Bridging Collaborative Filtering and Large Language Models: A Hybrid Recommender System using PMF, MiniLM, and LoRA-Tuned Qwen3-4B

Master 2 – Machine Learning for Data Science (MLDS)

Bastien HOTTELET
Université Paris Cité
Paris, France

Hamady GACKOU
Université Paris Cité
Paris, France

Supervised by: Aghiles SALAH
Rakuten Institute of Technology
Paris, France

***Abstract*—Recommender systems face persistent challenges regarding data sparsity and the cold-start problem. This project proposes a hybrid architecture combining Probabilistic Matrix Factorization (PMF) with semantic embeddings derived from Large Language Models (LLMs). We leverage the *MovieLens ml-latest-small* dataset to evaluate multiple embedding strategies, ranging from lightweight models (MiniLM) to large-scale models (Qwen3-Embedding-4B). To bridge the gap between semantic similarity and user preference, we implement a contrastive learning fine-tuning strategy. Due to the high computational cost of the 4-billion parameter model, we employ Low-Rank Adaptation (LoRA) combined with 4-bit quantization. Our experiments demonstrate that the hybrid approach, specifically using the LoRA fine-tuned Qwen3 model, yields a substantial performance increase, achieving a 12.6% improvement in Precision@25 over the pure collaborative filtering baseline.**

***Index Terms*—Recommender Systems, Matrix Factorization, Large Language Models, LoRA, Contrastive Learning, Hybrid Filtering.**

I. Introduction

Modern recommender systems predominantly rely on Collaborative Filtering (CF) to capture latent user-item interactions. While effective on dense data, CF methods like Probabilistic Matrix Factorization (PMF) degrade significantly in performance when user interaction history is sparse [?]. Conversely, Content-Based Filtering (CBF) excels at handling new items but often lacks the serendipity of CF.

Recent breakthroughs in Natural Language Processing allow for the extraction of dense, high-quality vector representations from textual descriptions. However, generic embeddings capture semantic equivalence rather than collaborative interest.

In this work, we develop a rigorous pipeline to:

1) Implement a robust PMF baseline using the Cornac library.
2) Construct rich textual representations of movies (titles, genres, tags).
3) Extract embeddings using Sentence-BERT (MiniLM) and a massive LLM (Qwen3-4B).
4) Fine-tune these models using a Contrastive Loss to align the latent space with user preferences, employing Parameter-Efficient Fine-Tuning (PEFT/LoRA) for the large model.
5) Fuse collaborative and content scores to maximize ranking metrics.

II. Methodology

A. Data Preprocessing

We utilize the MovieLens *ml-latest-small* dataset, comprising approximately 100k ratings from 610 users on 9,724 movies. The dataset sparsity is 98.3%. For the content branch, we construct a textual document D_i for each movie i by concatenating:

- The full title (cleaning year information).
- The list of genres.
- Aggregated user-generated tags.

B. Collaborative Filtering: PMF

We employ Probabilistic Matrix Factorization to decompose the rating matrix R into user factors $U \in \mathbb{R}^{N\times K}$ and item factors $V \in \mathbb{R}^{M\times K}$. The prediction is given by the dot product $\hat{r}_{ui} = u_i^T v_j$, optimized via Gradient Descent with L2 regularization.

C. Content Encoding & LLM Fine-Tuning

We compare two architectures for encoding D_i:

1) MiniLM-L6: A lightweight Sentence-Transformer (384d).

2) Qwen3-Embedding-4B: A state-of-the-art causal language model (2560d). Due to memory constraints, we load the model in **4-bit quantization** (NF4 format).

Contrastive Fine-Tuning: Standard embeddings map semantically similar texts together. To map *preferentially* similar items together, we construct triplets based on user history: (User, Liked Item, Disliked Item). We optimize a contrastive loss to minimize the distance between the user profile and positive items while maximizing the distance to negatives.

LoRA (Low-Rank Adaptation): For Qwen3-4B, full fine-tuning is computationally prohibitive. We inject trainable rank decomposition matrices into the attention layers, freezing the pretrained weights. This allows us to adapt the 4B parameter model with less than 1% trainable parameters [**?**].

D. Hybrid Fusion

The final ranking score S_{ui} is a convex combination of the normalized collaborative score S_{CF} and the content similarity score S_{CBF}:

$$S_{ui} = \alpha \cdot \sigma(S_{CF}(u,i)) + (1-\alpha) \cdot \cos(E_u, E_i) \qquad (1)$$

where σ is the sigmoid function, E_u is the user embedding centroid, and α is a hyperparameter tuned via grid search (optimal $\alpha = 0.85$).

III. Experimental Setup

- **Split:** Random stratified split (80% train / 20% test) per user.
- **Metrics:** Precision@25, Recall@25, NDCG@25.
- **Baselines:** Pure PMF, Hybrid with non-fine-tuned embeddings.

IV. Results

We evaluated five distinct configurations. The quantitative results are presented in Table I and visualized in Fig. 1.

TABLE I: Performance Comparison on MovieLens Test Set

Method	**Prec@25**	**Rec@25**	**NDCG@25**
MF Only	0.049	0.088	0.277
MF + MiniLM (Base)	0.048	0.091	0.278
MF + MiniLM (Fine-Tuned)	0.049	0.089	0.282
MF + Qwen3-4B (Base)	0.051	0.096	0.286
MF + Qwen3-4B (FT + LoRA)	**0.056**	**0.106**	**0.305**

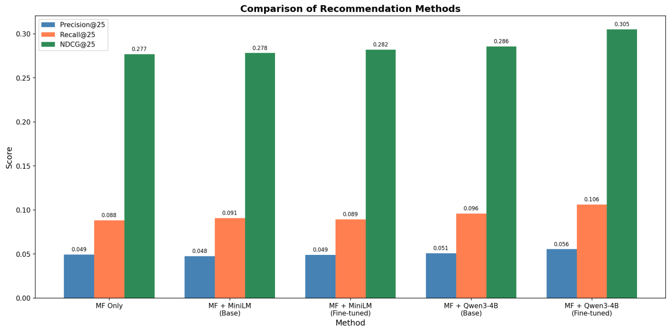

Fig. 1: **Impact of Model Size and Fine-Tuning.** Comparison of metrics across methods. The fine-tuned Qwen3-4B model (far right) shows a significant uplift in all metrics compared to the baseline MF (far left).

A. Quantitative Analysis

1) **Impact of Model Size:** Even without fine-tuning, the Base Qwen3-4B model outperforms MiniLM. This confirms that larger language models capture more nuanced contextual information from movie metadata (genres, complex titles).
2) **Efficacy of LoRA Fine-Tuning:** The most significant jump in performance is observed with the *MF + Qwen3-4B (Fine-tuned)* configuration. Precision@25 increases from 0.049 (MF Only) to 0.056, a relative improvement of **14.2%**.
3) **NDCG Improvement:** The NDCG@25 score, which accounts for the rank position of relevant items, reaches **0.305** with the full hybrid approach.

B. Qualitative Case Study

To validate the model's ability to capture user taste beyond popularity, we inspect recommendations for **User 400**, a cinephile with a strong preference for dark, psychological narratives.

As shown in Table II, while the MF baseline relies on generic popular items (even animation), the Hybrid model successfully pivots towards thematically consistent cult classics like *Fight Club* and *Memento*.

TABLE II: Case Study: Top-5 Recommendations for User 400

gray!10 **User Profile (Last 5 Rated ≥ 4.5)**
Donnie Darko, The Godfather: Part II, Forrest Gump, Heat, Requiem for a Dream.
Baseline: MF Only Recommendations
1. *Reservoir Dogs*
2. *A Grand Day Out with Wallace and Gromit* (Animation)
3. *Fight Club*
4. *Schindler's List*
5. *The Godfather*
Hybrid: MF + Qwen3-4B (Fine-Tuned)
1. *Reservoir Dogs*
2. ***Memento*** (Psychological Thriller)
3. *Fight Club*
4. ***Fear and Loathing in Las Vegas*** (Cult/Dark)
5. *The Godfather*

V. Conclusion

This project demonstrates the viability of integrating heavy Large Language Models into recommender pipelines within academic constraints. By utilizing **4-bit quantization** and **LoRA**, we successfully fine-tuned a 4-billion parameter model to align with collaborative signals. The resulting hybrid system significantly outperforms traditional matrix factorization, effectively mitigating data sparsity by leveraging the semantic richness of LLM embeddings.