Hybrid Movie Recommendation System: Combining Probabilistic Matrix Factorization with Fine-Tuned LLM Embeddings

Bastien HoTTELET* Hamady GACKOU*
*Master 2 – Machine Learning for Data Science (MLDS)
Université Paris Cité, UFR Sciences Fondamentales et Biomédicales, Paris, France
Supervised by: Aghiles Salah, Senior Research Scientist, Rakuten Institute of Technology

***Abstract*—This project investigates a hybrid movie recommendation system that combines Probabilistic Matrix Factorization (PMF) with modern Large Language Model (LLM) embeddings extracted from movie content. Using the MovieLens *ml-latest-small* dataset (∼100k ratings, ∼9k movies, 610 users), we compare a classical collaborative filtering baseline (PMF) with several hybrid variants that integrate MiniLM and Qwen3-Embedding-4B, both in base and fine-tuned forms. To adapt generic semantic embeddings to user preferences, we apply contrastive learning on positive and negative movie pairs derived from user histories, and we fine-tune Qwen3-4B using LoRA with 4-bit quantization for efficiency. Results on Precision@25, Recall@25 and NDCG@25 show that the best hybrid model (PMF + Qwen3-4B fine-tuned) outperforms pure PMF by up to 12.6% in Precision@25, confirming that LLM embeddings can substantially improve recommendation quality when aligned with interaction data.**

***Index Terms*—Recommender systems, matrix factorization, large language models, contrastive learning, LoRA, MovieLens.**

I. Introduction

Recommender systems are central to modern content platforms, where users must navigate catalogs of thousands of items. Collaborative Filtering (CF), and in particular matrix factorization, has become a standard approach due to its ability to capture latent user and item factors from implicit or explicit feedback. However, CF suffers from two major issues: (i) extreme sparsity of the user–item matrix, and (ii) the cold-start problem for new or rarely rated movies.

In parallel, recent advances in Natural Language Processing (NLP) have produced powerful Large Language Models (LLMs) capable of encoding rich semantic information from text. In movie recommender systems, this textual information comes from titles, genres and user-generated tags. Yet, generic LLM embeddings primarily reflect semantic similarity (e.g., “two movies are both war documentaries”) rather than similarity in user preferences.

The objective of this work is to design and evaluate a hybrid movie recommender that:

- retains the strengths of PMF on dense parts of the rating matrix;
- leverages LLM-based content embeddings to better handle sparse and long-tail items;
- re-aligns the embedding space with user preferences through contrastive fine-tuning, while remaining computationally feasible using LoRA.

II. Dataset and Problem Setting

We use the public MovieLens *ml-latest-small* dataset, which contains 100 836 ratings on a 0.5–5.0 scale, associated with 9 724 movies and 610 users. The rating matrix is about 98.3% sparse, making the cold-start and sparsity issues particularly relevant.

For each movie we construct a textual description by concatenating:

- the title (year removed),
- the list of genres,
- all user tags associated with the movie.

This text is then encoded into dense vectors by MiniLM and Qwen3-Embedding-4B.

The task is Top-K recommendation: for each user we predict a ranked list of unseen movies. A movie is considered relevant if the rating in the test set is at least 4.0. We report Precision@25, Recall@25 and NDCG@25 averaged over users.

III. Hybrid Model

A. Collaborative Branch: Probabilistic Matrix Factorization

PMF factorizes the observed rating matrix R into user and item latent factors:

$$\min_{U,V} \sum_{(u,i)\in\Omega} (R_{ui} - U_u^\top V_i)^2 + \lambda(\|U\|_F^2 + \|V\|_F^2), \tag{1}$$

where $U_u, V_i \in \mathbb{R}^d$ and Ω is the set of observed ratings. The PMF score is given by $s_{u,i}^{\mathrm{MF}} = U_u^\top V_i$, then normalized by a sigmoid.

B. Content Branch: LLM Embeddings

From the constructed movie text, we compute base embeddings:

- MiniLM Sentence Transformer (384 dimensions),
- Qwen3-Embedding-4B (2560 dimensions).

To represent a user, we aggregate embeddings of positively rated movies:

$$E_u = \frac{1}{Z_u} \sum_{i\in\mathcal{I}_u^+} w_{ui} e_i, \tag{2}$$

where e_i is the embedding of movie i, $\mathcal{I}_u^+$ are movies rated ≥ 4.0 by user u, and w_{ui} is a rating-based weight. Content scores are defined as cosine similarities $s_{u,i}^{\mathrm{C}} = \cos(E_u, e_i)$.

C. Contrastive Fine-Tuning with LoRA

Base embeddings do not necessarily reflect user preferences. We therefore fine-tune Qwen3-4B using a contrastive loss: positive pairs are movies co-liked by a user; negative pairs are movies liked and disliked by the same user.

To make this feasible on academic hardware, we apply LoRA adapters on attention projections and load the model in 4-bit quantization. Only about 0.14% of parameters are trainable, while the backbone remains frozen. We train for three epochs with AdamW and early stopping on validation NDCG@25.

D. Score Fusion

The final hybrid score for user u and movie i is a convex combination:

$$\hat{y}_{u,i} = \alpha\, \sigma(s^{\mathrm{MF}}_{u,i}) + (1-\alpha)\, s^{\mathrm{C}}_{u,i}, \tag{3}$$

where σ is a sigmoid and $\alpha \in [0,1]$. Grid search indicates that $\alpha = 0.85$ provides the best trade-off between collaborative and content signals.

IV. Experimental Protocol

We use a user-level 80/20 split: for each user, 80% of ratings are used for training and 20% for testing, ensuring that all users appear in both sets.

We evaluate the following methods:

- **MF Only** (PMF baseline),
- **MF + MiniLM (Base)**,
- **MF + MiniLM (Fine-Tuned)**,
- **MF + Qwen3 (Base)**,
- **MF + Qwen3 (Fine-Tuned with LoRA)**.

For each user we compute the Top-25 recommendations and evaluate Precision@25, Recall@25 and NDCG@25.

V. Results and Discussion

Table I summarizes the performance of all methods.

TABLE I
Performance on MovieLens *ml-latest-small*.

Method	**Prec@25**	**Rec@25**	**NDCG@25**
MF Only	0.1232	0.1951	0.2115
MF + MiniLM (Base)	0.1187	0.1879	0.2050
MF + MiniLM (FT)	0.1225	0.1941	0.2100
MF + Qwen3 (Base)	0.1270	0.2014	0.2168
MF + Qwen3 (FT + LoRA)	**0.1387**	**0.2209**	**0.2349**

Several trends emerge:

- Larger embeddings significantly help: Qwen3-4B (Base) already improves over MF and MiniLM, suggesting that richer content representations translate into better recommendations.
- Fine-tuning is crucial: the fine-tuned Qwen3 model delivers a 12.6% relative gain in Precision@25 compared to MF, and consistent gains on Recall and NDCG.
- Hybridization matters: the best performance is obtained when combining MF and content scores. Pure content-based ranking suffers from popularity bias and may over-recommend semantically similar but unpopular movies, while pure MF ignores semantic proximity among rarely rated items.

Qualitative inspection for heavy users confirms these observations: for cinephile profiles, the hybrid model tends to surface stylistically consistent movies (e.g., psychological dramas and auteur cinema) that are underrepresented in the MF-only top list.

VI. Conclusion

This project demonstrates that hybrid recommenders that combine PMF with fine-tuned LLM embeddings can substantially improve movie recommendation quality on sparse datasets. By leveraging LoRA and 4-bit quantization, we make the fine-tuning of a 4B-parameter embedding model computationally accessible in an academic setting.

Future work includes testing more advanced fusion mechanisms, incorporating multimodal content (posters and trailers), and extending the approach to other domains such as music or books.

Acknowledgment

The authors would like to thank Aghiles Salah for his guidance and feedback throughout the project, as well as the teaching staff of the M2 Machine Learning for Data Science program at Université Paris Cité.

References

[1] R. Salakhutdinov and A. Mnih, “Probabilistic Matrix Factorization,” *Advances in Neural Information Processing Systems*, 2007.

[2] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” *EMNLP*, 2019.

[3] E. Hu *et al.*, “LoRA: Low-Rank Adaptation of Large Language Models,” *arXiv preprint arXiv:2106.09685*, 2021.

[4] F. M. Harper and J. A. Konstan, “The MovieLens Datasets: History and Context,” *ACM Transactions on Interactive Intelligent Systems*, 2015.