Personalized Retrieval for YouTube Recommendations in a RAG System

Chantelle Chan

chan.ya@northeastern.edu

Dec 11, 2025

1. Overview

This model is a personalized retrieval module for space-related YouTube content, with the goal of understanding how user behavior can influence search and recommendation results. Instead of building a full end-to-end Retrieval-Augmented Generation (RAG) pipeline, the work focuses on the retrieval layer that a RAG system would rely on:

Given a user and a query, how can the system retrieve and rank the most relevant videos?

The starting point is a semantic search baseline using sentence embeddings of video titles and descriptions, built with a pre-trained Sentence-BERT (SentenceTransformer) model [2]. These embeddings are then extended with additional signals: view-count popularity and a personalization score derived from each user's viewing profile. Three ranking strategies are evaluated—semantic-only, semantic + popularity, and semantic + personalization—using Precision@5 and Recall@5, two standard measures in information retrieval [1]. A small qualitative case study is also conducted, in which two simulated users issue the same broad query ("space news") to examine whether the personalization layer produces different, preference-aligned rankings.

Overall, the findings are mixed but offer some useful perspectives. On the positive side, the personalization layer clearly behaves like a personalized system: different users receive different top-5 results that match their interests. However, quantitative ranking metrics on a small evaluation set show that this first version of the personalized re-ranker does not outperform a strong semantic baseline. In that sense, the work is best viewed as a proof of concept for personalized retrieval rather than a final optimization.

2. Data and Setup

The base collection of texts is a real dataset of space- and astronomy–related YouTube videos scraped between 2022 and 2025. Each record includes:

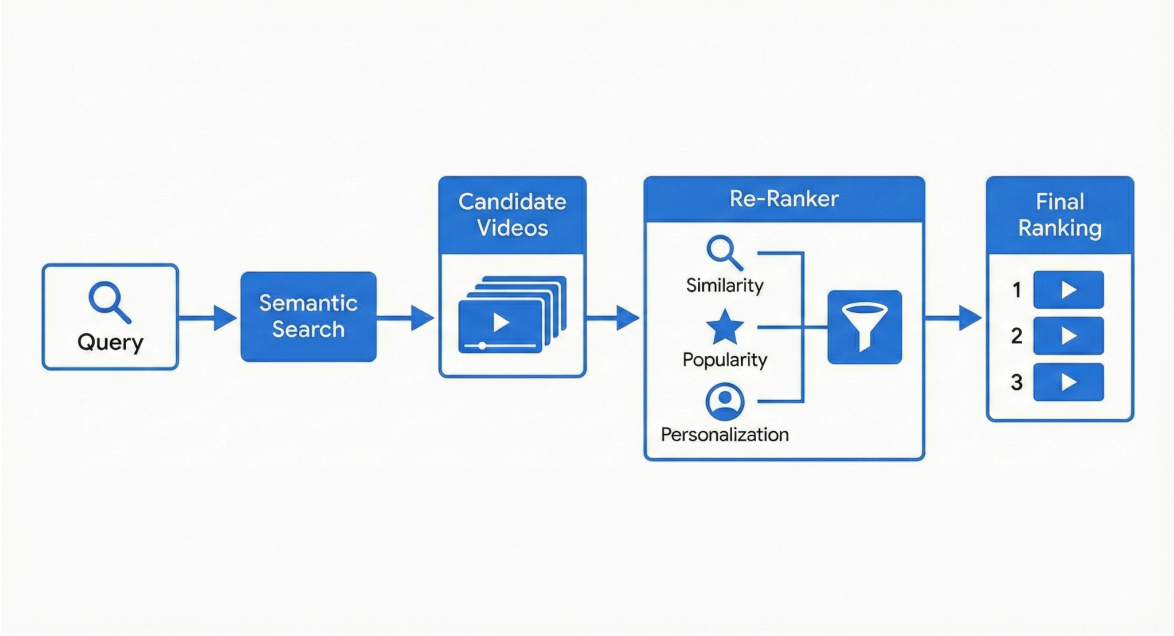
- Title and description

- Tags
- View count, like count, and comment count
- Duration and a derived is_short flag
- Video and thumbnail URLs

A Sentence-BERT model (all-MiniLM-L6-v2) is used to encode each video's combined title, description, and tags into a dense embedding vector, suitable for semantic search and similarity-based retrieval [2].
Because real user watch logs are not available, the project generates a synthetic interaction dataset. This dataset links simulated user IDs to video IDs with interaction types such as "view," "like," "share," and "comment." The synthetic log is not intended to perfectly mimic real user behavior, but it provides a controlled setting to test how personalization mechanisms would operate at the retrieval layer—the same layer where real interaction data would be used in practice.

3. Personalized Retrieval Design

The personalized retriever builds on a semantic search baseline and augments it with user profiles and additional ranking signals. A pre-trained Sentence-BERT model [2] is used to generate embeddings for video titles, descriptions, tags, and queries. This choice avoids the cost of training a text encoder from scratch and allows the model to focus on the design and evaluation of the personalized retrieval and re-ranking layer. Fine-tuning the encoder on this domain-specific text collection or on user interaction data is left as future work and could further improve retrieval quality.



Video titles, descriptions, and tags are encoded into dense vectors using this Sentence-BERT model, and queries are embedded in the same space [2]. Cosine similarity between the query embedding and each video embedding is used to retrieve the top-k nearest neighbors via a simple k-NN index, forming the semantic baseline, which ignores user identity and optimizes only for query relevance in the traditional vector-space information retrieval framework [1].

To incorporate user behavior, each simulated user is represented by a user profile embedding obtained by averaging the embeddings of all videos that user has interacted with in the synthetic log. This average vector summarizes the dominant themes in the user's history (for example, black-hole–heavy versus exoplanet-heavy viewing) and follows common practice in content-based and neural recommender models when a separate user encoder is not trained [2], [5].

On top of this, a re-ranking function is defined. For every candidate video, the system computes: (i) semantic similarity between the query and the video; (ii) a log-scaled view_count term that provides a small popularity boost, reflecting engagement-driven signals in large-scale recommenders such as YouTube's two-stage system [3]; and (iii) a personalization score given by the cosine similarity between the user profile embedding and the video embedding, measuring alignment with the user's historical preferences [5]. In addition, simple keyword-based boosts are applied when a video's tags overlap with user-specified interest keywords. These components are combined into a single re_rank_score, and three ranking modes are compared in the experiments: (1) semantic baseline (semantic similarity only), (2) view-count re-ranked (semantic similarity plus popularity), and (3) personalized (semantic similarity, popularity, personalization score, and keyword boosts).

4. Experiments and Results

An important design choice in the personalized retriever is how strongly the personalization score should influence the final ranking. In this version, that effect is controlled via numeric weights in the re-ranking function, and a small grid search is performed over a "personalization strength" parameter (alpha-like behavior) to observe how it affects metrics.
To evaluate ranking quality, the system uses Precision@5 and Recall@5. Precision@k measures the fraction of retrieved items in the top-k that are relevant, while Recall@k measures the fraction of all relevant items that appear in the top-k list [1]. Rather than labeling the entire dataset, a small dictionary is defined that explicitly maps each query to a list of video titles considered relevant for that query, serving as the ground truth for evaluation. This ground-truth dictionary is constructed for four representative queries:

- "how do black holes form?"
- "James Webb Space Telescope discoveries"
- "Explain the process of planet formation outside our solar system"
- "what is dark matter?"

For each query, a small set of titles in the dataset is selected as **ground-truth relevant items**. The notebook function evaluate_search_performance then (1) runs a given search function (semantic

baseline, view-count re-ranked, or personalized), (2) takes the top-5 results, and (3) counts how many of the top-5 overlap with the ground-truth titles (Precision@5) and how many of the ground-truth titles appear in the top-5 (Recall@5).

4.1 Quantitative Results: Baseline vs. Popularity vs. Personalization

Table 1 summarizes the evaluation results for the four queries and three ranking strategies.

Table 1. Precision@5 and Recall@5 for four queries and three ranking strategies.

Query	Strategy	Precision@5	Recall@5
how do black holes form?	Semantic baseline	0.6	1
	View-count re-ranked	0.40	0.67
	Hybrid personalized	0.20	0.33
James Webb Space Telescope discoveries	Semantic baseline	0.60	1
	View-count re-ranked	0.4	0.67
	Hybrid personalized	0.20	0.33
Explain the process of planet formation outside our solar system	Semantic baseline	0.60	1
	View-count re-ranked	0.20	0.33
	Hybrid personalized	0	0
what is dark matter?	Semantic baseline	0.4	1
	View-count re-ranked	0.4	1
	Hybrid personalized	0.20	0.50
Average over 4 queries	Semantic baseline	0.55	1
	View-count re-ranked	0.35	0.67
	Hybrid personalized	0.15	0.29

For this small evaluation set, both the view-count re-ranked and personalized rankings lower Precision@5 and Recall@5 compared to the semantic baseline. This highlights the importance of testing new ranking components against a strong baseline, which is standard practice in information retrieval and recommender evaluation [1], [5].

4.2 Qualitative Case Study: Same Query, Different Users

While the aggregate metrics show that this first version of personalization does not yet improve ranking quality, they do not reveal how behavior differs at the user level. To better understand system behavior, a small qualitative case study is performed.

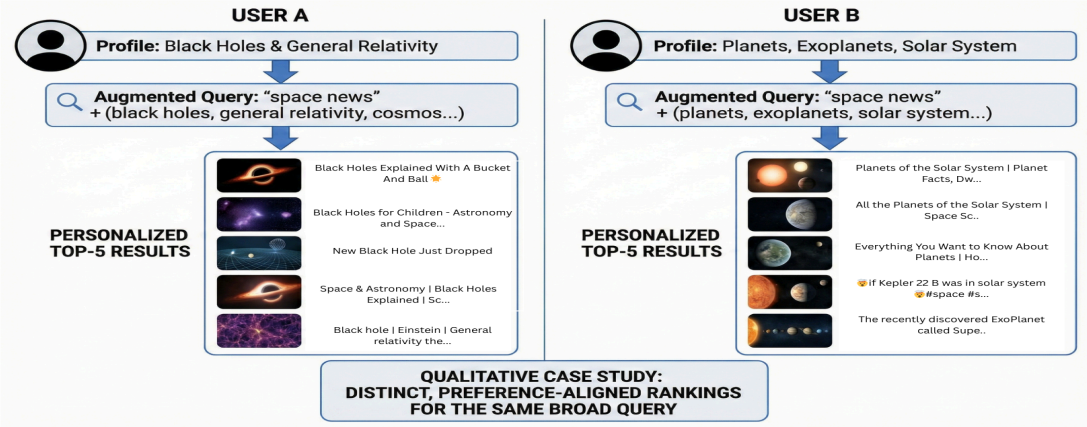
Two simulated users are defined:

- User A: profile dominated by black holes and general relativity

- User B: profile dominated by planets, exoplanets, and the solar system

Both users submit the same broad query, “space news.” Before retrieval, the query is augmented with terms reflecting each user’s interests (for example, “black holes, general relativity, cosmos…” for User A and “planets, exoplanets, solar system…” for User B). The personalized search function is then run and the top-5 results are inspected.

For User A, all top-5 results are videos about black holes or closely related topics. Semantic similarities and personalization scores are both high, and the combined re-ranking score favors these videos over more generic “space news” content. For User B, the top-5 results shift dramatically to focus on planets and exoplanets—such as planets of the solar system, everything you want to know about planets, and exoplanets, etc. Personalization scores are relatively high for these videos, and the re-ranking function lifts them above more general space content.



The results also suggest an important trade-off between strict semantic relevance and user-aligned personalization. In the hybrid model, user preferences are injected through both query augmentation and a personalization score, which can shift rankings toward the user’s long-term tastes and away from the narrow intent of a specific query. This is visible in cases such as the planet-formation query, where the personalized variant fails to retrieve any of the hand-labeled relevant videos, indicating that naive personalization can dilute query specificity. Future work should therefore decouple query intent from user preference signals and explore ways to optimize both accuracy and personalization simultaneously.

5. Summary of Implementation

The model implements a personalized retrieval module on top of a semantic search baseline. Video titles, descriptions, and tags are encoded with a Sentence-BERT model and indexed for cosine-similarity search, while synthetic interaction data are used to build user profile embeddings by averaging the embeddings of interacted videos. A hybrid re-ranking function then combines semantic similarity, a log-scaled popularity term, a personalization score, and simple keyword-based boosts into a single score, which is used to compare three ranking modes (semantic baseline, view-count re-ranked, and hybrid personalized) in the subsequent experiments.

6. Limitations and Future Work

Several limitations:

- User profiles are manually constructed from a small set of chosen videos, so they reflect simulated behavior rather than real watch histories.
- Profiles are built by simple averaging of past item embeddings, ignoring sequence effects, recency, and session structure.
- Personalization strength is controlled by a single global parameter (alpha) rather than a learned, context-aware mechanism.
- The current system focuses only on retrieval and re-ranking; it does not include any LLM components such as query rewriting or LLM-as-judge evaluation.

These limitations point to clear directions for future work:

- Incorporate sequence models (e.g., RNNs, Transformers, or Markov chains) to better model user viewing behavior over time.
- Learn query- and user-specific personalization strengths, possibly through a small learned model or LLM-based heuristics, fitting naturally into a hybrid recommender setting [5].
- Evaluate the system on more queries, more realistic interactions, or actual user logs if available.
- Integrate the personalized retrieval module into a full RAG pipeline, where the top-k retrieved videos are used as context for an LLM that generates grounded, user-tailored explanations or recommendations [4].

7. Conclusion

This model developed and analyzed a personalized retrieval module for space-related YouTube content, implemented as the retrieval component of a future Personalized Retrieval-Augmented Generation (RAG) system. Semantic search over a real video dataset was combined with user profile embeddings computed from synthetic interactions, and a hybrid re-ranking function was

designed to integrate semantic relevance, popularity, and personalization. The module was evaluated quantitatively, using Precision@5 and Recall@5, and qualitatively, through a case study where different users issued the same broad query.

The current version of personalization does not yet improve upon a strong semantic baseline on the small evaluation set, but it does produce clear user-specific behavior: different users receive distinct, preference-aligned rankings for the same query (for example, black-hole–focused results versus exoplanet-focused results for “space news”). This positions the personalization layer as a functional prototype for exploring personalized retrieval, rather than a fully optimized production system. The work clarifies how retrieval, user modeling, and re-ranking can be combined in a RAG-style system and highlights the importance of careful, metric-based evaluation when adding personalization to an already strong retriever. Overall, the behavior of this personalized retrieval module is consistent with how large-scale systems like YouTube’s recommendation engine are described in the literature: rather than returning a single “best” video for everyone, they compute rankings per user, aiming to surface the video that is most relevant for a specific viewer at a specific moment, given their viewing history and current intent [3].

References

[1] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[2] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019, pp. 3982–3992.

[3] P. Covington, J. Adams, and E. Sargin, “Deep Neural Networks for YouTube Recommendations,” in Proc. 10th ACM Conf. Recommender Systems (RecSys), Boston, MA, USA, 2016, pp. 191–198, doi: 10.1145/2959100.2959190.

[4] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Adv. Neural Inf. Process. Syst., vol. 33, pp. 9459–9474, 2020.

[5] R. Burke, “Hybrid Recommender Systems: Survey and Experiments,” User Modeling and User-Adapted Interaction, vol. 12, no. 4, pp. 331–370, 2002.