State-Driven Retrieval and Learned Re-Ranking for Conversational Music Recommendation

Nidhin Pattaniyil
npatta01@gmail.com
Independent Researcher
Hoboken, New Jersey, USA

Semih Yagli
semihyagli1@gmail.com
Independent Researcher
Hoboken, New Jersey, USA

Tanwir Zaman
zaman.tanwir@gmail.com
Independent Researcher
Hoboken, New Jersey, USA

Abstract

We describe team npatta01's submission to the RecSys Challenge 2026 conversational music recommendation task. The pipeline extracts a typed conversation state with an LLM, gathers candidates from eleven retrieval branches over a unified track index, re-ranks them with a LambdaMART model, and uses the state to generate a response. On the final Blind-B leaderboard it scored 0.3811 composite (nDCG@20 0.2537, LLM-judge 3.30), ranking 29th of 40 teams. We then examine why. The development estimates we selected on were computed in-sample and overstated performance. The training conversations are LLM-generated, and on many turns we were unsure that the single ground-truth track matched the request — often it repeats the just-played artist after an explicit request for someone else. We re-judged the turns with LLM judges and release that relabeling; a model trained on it scored lower against the original labels, so the submission kept them. Failure cases from the submitted run show extracted constraints the pipeline could not enforce. Code, models, and a full reproduction bundle are publicly released.

CCS Concepts

• **Information systems** → **Recommender systems**; *Retrieval models and ranking*; *Music retrieval*.

Keywords

conversational recommendation, music recommendation, learning to rank, rank fusion, label noise, RecSys Challenge

ACM Reference Format:
Nidhin Pattaniyil, Semih Yagli, and Tanwir Zaman. 2026. State-Driven Retrieval and Learned Re-Ranking for Conversational Music Recommendation. In *RecSys Challenge 2026 (RecSysChallenge '26), October 02, 2026, Minneapolis, MN, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3842413.3842420

1 Introduction

The Music-CRS challenge [6] asks systems to act as a conversational music recommender. Each session is a multi-turn dialog between a listener and an assistant; at every assistant turn the system must (i) retrieve a ranked list of 20 tracks from a shared 47,071-track catalog and (ii) write a short natural-language response presenting its top recommendation. Training data comprises 15,199 sessions built on the TalkPlayData 2 collection [2, 5], with per-track metadata (title, artist, album, a median of 17 genre/mood tags, popularity, release date) and precomputed embeddings (text, audio, cover art, collaborative filtering). Evaluation is on two 80-session blind splits of final turns; Blind-B additionally removes the conversation-goal annotations present in training data. Submissions are scored with the organizers' composite metric [6]:

$$0.50 \cdot \text{nDCG@20} + 0.10 \cdot \text{cat. div.} + 0.10 \cdot \text{lex. div.} + 0.30 \cdot \frac{\text{judge}-1}{4},$$

where "judge" is a 1–5 rating of the generated response from the organizers' LLM-based evaluator [7]. In addition to the training sessions, a 1,000-session development split with public per-turn ground truth (one target track per turn) supports offline iteration.

Our system is a retrieve-then-rerank pipeline conditioned on an explicitly compiled conversation state. The deployed reranker is *goal-free* at inference: no conversation-goal annotation is an input feature, though goal-progress annotations did shape its training labels (Section 2.3). Section 2 documents the pipeline, Section 3 the results, and Section 4 the errors behind the final score.

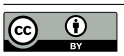
Related work. The design combines four established lines rather than proposing a new method. Conversational recommenders commonly track dialog state explicitly and condition retrieval on it [8]; we follow that pattern but compile the state with a prompted LLM instead of a trained tracker. Large-catalog systems retrieve then rerank, pooling multiple sources with reciprocal-rank fusion [3] and ranking with LambdaMART [1, 9]. Music retrieval adds modality-specific encoders — contrastive language-audio pretraining [17] and vision-language models over cover art [16] — alongside collaborative-filtering embeddings [14] and general-purpose text embeddings [19]. What we report is how that combination behaved, not a new technique.

2 Approach

Figure 1 shows the pipeline end to end; the subsections below follow it in order.

2.1 State extraction

At each turn we prompt DeepSeek-V4-Flash (build 0423) [4] with the dialog so far — the turns themselves and the IDs of the tracks already played — to emit a schema-constrained conversation state. None of the dataset's supplied annotations — conversation goals, goal progress, or user profiles — reach the extractor, so the state is whatever an LLM can read off the conversation. Its fields cover the current request and intent mode, track feedback and pinned references, mentioned entities, hard filters and explicit rejections, process constraints, routing flags, lyrical theme, and a release-year window; Figure 2 shows the schema populated on a real turn. A



RecSysChallenge '26, Minneapolis, MN, USA

ACM ISBN 979-8-4007-2863-1/2026/10
https://doi.org/10.1145/3842413.3842420

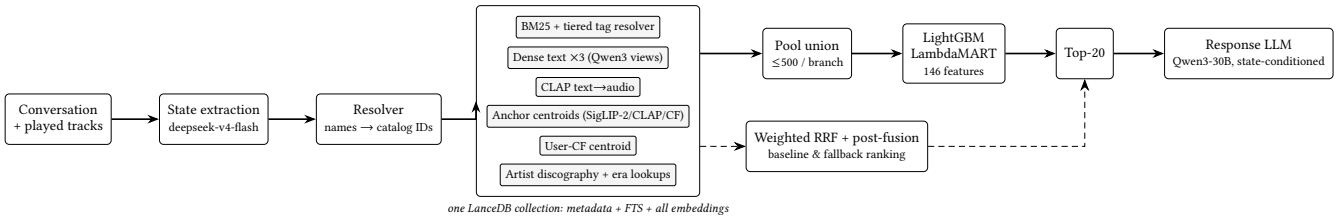


Figure 1: System overview. The conversation is compiled into a typed state; eleven retrieval branches query one LanceDB collection; the LambdaMART reranker scores the branch-pool union and emits the top-20; a response LLM presents the top track. Weighted RRF over the same branches provides the baseline and fallback ranking (dashed).

fuzzy resolver grounds surface names to catalog IDs so downstream stages operate on identifiers: fuzzy string matching of artist and track names against the catalog name lists, keeping matches above 80/100. Cheaper extractors under-filled optional fields such as rejections and facets, so all final runs use it.

User (turns 2–3): "…subdued, female singer, a sense of place in the lyrics, stark almost a cappella delivery…not it either."

Extracted state (abridged):

```
request_type: hidden_target
anchor artist: Neko Case (must_use)
rejected tracks: "Calling Cards", "Man"
facets: mood=subdued; lyrical_theme=sense of place;
        sonic=stark almost a cappella
```

Top-1: Neko Case — "Bracing For Sunday"

Submitted response: "You're looking for something subdued with a strong sense of place and a stark, almost a cappella delivery—and *Bracing For Sunday* fits that mood perfectly. …"

Figure 2: A Blind-B turn: extracted state and submitted output for a hidden-target request.

2.2 Candidate retrieval

All per-track information lives in a single LanceDB collection [10]: metadata fields, a full-text index, and every embedding column — three Qwen3-Embedding text views (lyrics from the 0.6B model at 1024-d; metadata and attributes re-indexed with the 8B model), LAION-CLAP audio (512-d) [17], SigLIP-2 cover art (768-d) [16], BPR collaborative-filtering vectors (128-d) [14], and our bi-encoder document vectors (2560-d, Section 2.4). The goal was to express every filter in one space: each branch is a differently-shaped query against the same table, so retrieval could in principle collapse into a single composed query. We never got there (Section 4.3).

Table 1 lists the eleven branches the reranker scores. A twelfth, a SigLIP-2 text-to-cover-art branch, is configured but gated: it fires only when the state flags an image or visual request, and the reranker carries no features for it. In the BM25 branch [15], free-text mood and genre phrases first pass through a tiered tag resolver: exact, alias, and substring matching against the catalog tag vocabulary, with an embedding nearest-neighbor fallback (cosine ≥ 0.6, top-3). Weighted reciprocal-rank fusion [3] (k=60, top-1000) plus post-fusion multiplicative adjustments (rejected artists/tracks dropped, played tracks dropped, exact-membership tag promotes/demotes, year-range decay) yields an explicit ranking that serves as baseline and fallback.

Table 1: The eleven retrieval branches (all queries against the one collection).

Branch	Space	Query
BM25 (fielded)	lexical	state text + resolved tags
Dense text ×3	Qwen3 views	state-built strings
CLAP text→audio	audio	sonic description
Anchor centroids ×3	SigLIP-2/CLAP/CF	liked-track centroids†
User-CF centroid	CF-BPR	the user's own vector
Lookups ×2	catalog	discography; era popularity

† Gated by intent mode: a pivot suppresses anchor centroids.

Table 2: Reranker feature families (share of total gain, deployed model).

Family	#	Gain	Examples
Bi-encoder cosine	1	59.0%	`b1_cos`
Per-branch rank/score	66	11.5%	BM25 rank, dense margins
Session/artist affinity	18	10.7%	same-artist, played count
Other similarity cosines	17	8.9%	CLAP/SigLIP centroids
Popularity	7	5.0%	pop. percentile
State/intent	14	2.2%	request type, intent mode
Tag/lexical overlap	10	2.0%	IDF tag overlap
Constraints/rejections	13	0.7%	rejected-artist flag

2.3 Candidate re-ranking

The submitted system replaces the fused order: a LightGBM LambdaMART [1, 9] model scores the union of the branch pools (each truncated to its top 500) and emits the final top-20. It sees 146 features per (turn, track) pair: per-branch rank/score/margin/hit features, pool-normalized z-scores and percentiles, content and behavioral cosines, session/artist-affinity counters, tag and lexical overlap, popularity, extracted-state categoricals, and a constraint sidecar of played/rejected/violation flags. Table 2 groups them by family with each family's share of total split gain in the deployed model; one feature — the bi-encoder cosine `b1_cos` — carries 59.0% of all gain. After scoring, two rule-based guards apply: an exact-track pin when the user names a specific track, and a final-artist constraint check.

The deployed model is *goal-free*: no conversation-goal annotation enters it as a feature, since Blind-B removes them. Its training data is the development split only — the same 8,000 turns used

for offline evaluation (~20.6M candidate rows). Per-turn state extraction runs through a paid LLM API, so we never built features over the ~121.6k-turn training split; Section 4.1 discusses what that reuse cost us. Training treats the per-turn ground-truth track as a binary label under a LambdaRank objective.[1] Because the labels are noisy (Section 4.2), we down-weight a turn when the *next* turn contradicts it: ×0.3 when the user rejects the labelled track or the goal-progress annotation says it did not help, ×0.6 when they reject only its artist, floored at 0.2. We chose these magnitudes by hand and did not tune them. That annotation shapes training labels — these weights and the bi-encoder positives of Section 2.4 — but is never itself a feature.

2.4 The bi-encoder feature b1_cos

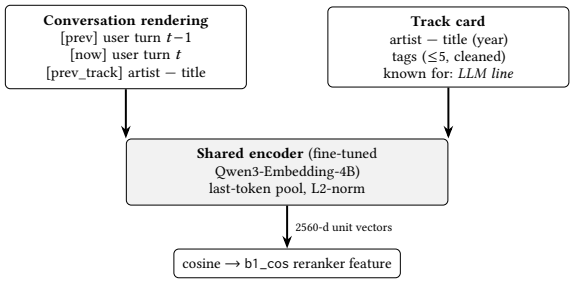


Figure 3: The b1 bi-encoder: one shared fine-tuned encoder embeds the conversation and the track card; their cosine is the reranker's dominant feature.

The reranker's dominant feature comes from a two-tower bi-encoder fine-tuned from Qwen3-Embedding-4B [19]. The query tower renders the conversation compactly behind a retrieval instruction prefix and the document tower renders a track card (both formats in Figure 3); the card's one-line LLM-written "known for" artist description imports static world knowledge from larger models into the embedding space. One shared encoder E_θ serves both towers; with $v_x = E_\theta(x)/\|E_\theta(x)\|_2$ (last-token pooling, 2560-d), the feature is the cosine $\texttt{b1\_cos}(q,d) = v_q^\top v_d$. Training minimizes the in-batch contrastive loss [13] over batches of positive pairs (q_i, d_i^+) with four mined hard negatives per positive,

$$\mathcal{L} = -\frac{1}{B}\sum_i \log \frac{\exp(v_{q_i}^\top v_{d_i^+}/\tau)}{\sum_{d\in\mathcal{D}_i}\exp(v_{q_i}^\top v_d/\tau)}, \qquad \tau = \tfrac{1}{20},$$

where $\mathcal{D}_i$ contains the in-batch documents plus the hard negatives. Positives are conservative: the 53,885 (turn, track) pairs whose next-turn goal-progress annotation says the recommendation *moved the session forward*. Document vectors for all 47,071 tracks are precomputed into the collection; the conversation vector is encoded live with caching.

We deploy the bi-encoder as a reranker feature only. In an earlier experiment we tried it as a twelfth retrieval branch, where it added little to the union's coverage and was dropped. As a feature its cosine dominates the fitted model — 59.0% of split gain (Table 2) — but that is a diagnostic of the fit, not a measure of recommendation quality; its incremental effect is much smaller, moving leakage-safe out-of-fold nDCG@20 from 0.1970 to 0.2032 (Table 3). We did not isolate which parts of the rendering mattered, or test whether the effect carries to the blind splits.

[1]Learning rate 0.025, 127 leaves, truncation 200, 5-fold user-grouped cross-validation.

Table 3: Development split. The out-of-fold rows are leakage-safe. The in-sample row is a historical goal-aware v10 model without b1_cos and is not directly comparable.

System (how measured)	nDCG@20
Weighted RRF fusion (no training)	0.1492
LambdaMART, out-of-fold CV	0.1970
+ b1_cos (deployed feature set)	0.2032
Historical v10 LambdaMART, in-sample	0.3844

Official Blind-B nDCG@20: 0.2537 — above the out-of-fold estimates and far below the in-sample replay.

Table 4: Official CodaBench scores. Section 1 defines the composite. The Blind-A row is our final development-phase submission.

Split	nDCG	Cat.	Lex.	Judge	Comp.
Blind-A (dev)	0.4380	0.0313	0.8028	4.70	0.5799
Blind-B (final)	0.2537	0.0315	0.7862	3.30	**0.3811**

2.5 Response generation

The response generator is `Qwen3-30B-A3B-Instruct-2507` [12, 18] at temperature 0: a single pass over the top-1 track rendered as a delimited XML block (≤10 tags) plus the *latest* extracted state. The deployed template's full style instruction is:

> *"Write 1–2 concise sentences about only the selected track. Prioritize the latest user request and extracted state over older conversation history. If the track is reasonably aligned, explain the fit with one specific supported reason. If it clearly conflicts with an explicit avoid/new-artist constraint, do not oversell it or blame the system; briefly frame the limitation and the closest supported reason."*

Over a frozen Blind-A retrieval output, a sweep of eight response variants (varying state conditioning, history depth, item rendering, concision, and the generator model) reached the 4.70 judge score of Table 4 without changing the recommendations; we selected on the leaderboard score. There is no multi-draft sampling, selection, checking, or repair pass.

Code, configs, trained models, cached extraction states, and a frozen-LLM-cache reproduction bundle that replays our submissions without API credentials are public.[2]

3 Results

The development split (1,000 sessions × 8 turns) has one ground-truth track per turn. Table 3 reports development-split numbers, separating leakage-safe estimates from in-sample ones; Table 4 reports the official leaderboard results.

The evaluated static-fusion and learned-ranker configurations used different candidate-pool depths, so the leakage-safe scores

[2]https://github.com/npatta01/music-crs-2026; models and data: https://huggingface.co/datasets/Npatta01/music-crs-repro-2026.

(0.149 versus 0.197–0.203) do not isolate the effect of learned reranking; Section 4.1 discusses the in-sample row.

4 Error Analysis

4.1 In-sample development evaluation

We never built features over the training split, so the development split served as both training and evaluation data and we had no clean held-out set. The numbers we selected on (0.38–0.46) were replays of turns the model had trained on; our out-of-fold estimates were much lower (0.197–0.203). The blind score (0.2537) came out between the two — above the out-of-fold estimates, well below the in-sample replays. We overfit our own selection process. Even the out-of-fold numbers are not independent: we chose the design against the same 8,000 turns.

4.2 Label disagreement and relabeling

Our own model shows anchoring behaviour, and we think it learns it from the labels. The per-turn ground truth is simply the track played next in the session, which is itself LLM-generated, and on many turns it does not fit the request: the label stays with the just-played artist even against an explicit request for someone else (Figure 4).

User: "…something with a similar chill, electronic vibe, but **from a different artist**?" (just played: *Bonobo*)
Ground truth: Bonobo — "Jets" *annotation: MOVES toward goal*

User (next turn): " 'Jets' is cool, but I was actually hoping for something **from a different artist this time**…"
Ground truth: Bonobo — "Pieces" *annotation: MOVES toward goal*

Figure 4: The anchoring pattern: two consecutive explicit different-artist requests, yet the label stays on Bonobo and the goal-progress annotation records both turns as moving toward the goal.

We tried to train on cleaner labels instead. Two inexpensive LLM judges (Gemma-4-26B and DeepSeek-V4-Flash) re-judged every labelable training and development turn (113,393) on two axes — whether the user asked for a different artist, and whether the track fits the request. They disagreed on about 20% of the 106,393 training turns, which Claude Opus arbitrated. Across the same turns they did not settle on the played track being a good fit for 61%, and 17% (18,222) are anchoring cases. Training on the relabeled data lowered our development score — but that score uses the original labels, so it could not have favoured the relabeling. We kept the original labels. Our judges may be wrong too; we release the relabeling so others can check it.

4.3 Failure cases

We used two label-free diagnostics on the submitted Blind-B run. A single LLM judge (DeepSeek-V4-Flash) rated 78% of turns weak-or-bad. Separately, a deterministic trace audit checked constraints and searched the logged candidate pool using positive-term metadata matches plus popularity; Table 5 diagnoses 32 turns. These are examples and audit categories, not measures of what each cost us — Blind-B exposes no relevance labels, so only the validation failure of Section 4.1 is actually measured.

Table 5: Deterministic label-free trace audit of the submitted Blind-B run. Categories overlap: 43 diagnoses across 32 distinct turns.

Diagnosis	Turns
Heuristic alternative below top-20	31
Under-applied constraint	8
State-extraction miss	3
Entity-resolution miss	1

- **The constraint was under-applied.** Asked for "the Kamelot track from *Silverthorn*," the system matched the artist but not the album, returning *Fallen Star* from Kamelot's *Haven* (*Silverthorn* is in the catalog); asked to branch out to other artists, it stayed on the current one (10 of the top 20 in one Ryan Adams session).
- **Nothing in our data encodes the request.** An exact "126.70 bpm and G minor" (no tempo or key field), or "a song from the *Breaking Bad* soundtrack" (no film/TV metadata), which no branch can match on.
- **Response generation.** The single-pass generator did not properly consider the query — in particular an exact-track request for a title we do not have. On several turns a named track was read correctly as an exact-track request but is absent from the 47,071-track catalog; rather than flag it unavailable, the generator described a different song by the same artist as if it were the one asked for — for " 'Watercolors' by Pat Metheny," it returned Metheny's "Alfie."

The limitation we keep returning to is that too little of the user's intent became executable constraints. The state records what the user wants, but few fields became filters or ranked evidence — each branch consumed a slice of the state, and tag matching fell back to exact strings outside the BM25 branch — while no co-occurrence, transition, or live-reasoning lane existed to compensate. Noisy candidate pools therefore reached the reranker; on 31 flagged turns, the audit heuristic found a non-violating alternative below our top-20, selected by positive-term metadata matches plus popularity — Table 5's largest bucket, but an audit signal rather than a relevance judgment.

Our analysis has its own limits. The relabeling was never checked against human judgments, the Blind-B quality audit used a single LLM judge, the trace audit was heuristic, the conversations are LLM-generated rather than observed listening sessions, and neither blind split exposes relevance labels — so the blind-set failure rates and the relabeling rates are diagnostic rather than ground truth. The development-split and leaderboard numbers in Section 3 are separate measurements.

5 Conclusion

We documented a state-compiled retrieve-then-rerank conversational music recommender and finished 29th of 40. Four gaps stand out among the failures we could diagnose:

(1) model selection relied on in-sample development replays, and we had no clean held-out set to check them against;

(2) the extracted conversation state was rich, but too little of it became executable filters or ranking evidence, so noisy candidate pools reached the reranker;
(3) for retrieval and ranking we had little information about tracks and artists beyond the provided catalog metadata — no external source, and no LLM-generated knowledge beyond the one-line artist description — so requests naming anything the catalog does not encode had nothing to match against;
(4) responses were generated in a single unchecked pass that could reliably flag neither unavailable tracks nor unmet constraints.

Separately, we were unsure about many of the training labels, and re-judged the 113,393 labelable turns with LLM judges (Section 4.2). We release the pipeline, the models, a replay bundle, and the re-judged labels [11].

Acknowledgments

Generative AI tools (OpenAI Codex/ChatGPT) assisted with coding, debugging, analysis of experimental outputs, drafting and editing this paper, and camera-ready formatting and reference normalization. All AI-generated content was reviewed by the authors, who take responsibility for the final work.

References

[1] Christopher J. C. Burges. 2010. *From RankNet to LambdaRank to LambdaMART: An Overview.* Technical Report MSR-TR-2010-82. Microsoft Research. https://www.microsoft.com/en-us/research/publication/from-ranknet-to-lambdarank-to-lambdamart-an-overview/
[2] Keunwoo Choi, Seungheon Doh, and Juhan Nam. 2025. TalkPlayData 2: An Agentic Synthetic Data Pipeline for Multimodal Conversational Music Recommendation. arXiv:2509.09685 doi:10.48550/arXiv.2509.09685
[3] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 758–759. doi:10.1145/1571941.1572114
[4] DeepSeek-AI. 2026. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence. arXiv:2606.19348 [cs.CL] doi:10.48550/arXiv.2606.19348
[5] Seungheon Doh, Keunwoo Choi, and Juhan Nam. 2025. TalkPlay: Multimodal Music Recommendation with Large Language Models. arXiv:2502.13713 doi:10.48550/arXiv.2502.13713
[6] Seungheon Doh, Sergio Oramas, Bruno Sguerra, Abhinav Bohra, Claudio Pomo, and Francesco Barile. 2026. RecSys Challenge 2026: Conversational Music Recommendation. In *Proceedings of the 20th ACM Conference on Recommender Systems, RecSys 2026, Minneapolis, MN, USA, September 27 - October 02, 2026 (RecSys '26)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3773078.3841245
[7] Seungheon Doh, Bruno Sguerra, Sergio Oramas, Elena V. Epure, and Juhan Nam. 2026. LLM-as-a-Judge for Evaluating System Responses in Conversational Music Recommendation. In *Proceedings of the 20th ACM Conference on Recommender Systems* (Minneapolis, MN, USA) *(RecSys '26)*. Association for Computing Machinery, New York, NY, USA. doi:10.48550/arXiv.2607.25640
[8] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A Survey on Conversational Recommender Systems. *Comput. Surveys* 54, 5, Article 105 (2021), 36 pages. doi:10.1145/3453154
[9] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., Red Hook, NY, USA, 3146–3154. https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html
[10] LanceDB. 2026. LanceDB: Multimodal Lakehouse for AI. https://docs.lancedb.com/. Accessed August 11, 2026.
[11] Nidhin Pattaniyil, Semih Yagli, and Tanwir Zaman. 2026. music-crs-2026: Code, Models, and Reproduction Bundle. https://github.com/npatta01/music-crs-2026. Accessed August 11, 2026.
[12] Qwen Team. 2025. Qwen3-30B-A3B-Instruct-2507. Hugging Face model card. Accessed August 19, 2026. https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507
[13] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 3982–3992. doi:10.18653/v1/D19-1410
[14] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI '09)*. AUAI Press, Arlington, VA, USA, 452–461. https://auai.org/uai2009/papers/UAI2009_0139_48141db02b9f0b02bc7158819ebfa2c7.pdf
[15] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* 3, 4 (2009), 333–389. doi:10.1561/1500000019
[16] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. arXiv:2502.14786 doi:10.48550/arXiv.2502.14786
[17] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*. IEEE, Piscataway, NJ, USA, 1–5. doi:10.1109/ICASSP49357.2023.10095969
[18] An Yang et al. 2025. Qwen3 Technical Report. arXiv:2505.09388 doi:10.48550/arXiv.2505.09388
[19] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. arXiv:2506.05176 doi:10.48550/arXiv.2506.05176