Proceedings of The ACM Web Conference 2022*, pages 2388–2399.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. 2023. Large language models are human-level prompt engineers. In *Proceedings of ICLR 2023*.

Grattafiori, A. et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud ,L. R., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., and El Sayed, W. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of NeurIPS 2020*.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Xiao, S., Liu, Z., Zhang, P., and Muennighoff, N. 2024. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of SIGIR 2024*.

A Terminology

B Expanded Related Work

Sequential recommendation has moved from recurrent session models such as GRU4Rec (Hidasi et al., 2016) to self-attentive architectures such as SASRec (Kang and McAuley, 2018) and BERT4Rec (Sun et al., 2019). Transferable and text-enhanced recommenders address a related representation problem: UniSRec (Hou et al., 2022) learns transferable sequential representations from textual item features, VQ-Rec (Hou et al., 2023b) replaces textual IDs with vector-quantized codes, and RecFormer (Li et al., 2023) fuses item text into a sequential transformer. These methods assume item text or a learned codebook, while our setting first constructs descriptor text from non-textual signals.

LLMs have been used as rankers, identifier generators, and recommender backbones. LLMRank reranks candidate items (Hou et al., 2024); BIGRec generates item identifiers (Bao et al., 2024); and LlamaRec adapts LLaMA for recommendation (Yue et al., 2023). Our LLM is not the ranker itself: it produces a compact intent state that is embedded and fused with a collaborative sequence representation.

Music-language models connect text to audio generation or understanding. MusicLM (Agostinelli et al., 2023) and MusicGen (Copet et al., 2023) generate music from text or melody-conditioned prompts, while LLark (Gardner et al., 2024) and M^2UGen (Hussain et al., 2023) adapt multimodal LLMs for music captioning, question answering, editing, and generation. CLAP (Wu et al., 2023; Elizalde et al., 2023) learns joint audio-text embeddings, and MusicCaps (Agostinelli et al., 2023) provides human-written music captions.

Paper term	Interpretation	Operationalization
Item-level pseudo-metadata	Track descriptor text	Audio cluster, artist community, duration, completion, discovery
Intent state	Contextual listening intent	Short verbalization from recent descriptors
Intent embedding / VIE	Semantic retrieval vector	Sentence embedding of the generated intent
Descriptor consistency	Grounded generated intent	Claims supported by the descriptor window
Acoustic style	Timbral neighborhood	k-means cluster over audio embeddings
Artist community	Co-listening affinity	Artist–cluster co-occurrence community
Energy	Coarse intensity cue	Derived from acoustic labels and completion cues
Listening completion	Play-through signal	Binned `played_ratio_pct` from training data
Discovery mode	Exposure source	Binned `is_organic` rate from training data

Table 10: Terminology used in the paper. VIE denotes verbalized intent embedding.

These resources motivate future automated descriptor labeling, but our experiments construct descriptors from interaction and audio evidence rather than curated audio-text pairs.

Faithfulness work in generated text focuses on source-grounded summarization and factual precision (Maynez et al., 2020; Min et al., 2023). Contextual listening-intent generation differs because the evidence is partly non-textual and unsupported claims may concern acoustic style, energy, playback behavior, or artist-community structure. We combine descriptor-consistency analysis, where hallucination is used only as shorthand for unsupported claims, with prompt ablation (Wei et al., 2022; Zhou et al., 2023) and probing methods (Conneau et al., 2018; Hewitt and Manning, 2019; Belinkov, 2022) to evaluate whether generated intent language preserves useful retrieval structure.

C Descriptor Construction Details

Yambda-50M contains 934,057 unique item IDs, with a heavy-tailed item distribution: the median item has only 2 interactions. We restrict descriptor construction to the 58,903 items appearing in at least 50 training-period LISTEN+ events. Of these, 56,102 (95.2%) have 128-dimensional audio embeddings generated by a contrastive CNN on audio spectrograms (Spijkervet and Burgoyne, 2021). We L2-normalize these embeddings and run k-means with $k = 20$, yielding silhouette score 0.184 and mean within-cluster cosine similarity 0.63.

From the artist-item mapping, we compute an artist-to-cluster affinity matrix and apply spectral clustering to obtain 12 artist communities. Since Yambda withholds genre metadata, artist co-occurrence is used as a structural proxy for broad scene affinity. Two authors inspected candidate acoustic-cluster labels using dominant artist communities, play-count distributions, and listening-completion patterns; the retained artifact contains the resolved label mapping but not the initial agreement count.

For each track, we aggregate mean `played_ratio_pct`, track length, and mean `is_organic` from training interactions only, then discretize these values into completion, duration, and discovery buckets. Yambda has no native energy tag, so energy is treated as a coarse derived attribute grounded primarily by acoustic-style label vocabulary and secondarily by recent completion cues. Energy inconsistency rates therefore measure descriptor consistency rather than external factual accuracy. For Music4All-Onion, native genres, Last.fm tags, and Spotify audio features allow richer descriptors such as genre, tags, tempo, valence, and energy.

D Prompt Templates

This appendix lists the short prompt templates used in the prompt-family ablation. The production prompts additionally include the recent descriptor bundle and, for one-shot and three-shot settings, the corresponding in-context examples.

P1 (free-form). *Given the user's last k tracks with these descriptors: [descriptors]. In one sentence, describe what this listener wants to hear next. Keep the answer under 25 tokens.*

P2 (query-like). *Based on these listening descriptors: [descriptors]. Write a short search query that would retrieve the ideal next track. Keep the answer under 25 tokens.*

P3 (structured). *Based on these listening descriptors: [descriptors]. Describe the listener's intent in this exact format: Acoustic style: [style]. Energy: [level]. Listening-completion pattern: [brief phrase]. Keep the answer under 25 tokens.*

Model	Stratum	PPL ↓	BS F1 ↑	D-2 ↑	D-3 ↑	Avg tok
Q3-8B	Short	18.7	0.761	0.412	0.538	17.8
Q3-8B	Medium	15.4	0.792	0.438	0.561	18.4
Q3-8B	Long	13.8	0.817	0.451	0.579	18.9
L3-8B	Short	20.1	0.742	0.396	0.519	17.2
L3-8B	Medium	16.8	0.773	0.421	0.546	17.9
L3-8B	Long	15.2	0.801	0.436	0.563	18.3
M7B	Short	22.6	0.718	0.374	0.493	16.6
M7B	Medium	18.9	0.751	0.401	0.522	17.1
M7B	Long	16.7	0.779	0.419	0.548	17.7

Table 11: Full automatic linguistic quality by model and history stratum.

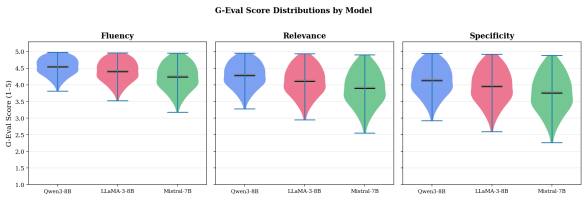

Figure 4: Violin plots of G-Eval score distributions by model for fluency, relevance, and specificity.

Model	Flu. ↑	Rel. ↑	Spec. ↑	Hum. Acc. ↑	κ
Q3-8B	4.62 ± 0.31	4.36 ± 0.42	4.21 ± 0.50	86.0	0.72
L3-8B	4.48 ± 0.37	4.18 ± 0.48	4.03 ± 0.56	82.7	0.68
M7B	4.31 ± 0.44	3.96 ± 0.55	3.82 ± 0.61	78.0	0.65

Table 12: G-Eval scores (mean $\pm$ std) and human validation results.

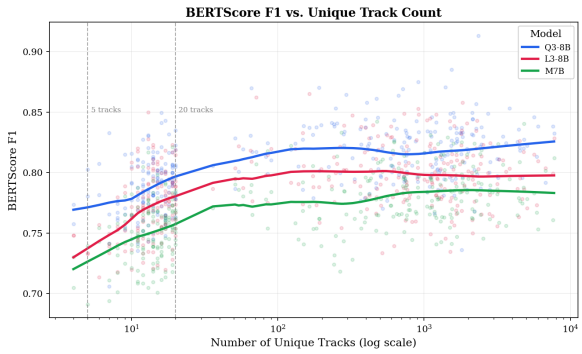

Figure 5: BERTScore F1 versus number of unique tracks (log-scale x-axis) with LOESS trend lines per model. Vertical dashed lines mark the 5 and 20 track boundaries.

E Linguistic Quality Details

F Descriptor-Consistency Annotation Protocol

Human validation used a binary descriptor-consistency question, per-type labels for acoustic, energy, artist-group, and playback-pattern claims, and three annotators over 600 intent states. Annotators saw the last five descriptor strings for the user, the generated intent state, and the four descriptor-inconsistency definitions from Section 4.2. They were instructed to mark a claim as *supported* if it was entailed by at least one recent descriptor or by a stable pattern across the descriptors, *inconsistent* if it introduced an incompatible or absent attribute, and *vague* if the claim was too underspecified to verify. Vague claims were not counted as inconsistent. The binary descriptor-consistency label was positive only when the intent contained no inconsistent claim under this definition.

For adjudication, the final human label was the majority vote among the three annotators. Cases with three-way disagreement among *supported*, *inconsistent*, and *vague* were resolved as *vague* for the per-claim label and as consistent for the aggregate rate, a conservative choice that avoids inflating inconsistency counts. The automated descriptor-consistency pipeline first matched acoustic-style and energy claims against descriptor vocabularies, then passed artist-group and playback-pattern claims to a gpt-oss-20B consistency checker with the recent descriptor bundle as context. Claims were labeled as supported, inconsistent, or vague; only inconsistent claims counted toward the rates in

Table 7. We use the same fixed gpt-oss-20B checkpoint (OpenAI, 2025) for both G-Eval scoring and consistency checking, with task-specific prompts and fixed output schemas; these auxiliary judgments are interpreted alongside human validation and rule-based checks rather than as stand-alone ground truth.

The retained annotation export supports model-level agreement (Table 12), per-stratum agreement (Table 33), and human-versus-automated agreement (Table 34). It does not retain separate per-type inter-annotator agreement, so we do not report per-type κ values or make per-type reliability claims.

We recognize that the absence of per-type κ values limits the validation of individual descriptor-inconsistency categories. In particular, the distinction between acoustic-style and energy inconsistency may be difficult for annotators to make reliably, given that energy is a derived attribute grounded partly in acoustic-style labels. Future annotation efforts should collect per-type agreement and include a calibration round where annotators discuss boundary cases before independent labeling.

G Additional Descriptor-Consistency Analyses

Model	ρ(Incon., NDCG@10)	ρ(Incon., Rec@10)	p
Q3-8B	-0.34	-0.29	0.004
L3-8B	-0.37	-0.32	0.002
M7B	-0.31	-0.27	0.011
All pooled	**-0.38**	**-0.33**	< 0.001

Table 13: Spearman correlation between per-user descriptor-inconsistency rate and retrieval metrics.

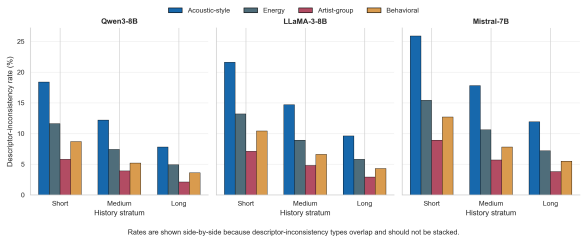

Figure 6: Descriptor-inconsistency type rates by model and history stratum. Acoustic-style inconsistency is measured against the derived acoustic-style cluster labels used in descriptor construction. Rates are shown side-by-side rather than stacked because the four inconsistency types can overlap within the same intent state.

H Additional Qualitative Examples

This appendix provides the qualitative cases referenced in Section 4.2. The figure is intended for error inspection rather than quantitative comparison.

I Intent-Embedding Visualizations

This appendix provides the multi-model intent-embedding view that complements the single-model visualization in Figure 3.

J Prompt Interaction Analysis

This appendix reports the ANOVA summary used to contextualize the prompt ablation in Section 4.1.

K Retrieval Ablations and Robustness

This appendix collects retrieval controls that are too detailed for the main text the full retrieval table, cold-start breakdown, descriptor-window and popularity stratifications, descriptor-only baselines, fusion architecture ablation, descriptor sensitivity, controlled descriptor-inconsistency interventions, runtime profile, bootstrap intervals, and pairwise significance tests. The fusion ablation tests whether intent should be appended to the collaborative state or used as a query over it; cross-attention is strongest and is therefore the default. The descriptor-sensitivity table separates semantic labels, neutral cluster IDs, random strings, and acoustic-style removal to test whether retrieval depends on author-written music labels. The runtime profile shows that generation, not sentence encoding or retrieval, dominates cost, which motivates caching and asynchronous refresh in deployment. The Music4All-Onion rows should be read as a secondary evaluation because they use the source paper's random 60/20/20 split with binarized interactions, which differs from Yambda GTS in split method and interaction definition.

L Acoustic Cluster Label Mapping

Table 29 provides the resolved mapping from k-means cluster IDs to author-assigned verbal labels. Labels were assigned by inspecting dominant artist communities, high-play-count tracks, and listening-completion patterns, with the primary criterion that the label should be interpretable to a music listener unfamiliar with the specific anonymized tracks. The mapping was then checked by three external music-aware annotators over all 20 clusters

Model	Short ρ (p)	Medium ρ (p)	Long ρ (p)	Partial ρ (p)
Q3-8B	-0.19 (0.007)	-0.23 (0.001)	-0.27 (< 0.001)	-0.22 (0.002)
L3-8B	-0.17 (0.016)	-0.21 (0.003)	-0.25 (< 0.001)	-0.20 (0.005)
M7B	-0.14 (0.048)	-0.18 (0.011)	-0.22 (0.002)	-0.16 (0.024)
All pooled	—	—	—	$\mathbf{-0.24}$ ($< \mathbf{0.001}$)

Table 14: Within-stratum and partial Spearman correlations between descriptor-inconsistency rate and NDCG@10. Partial correlations control for log unique-track count.

(Table 3); this validates label fit as prompt-facing pseudo-metadata, not as native genre ground truth.

M Additional Probing Controls

This appendix reports probing controls for descriptor-label removal and for the metadata-rich Music4All-Onion setting.

N Additional Human Evaluation Details

This appendix reports the human-evaluation summaries retained from the annotation export.

O Reproducibility Details

All experiments use public benchmark datasets. Yambda and Music4All-Onion should be obtained from their original distributors under the corresponding dataset terms; we do not redistribute raw interaction logs or audio embeddings. For the camera-ready version, we will release the descriptor-construction scripts, prompt templates, model configuration files, split-generation code, evaluation scripts, and generated intent states for item and user identifiers that can be shared under the source dataset licenses. During anonymous review, the repository URL is withheld and will be supplied as an anonymized artifact if requested by the reviewers.

This appendix records the main implementation settings used for reproduction. All random seeds are drawn from $\{42, 123, 456, 789, 2024\}$, with `torch`, `numpy`, `random`, and `torch.cuda` seeded; deterministic cuDNN is enabled and benchmarking is disabled. Yambda uses the official GTS protocol: days 1–300 for training, a 30-minute gap, a 1-day test window, Listen+ interactions, and the last 10% of the training period for development. Music4All-Onion uses the source paper's random 60/20/20 train/development/test split with binarized interactions on 79,072 items with complete features (Moscati et al., 2022).

SASRec uses two blocks, two heads, hidden size 256, maximum sequence length 50, dropout 0.2, Adam with learning rate 10^{-3}, batch size 256, maximum 200 epochs, and early stopping patience 10 on development NDCG@10. Fusion training freezes the sentence encoder and SASRec backbone, then trains the cross-attention module with Adam, learning rate 10^{-3}, batch size 256, maximum 100 epochs, and early stopping patience 5. Evaluation ranks against the full catalog rather than sampled negatives, using a FAISS IVF-Flat index with 1024 centroids, $nprobe = 32$, and inner product over L2-normalized 128-dimensional query vectors.

Intent generation uses greedy decoding with temperature 0, max 25 new tokens, the last five descriptors, and the P3 one-shot structured prompt for the default configuration. We use temperature 0 to make intent states deterministic and to prioritize descriptor-consistency over lexical diversity. The LLMs run in float16 with batch size 32. Sentence encoders are frozen; MiniLM uses mean pooling, while BGE-small and E5-small use CLS pooling with maximum token length 128.

Model	Prompt	Shot	Hist	Budget	Source	BS F1 ↑	G-Rel ↑	NDCG@10 ↑	Rec@10 ↑
Q3-8B	**P3**	**1**	**5**	**25**	**org-w**	**0.792**	**4.36**	**0.0827**	**0.0360**
	P1	1	5	25	org-w	0.763	4.08	0.0795	0.0347
	P2	1	5	25	org-w	0.774	4.15	0.0803	0.0350
	P3	0	5	25	org-w	0.779	4.21	0.0810	0.0353
	P3	3	5	25	org-w	0.789	4.33	0.0821	0.0357
	P3	1	5	25	all	0.786	4.28	0.0817	0.0355
	P3	1	5	25	org	0.783	4.25	0.0813	0.0352
	P3	1	3	25	org-w	0.771	4.17	0.0807	0.0352
	P3	1	10	25	org-w	0.786	4.29	0.0819	0.0356
	P3	1	5	10	org-w	0.768	4.12	0.0801	0.0349
	P3	1	5	50	org-w	0.785	4.27	0.0818	0.0356
L3-8B	**P3**	**1**	**5**	**25**	**org-w**	**0.773**	**4.18**	**0.0820**	**0.0358**
	P1	1	5	25	org-w	0.742	3.92	0.0788	0.0343
	P2	1	5	25	org-w	0.756	4.01	0.0797	0.0348
	P3	0	5	25	org-w	0.759	4.04	0.0805	0.0351
	P3	3	5	25	org-w	0.769	4.14	0.0814	0.0354
	P3	1	5	25	all	0.766	4.11	0.0812	0.0353
	P3	1	5	25	org	0.763	4.08	0.0808	0.0350
	P3	1	3	25	org-w	0.752	3.98	0.0799	0.0349
	P3	1	10	25	org-w	0.766	4.11	0.0811	0.0353
	P3	1	5	10	org-w	0.748	3.95	0.0793	0.0346
	P3	1	5	50	org-w	0.764	4.09	0.0809	0.0352
M7B	**P3**	**1**	**5**	**25**	**org-w**	**0.751**	**3.96**	**0.0812**	**0.0353**
	P1	1	5	25	org-w	0.721	3.71	0.0776	0.0339
	P2	1	5	25	org-w	0.733	3.79	0.0785	0.0343
	P3	0	5	25	org-w	0.738	3.84	0.0795	0.0348
	P3	3	5	25	org-w	0.747	3.92	0.0805	0.0351
	P3	1	5	25	all	0.744	3.89	0.0804	0.0349
	P3	1	5	25	org	0.741	3.86	0.0800	0.0346
	P3	1	3	25	org-w	0.731	3.77	0.0789	0.0345
	P3	1	10	25	org-w	0.744	3.89	0.0803	0.0350
	P3	1	5	10	org-w	0.727	3.74	0.0782	0.0342
	P3	1	5	50	org-w	0.742	3.87	0.0801	0.0349

Table 15: Full prompt ablation on Yambda-50M. Best configuration per model is shown in bold.

Factor	F	p-value	η^2
Prompt style (P)	38.7	< 0.001	0.42
Shot count (S)	12.4	0.002	0.14
History (H)	8.1	0.009	0.09
Budget (B)	6.3	0.018	0.07
Source (Src)	4.9	0.034	0.05
P $\times$ S	2.1	0.142	0.02
P $\times$ H	1.4	0.261	0.01
S $\times$ Src	0.8	0.458	0.01

Table 16: Two-way ANOVA summary over the reported ablation configurations using NDCG@10 as the dependent variable. Prompt style is the dominant main effect, while the tested pairwise interactions are not significant.

Qualitative Examples: Descriptors → Intent States

Two sparse-history and two dense-history Yambda-50M users · Red-highlighted text = descriptor-inconsistent attributes not grounded in recent history

(a) Sparse-history example 1 · User ID: 223300 · 5 unique tracks · high descriptor-inconsistency risk

RECENT TRACK DESCRIPTORS

T1	Acoustic style: C10. Duration: short. Engagement: medium. Discovery: organic-dominant.
T2	Acoustic style: C10. Duration: short. Engagement: medium. Discovery: organic-dominant.
T3	Acoustic style: C10. Duration: short. Engagement: medium. Discovery: organic-dominant.
T4	Acoustic style: C10. Duration: short. Engagement: medium. Discovery: organic-dominant.
T5	Acoustic style: C10. Duration: medium. Engagement: low. Discovery: organic-dominant.

GENERATED INTENT STATES · P3 structured format

Q3-8B	Acoustic style: C10. Energy: medium. Engagement pattern: organic-dominant. short sessions.	consistent
L3-8B	Acoustic style: C10. Energy: high. Engagement pattern: organic-dominant. listening with variety seeking.	1 desc.-incon.
M7B	Acoustic style: C0. Energy: high. Engagement pattern: algorithmic-dominant deep dives with extended focus.	3 desc.-incon.

(b) Sparse-history example 2 · User ID: 464400 · 5 unique tracks · high descriptor-inconsistency risk

RECENT TRACK DESCRIPTORS

T1	Acoustic style: C9. Duration: short. Engagement: low. Discovery: organic-dominant.
T2	Acoustic style: C9. Duration: medium. Engagement: medium. Discovery: mixed.
T3	Acoustic style: C15. Duration: medium. Engagement: medium. Discovery: mixed.
T4	Acoustic style: C19. Duration: medium. Engagement: medium. Discovery: mixed.
T5	Acoustic style: C19. Duration: medium. Engagement: medium. Discovery: mixed.

GENERATED INTENT STATES · P3 structured format

Q3-8B	Acoustic style: C9. Energy: medium. Engagement pattern: mixed. short sessions.	consistent
L3-8B	Acoustic style: C9. Energy: high. Engagement pattern: mixed. listening with variety seeking.	1 desc.-incon.
M7B	Acoustic style: C0. Energy: high. Engagement pattern: algorithmic-dominant deep dives with extended focus.	3 desc.-incon.

(c) Dense-history example 1 · User ID: 665600 · 2504 unique tracks · mostly consistent outputs

RECENT TRACK DESCRIPTORS

T1	Acoustic style: C10. Duration: medium. Engagement: medium. Discovery: mixed.
T2	Acoustic style: C18. Duration: medium. Engagement: medium. Discovery: algorithmic-dominant.
T3	Acoustic style: C8. Duration: medium. Engagement: medium. Discovery: organic-dominant.
T4	Acoustic style: C10. Duration: short. Engagement: medium. Discovery: mixed.
T5	Acoustic style: C18. Duration: medium. Engagement: medium. Discovery: mixed.

GENERATED INTENT STATES · P3 structured format

Q3-8B	Acoustic style: C10. Energy: medium. Engagement pattern: sustained mixed. listening.	consistent
L3-8B	Acoustic style: C10. Energy: medium. Engagement pattern: mixed. exploration, extended sessions.	consistent
M7B	Acoustic style: C10. Energy: high. Engagement pattern: mixed. listening with calm focus.	1 desc.-incon.

(d) Dense-history example 2 · User ID: 997000 · 1051 unique tracks · mostly consistent outputs

RECENT TRACK DESCRIPTORS

T1	Acoustic style: C16. Duration: short. Engagement: high. Discovery: algorithmic-dominant.
T2	Acoustic style: C7. Duration: medium. Engagement: high. Discovery: algorithmic-dominant.
T3	Acoustic style: C12. Duration: medium. Engagement: high. Discovery: algorithmic-dominant.
T4	Acoustic style: C1. Duration: medium. Engagement: high. Discovery: algorithmic-dominant.
T5	[no descriptor for 5349284]

GENERATED INTENT STATES · P3 structured format

Q3-8B	Acoustic style: C16. Energy: high. Engagement pattern: sustained algorithmic-dominant. listening.	consistent
L3-8B	Acoustic style: C16. Energy: high. Engagement pattern: algorithmic-dominant. exploration, extended sessions.	consistent
M7B	Acoustic style: C16. Energy: low. Engagement pattern: algorithmic-dominant. listening with calm focus.	1 desc.-incon.

Figure 7: Appendix view of user-level qualitative examples showing the last five descriptor strings and the corresponding intent states produced by all three models for two low-history and two dense-history users. Red-highlighted terms indicate descriptor-inconsistent attributes not supported by the recent descriptor bundle.

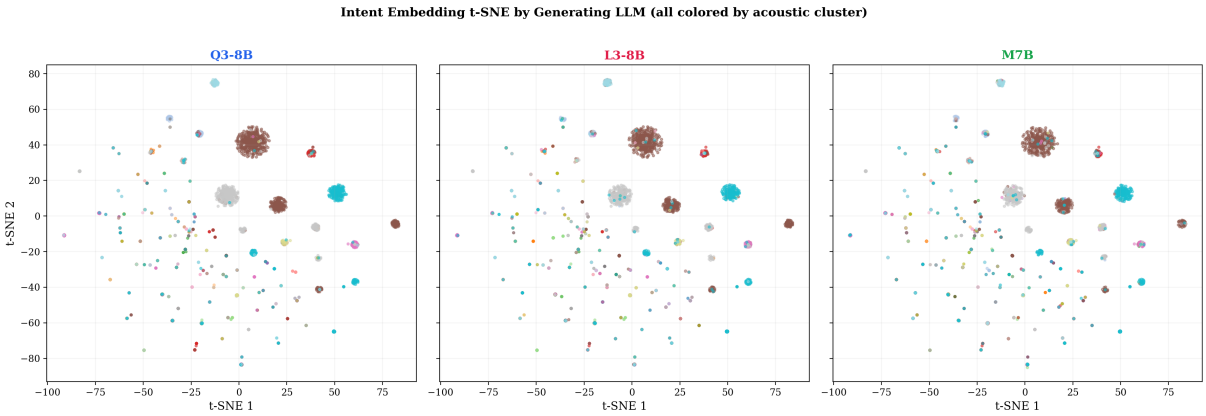

Figure 8: t-SNE projections of intent embeddings for Qwen3-8B, LLaMA-3-8B, and Mistral-7B (all BGE-small), colored by dominant acoustic cluster.

Dataset	Method	Encoder	NDCG@10 ↑	NDCG@100 ↑	Rec@10 ↑	Rec@100 ↑	Cov@10 ↑	Protocol
Yambda-50M								
Yambda-50M	MostPop	—	0.0186	0.0249	0.0064	0.0321	0.0000	GTS-REF
Yambda-50M	ItemKNN	—	0.0781	0.0934	0.0373	0.1297	0.0180	GTS-REF
Yambda-50M	iALS	—	0.0407	0.0642	0.0128	0.0808	0.0044	GTS-REF
Yambda-50M	BPR	—	0.0389	0.0641	0.0139	0.0836	0.0257	GTS-REF
Yambda-50M	SASRec	—	0.0748 ± 0.0011	0.0764 ± 0.0013	0.0325 ± 0.0008	0.1026 ± 0.0018	0.0130	GTS
Yambda-50M	BERT4Rec	—	0.0758 ± 0.0013	0.0825 ± 0.0014	0.0332 ± 0.0009	0.1040 ± 0.0019	0.0125	GTS
Yambda-50M	BERT4Rec+Q3-8B+BGE	BGE-sm	$0.0865 \pm 0.0011^{\ddagger}$	0.0885 ± 0.0012	0.0374 ± 0.0008	0.1168 ± 0.0016	0.0215	GTS
Yambda-50M	FMLP-Rec	—	0.0778 ± 0.0012	0.0845 ± 0.0013	0.0342 ± 0.0008	0.1058 ± 0.0017	0.0132	GTS
Yambda-50M	LLMRank	—	0.0710 ± 0.0014	0.0775 ± 0.0015	0.0318 ± 0.0009	0.1012 ± 0.0020	0.0096	GTS
Yambda-50M	Descriptor-concat	BGE-sm	$0.0808 \pm 0.0012^{\ddagger}$	0.0840 ± 0.0014	0.0352 ± 0.0008	0.1100 ± 0.0018	0.0175	GTS
Yambda-50M	Mean-desc-embed	BGE-sm	0.0795 ± 0.0013	0.0828 ± 0.0015	0.0346 ± 0.0009	0.1085 ± 0.0019	0.0160	GTS
Yambda-50M	Rule-template	BGE-sm	$0.0815 \pm 0.0011^{\ddagger}$	0.0848 ± 0.0013	0.0355 ± 0.0008	0.1108 ± 0.0017	0.0182	GTS
Yambda-50M	SASRec+audio	—	$0.0805 \pm 0.0012^{\ddagger}$	0.0855 ± 0.0014	0.0352 ± 0.0008	0.1098 ± 0.0018	0.0225	GTS
Yambda-50M	Q3-8B+BGE	BGE-sm	$0.0868 \pm 0.0009^{\ddagger}$	0.0878 ± 0.0012	0.0378 ± 0.0007	0.1175 ± 0.0016	0.0145	GTS
Yambda-50M	L3-8B+BGE	BGE-sm	$0.0840 \pm 0.0011^{\ddagger}$	0.0862 ± 0.0013	0.0370 ± 0.0008	0.1155 ± 0.0017	0.0140	GTS
Yambda-50M	M7B+BGE	BGE-sm	$0.0830 \pm 0.0012^{\ddagger}$	0.0850 ± 0.0014	0.0362 ± 0.0008	0.1138 ± 0.0018	0.0135	GTS
Yambda-50M	Q3-8B+audio+text	BGE-sm	$\mathbf{0.0875 \pm 0.0010}^{\ddagger}$	$\mathbf{0.0898 \pm 0.0012}$	$\mathbf{0.0390 \pm 0.0007}$	$\mathbf{0.1200 \pm 0.0016}$	**0.0412**	GTS
Yambda-500M								
Yambda-500M	ItemKNN	—	0.0708	0.0771	0.0320	0.1105	0.0257	GTS-REF
Yambda-500M	SASRec	—	0.0754 ± 0.0010	0.0884 ± 0.0012	0.0336 ± 0.0007	0.1065 ± 0.0018	0.0347	GTS
Yambda-500M	Descriptor-concat	BGE-sm	0.0812 ± 0.0011	0.0942 ± 0.0013	0.0358 ± 0.0008	0.1122 ± 0.0017	0.0415	GTS
Yambda-500M	SASRec+audio	—	0.0810 ± 0.0011	0.0945 ± 0.0013	0.0365 ± 0.0008	0.1148 ± 0.0017	0.0480	GTS
Yambda-500M	Q3-8B+BGE	BGE-sm	$\mathbf{0.0860 \pm 0.0009}$	$\mathbf{0.1000 \pm 0.0011}$	$\mathbf{0.0392 \pm 0.0007}$	$\mathbf{0.1189 \pm 0.0018}$	**0.0554**	GTS
Music4All-Onion								
Music4All-Onion	MostPop	—	0.0537	0.0610	0.0524	0.0785	0.0005	M4A-REF
Music4All-Onion	BPR	—	0.0534	0.0620	0.0515	0.0810	0.0160	M4A-REF
Music4All-Onion	BiVAE	mixed	0.0641	0.0755	0.0646	0.0980	0.0215	M4A-REF
Music4All-Onion	SASRec	—	0.0592 ± 0.0014	0.0710 ± 0.0013	0.0268 ± 0.0010	0.0740 ± 0.0018	0.0190	60/20/20
Music4All-Onion	Descriptor-concat	BGE-sm	0.0640 ± 0.0013	0.0770 ± 0.0012	0.0295 ± 0.0009	0.0805 ± 0.0017	0.0260	60/20/20
Music4All-Onion	Q3-8B+BGE	BGE-sm	$\mathbf{0.0680 \pm 0.0012}$	$\mathbf{0.0820 \pm 0.0011}$	$\mathbf{0.0312 \pm 0.0008}$	$\mathbf{0.0865 \pm 0.0016}$	**0.0365**	60/20/20

Table 17: Full retrieval results on Yambda-50M, Yambda-500M, and Music4All-Onion. Light green marks the strongest cells; lighter green marks the headline intent-only configuration when the audio+text variant is higher. All Yambda rows use the official LISTEN+ target from the Yambda benchmark; the separate LIKE task is not evaluated in this work. GTS-REF rows are official source-paper Yambda references under GTS; GTS rows are our GTS runs over five seeds; 60/20/20 marks our Music4All-Onion runs using the source paper's random 60/20/20 split with binarized interactions; and M4A-REF marks source-paper reference numbers under the same protocol. Means report $\pm$ standard deviation over seeds $\{42, 123, 456, 789, 2024\}$. † and ‡ indicate user-level paired bootstrap significance versus SASRec at $p < 0.05$ and $p < 0.01$, respectively. Note: our Music4All-Onion runs and M4A-REF rows are protocol-matched within Music4All-Onion on the 79,072 items with complete features, so they can be compared within that dataset but not to Yambda GTS rows. We do not run paired significance tests against source-paper baselines because their per-user outputs are unavailable.

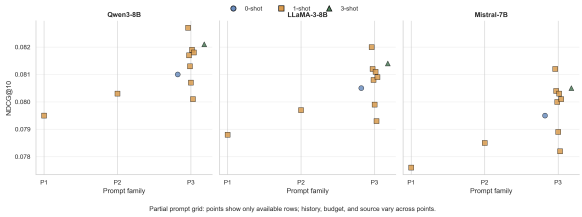

Figure 9: Prompt-ablation summary by model from the reported ablation rows. Points show observed NDCG@10 values by prompt family and shot count, while history length, budget, and interaction source may vary across points.

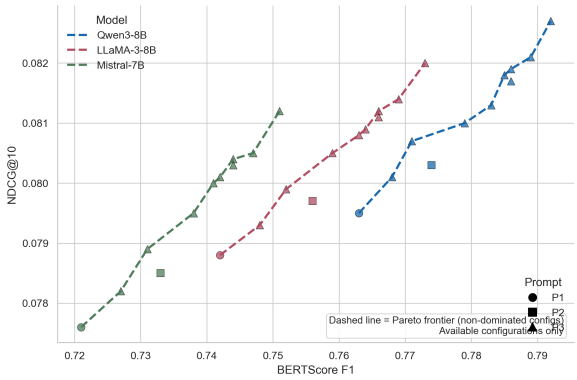

Figure 10: BERTScore F1 versus NDCG@10 across the reported prompt configurations. Marker shape denotes prompt family, color denotes generating LLM, and the dashed line marks the Pareto frontier of non-dominated configurations within each model.

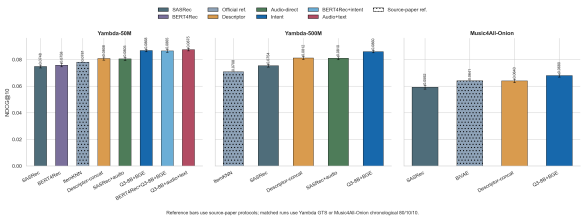

Figure 11: NDCG@10 across Yambda-50M, Yambda-500M, and Music4All-Onion. Source-paper reference bars are contextual comparators; matched-run error bars show ± 1 s.d. over five seeds.

History bucket	SASRec	Desc-concat	Q3-8B+BGE	Q3-8B+audio+text	Δ
< 5	0.0450 ± 0.0018	0.0505 ± 0.0016	0.0555 ± 0.0015	$\mathbf{0.0575 \pm 0.0014}$	$+0.0105^{\ddagger}$
5–11	0.0585 ± 0.0015	0.0635 ± 0.0013	0.0680 ± 0.0012	$\mathbf{0.0695 \pm 0.0012}$	$+0.0095^{\ddagger}$
12–24	0.0712 ± 0.0013	0.0755 ± 0.0012	0.0790 ± 0.0011	$\mathbf{0.0802 \pm 0.0011}$	$+0.0078^{\ddagger}$
25–54	0.0823 ± 0.0011	0.0852 ± 0.0010	0.0875 ± 0.0010	$\mathbf{0.0885 \pm 0.0009}$	$+0.0052^{\ddagger}$
> 54	0.0916 ± 0.0009	0.0932 ± 0.0008	0.0945 ± 0.0008	$\mathbf{0.0950 \pm 0.0008}$	$+0.0029^{\dagger}$

Table 18: Cold-start NDCG@10 by unique-track bucket on Yambda-50M under GTS. Δ is Q3-8B+BGE versus SASRec.

Last-5 coverage	SASRec	Q3-8B+BGE	Rel. gain
0 descriptors	0.073	0.073	0%
1–2 descriptors	0.070	0.077	+7%
3–4 descriptors	0.074	0.085	+12%
5 descriptors	0.076	0.089	+15%

Table 19: Retrieval stratified by descriptor coverage in the last-5 history window.

Item bucket	SASRec	Q3-8B+BGE	Rel. gain
Head items	0.090	0.103	+12%
Mid-tail items	0.055	0.061	+8%
Long-tail items	0.025	0.027	+3%

Table 20: Item-popularity stratified NDCG@10 on Yambda-50M. Gains are strongest for head items, where descriptors are most stable, but remain non-negative for mid-tail and long-tail items.

Fusion method	NDCG@10 $\uparrow$	95% CI	Δ vs Concat
Concat + Linear	0.0812 ± 0.0012	[0.0785, 0.0839]	—
2-layer MLP	0.0830 ± 0.0011	[0.0804, 0.0856]	+0.0018
Gated fusion	0.0858 ± 0.0010	[0.0833, 0.0883]	$+0.0046^{\ddagger}$
Cross-attention (default)	$\mathbf{0.0868 \pm 0.0009}$	[0.0845, 0.0891]	$+0.0056^{\ddagger}$

Table 21: Fusion architecture ablation on Yambda-50M under GTS. Cross-attention is the default configuration used for the main retrieval runs.

Descriptor variant	NDCG@10 $\uparrow$	Rel. vs SASRec
SASRec	0.0748	—
Semantic cluster labels	**0.0868**	+16.0%
Neutral cluster IDs	0.0850	+13.6%
Random cluster strings	0.0780	+4.3%
No acoustic-style field	0.0820	+9.6%

Table 22: Descriptor sensitivity ablation on Yambda-50M. Neutral cluster IDs retain most of the semantic-label benefit, while random strings substantially weaken the result.

Intent variant	Inconsistency	NDCG@10
Faithful Q3-8B intent	13%	0.0868
Constrained generation	$\leq$8%	0.0860
Style-corrupted intent	$\geq$70%	0.0810
Energy-corrupted intent	$\geq$60%	0.0830
Random-corrupted intent	$\geq$90%	0.0750

Table 23: Controlled descriptor-inconsistency intervention on Yambda-50M. Corrupting unsupported style and energy fields reduces retrieval quality, while constrained generation lowers inconsistency without a meaningful retrieval loss.

Component	Latency (ms)
LLM intent generation (Qwen3-8B, 25 tokens)	94.2
Sentence encoding (BGE-small)	4.7
FAISS retrieval (IVF-Flat, 1024 centroids)	2.4
SASRec-only inference	6.9
Full pipeline overhead	101.3

Table 24: Per-user wall-clock latency on a Xeon 40-core CPU server with 128GB RAM and an NVIDIA RTX A6000 (48GB) GPU.

Method	NDCG@10 ↑	Δ vs SASRec	Δ vs Desc-concat	Significance vs Desc-concat
SASRec (no text path)	0.0748 ± 0.0011	—	—	—
Random text + fusion	0.0753 ± 0.0012	+0.0005	—	ns
Mean descriptor embed	0.0795 ± 0.0013	+0.0047	-0.0013	—
Descriptor concatenation	0.0808 ± 0.0012	+0.0060	reference	—
Rule-based template	0.0815 ± 0.0011	+0.0067	+0.0007	0.31 (ns)
Q3-8B intent (cross-attn)	$\mathbf{0.0868 \pm 0.0009}$	$+0.0120^{\ddagger}$	+0.0060	$< 0.01^{\ddagger}$
L3-8B intent	0.0840 ± 0.0011	$+0.0092^{\ddagger}$	+0.0032	$0.012^{\dagger}$
M7B intent	0.0830 ± 0.0012	$+0.0082^{\ddagger}$	+0.0022	$0.038^{\dagger}$
Q3-8B+audio+text	$\mathbf{0.0875 \pm 0.0010}$	$+0.0127^{\ddagger}$	+0.0067	$< 0.001^{\ddagger}$

Table 25: Descriptor-only and no-LLM ablations on Yambda-50M under GTS. All rows use BGE-small where a text path is present and five seeds. Cross-attention is the default fusion module for the intent rows.

Dataset	Method	NDCG@10	95% Bootstrap CI	Overlap w/ SASRec
Yambda-50M	SASRec	0.0748	[0.0721, 0.0775]	—
Yambda-50M	FMLP-Rec	0.0778	[0.0750, 0.0806]	Partial
Yambda-50M	ItemKNN (official)	0.0781	[—, —]	—
Yambda-50M	Descriptor-concat	0.0808	[0.0780, 0.0836]	No
Yambda-50M	Rule-template	0.0815	[0.0788, 0.0842]	No
Yambda-50M	SASRec+audio	0.0805	[0.0778, 0.0832]	No
Yambda-50M	Q3-8B+BGE (cross-attn)	0.0868	[0.0845, 0.0891]	No
Yambda-50M	L3-8B+BGE	0.0840	[0.0814, 0.0866]	No
Yambda-50M	M7B+BGE	0.0830	[0.0803, 0.0857]	No
Yambda-50M	Q3-8B+audio+text	0.0875	[0.0850, 0.0900]	No
Yambda-500M	SASRec	0.0754	[0.0728, 0.0780]	—
Yambda-500M	Q3-8B+BGE	0.0860	[0.0836, 0.0884]	No
Music4All-Onion (60/20/20)	SASRec	0.0592	[0.0558, 0.0626]	—
Music4All-Onion (60/20/20)	Q3-8B+BGE	0.0680	[0.0650, 0.0710]	No

Table 26: Bootstrap confidence intervals for retrieval rows. Intervals use 10,000 user-level resamples with the percentile method.

Comparison	Δ NDCG@10	95% bootstrap CI	p-value
Q3-8B+BGE vs SASRec	+0.0120	$[+0.0104, +0.0136]$	< 0.001
Q3-8B+BGE vs BERT4Rec	+0.0110	$[+0.0093, +0.0127]$	< 0.001
Q3-8B+BGE vs Descriptor-concat	+0.0060	$[+0.0044, +0.0076]$	< 0.001
Q3-8B+BGE vs SASRec+audio	+0.0063	$[+0.0047, +0.0079]$	< 0.001
Q3-8B+BGE vs Rule-template	+0.0040	$[+0.0025, +0.0070]$	< 0.01
Q3-8B+audio+text vs Q3-8B+BGE	+0.0005	$[-0.0001, +0.0018]$	0.05

Table 27: User-level paired bootstrap significance tests on Yambda-50M. The audio+text comparison is reported conservatively because the interval crosses zero.

Comparison	Δ NDCG@10	95% bootstrap CI	p-value
Q3-8B+BGE vs SASRec	+0.0106	$[+0.0092, +0.0120]$	< 0.001
Q3-8B+BGE vs Descriptor-concat	+0.0048	$[+0.0035, +0.0062]$	< 0.001
Q3-8B+BGE vs SASRec+audio	+0.0050	$[+0.0036, +0.0064]$	< 0.001

Table 28: User-level paired bootstrap significance tests on Yambda-500M.

Cluster ID	Author-assigned label
0	Pop / Upbeat
1	Electronic / High-Energy
2	Acoustic / Mellow
3	Hip-Hop / Rhythmic
4	Orchestral / Extended
5	Ambient / Low-Energy
6	Rock / Driving
7	R&B / Smooth
8	Folk / Organic
9	Dance / Club
10	Metal / Intense
11	Jazz / Complex
12	Classical / Calm
13	Indie / Atmospheric
14	Latin / Percussive
15	Soul / Warm
16	Electronic / Chill
17	Punk / Raw
18	World / Eclectic
19	Synthwave / Retro

Table 29: Resolved acoustic-cluster ID to verbal-label mapping used for Yambda descriptor construction.

LLM	Encoder	NMI ↑	Purity ↑	NN sim ↑	NN rand	Δ cos	Probe acc ↑	Majority
Q3-8B	MiniLM	0.342	0.514	0.621	0.318	0.303	0.462	0.184
Q3-8B	BGE-small	**0.386**	**0.552**	**0.658**	0.321	**0.337**	**0.497**	0.184
Q3-8B	E5-small	0.371	0.539	0.644	0.319	0.325	0.482	0.184
L3-8B	MiniLM	0.318	0.493	0.594	0.315	0.279	0.438	0.184
L3-8B	BGE-small	0.359	0.528	0.632	0.318	0.314	0.471	0.184
L3-8B	E5-small	0.346	0.516	0.617	0.317	0.300	0.456	0.184
M7B	MiniLM	0.291	0.468	0.563	0.312	0.251	0.407	0.184
M7B	BGE-small	0.327	0.502	0.601	0.316	0.285	0.439	0.184
M7B	E5-small	0.314	0.487	0.586	0.314	0.272	0.426	0.184
Random text	*MiniLM*	*0.074*	*0.216*	*0.286*	*0.281*	*0.005*	*0.193*	*0.184*

Table 30: Full embedding probing results.

LLM	Encoder	NMI ↑	Purity ↑	Δ cos	Probe acc ↑	Majority
Q3-8B	BGE-small	0.217	0.381	0.189	0.312	0.184
L3-8B	BGE-small	0.194	0.356	0.168	0.289	0.184
M7B	BGE-small	0.168	0.329	0.143	0.261	0.184
Random text	*BGE-small*	*0.074*	*0.216*	*0.005*	*0.193*	*0.184*

Table 31: Probing results after removing acoustic-style labels from descriptors and retaining only duration, listening-completion, and discovery cues.

LLM	Encoder	NMI ↑	Purity ↑	Probe acc ↑	Majority
Q3-8B	BGE-small	0.438	0.601	0.547	0.184
L3-8B	BGE-small	0.411	0.578	0.521	0.184
M7B	BGE-small	0.382	0.549	0.491	0.184
Random text	*BGE-small*	*0.081*	*0.224*	*0.198*	*0.184*

Table 32: Music4All-Onion probing results against native genre labels. Richer native metadata yields stronger intent-embedding structure than in the metadata-free Yambda setting.

Model	Short κ	Med κ	Long κ	Short Acc	Med Acc	Long Acc
Q3-8B	0.65	0.74	0.78	82	87	89
L3-8B	0.63	0.70	0.74	78	84	86
M7B	0.60	0.66	0.69	72	79	83

Table 33: Per-stratum human evaluation agreement and accuracy. Agreement improves monotonically with history density for all three models.

	Auto=Incon.	Auto=OK
Human=Incon.	336	36
Human=OK	44	184

Table 34: Confusion matrix between human descriptor-inconsistency labels and the automated descriptor-consistency pipeline. Precision is 0.88, recall is 0.90, and Cohen's κ is 0.73.

Component	Y-50M	Y-500M	M4A-Onion
SASRec training (5 seeds)	2.5 hr	15.0 hr	1.5 hr
Intent generation (3 LLMs)	3.0 hr	8.0 hr	1.0 hr
Fusion training (5 seeds)	1.5 hr	5.0 hr	1.0 hr
Per-dataset total	$\sim$7.5 hr	$\sim$29.0 hr	$\sim$4.0 hr

Table 35: Approximate RTX A6000 GPU-hours by dataset. Grand total is approximately 40.5 RTX A6000 GPU-hours (excluding the re-run experiments).