LLM-Based Movie Recommendation Systems

Practical Work Report

Author: Tamas Scheidl
Date: 09.06.2025

1. Introduction

This practical work investigates the application of large language models (LLMs) for movie recommendation tasks using the MovieLens 1M dataset. The goal is to explore a range of prompting strategies and ranking methods that leverage semantic reasoning, surprise, diversity, and genre alignment, while balancing relevance and novelty.

The system builds on and adapts techniques from Delbar et al. (2024) [1] and extends the GitHub implementation Benchmark_RecLLM_Fairness. The project is implemented using Python and evaluated on a local Apple Silicon environment (MPS backend), using a Hugging Face implementation of the Phi-3.5-mini-instruct model.

2. Model Configuration

Attribute	Description
LLM	`microsoft/Phi-3.5-mini-instruct`
Architecture	Decoder-based (small LLM, ~1.8B params)
Platform	Local inference with Hugging Face
Device	Apple Mac (MPS backend)

3. Dataset

- Dataset: MovieLens 1M
- User-Movie Interactions: 1M ratings (userId, itemId, rating, timestamp)
- Metadata:
    - df_users_ml: demographics
    - df_items_ml: movie titles, genres
- Evaluation:

- 10 users × 120 candidate movies per user
- 5 movies as input history, remainder as ground truth

4. Methodology

Each model (S1–S7) is implemented as a separate Python script located in the scripts/ directory. All models take as input a short history of previously watched movies and output a top-10 ranked list of recommendations. Depending on the strategy, candidates are filtered by genre, popularity, or model scores.

The recommendations are then evaluated using both standard accuracy metrics and diversity indicators. Additional visualizations are produced via Plot.ipynb to support interpretation.

5. Model Architectures

ID	Name	Strategy Description
S1	Baseline	Input history → predict top 10
S2	Genre-Focused	Match user genres
S3	Genre-Contrast Diversity	Select unseen or unusual genres
S4	Quality + Novelty	Recommend lesser-known high-quality films
S5	Surprise	High deviation from history to induce surprise
S6	Motivate Reasoning	Thematically motivated recommendations
S7	Chain-of-Thought Reasoning(COT)	Step-by-step reasoning + thematic alignment

LLMs are used in S6 and S7 via prompt-based generation using Hugging Face Transformers with Phi-3.5-mini-instruct.

6. Evaluation Metrics

The models are evaluated on:

- Hit Rate: whether any ground-truth item appears in top-K
- Average Rank: average position of ground-truth items
- Recall@5, NDCG@5: Top-K relevance
- Diversity:
    - HHI (Herfindahl-Hirschman Index)
    - Entropy (user-level & system-level)
    - Gini Index (distribution inequality)

All scores are aggregated across the first 10 users.

Evaluation Metrics Table

Core Metrics (Averaged over First 10 Users)

Model	Hit Rate	Avg. Rank	HHI	Entropy	Gini
S1 (Simple)	1.0000	3.00	0.0104	6.6136	0.0285
S2 (Genre-Focused)	0.8333	4.67	0.1000	3.3219	0.2557
S3 (Diversify + xLSTM)	0.7667	4.00	0.1000	3.3219	0.4427
S4 (Diversify + Noise)	0.1000	10.20	0.1000	3.3219	0.0483
S5 (Surprise)	0.3000	9.40	0.1000	3.3219	0.0817
S6 (Motivate Reasoning)	0.1000	10.60	0.1000	3.3219	0.1924
S7 (Chain-of-Thought)	0.1000	2.10	0.1533	0.5907	0.0104

Recall and NDCG Results

Model	Recall@5	NDCG@5
S1 (Simple)	0.0208	0.2346
S2 (Genre-Focused)	0.0208	0.1461
S3 (Diversify + xLSTM)	0.0015	0.0301
S4 (Diversify + Noise)	0.0008	0.0110
S5 (Surprise)	0.0079	0.0362
S6 (Motivate Reasoning)	0.0032	0.0073
S7 (Chain-of-Thought)	0.0125	0.0098

System-Level Entropy (Across All Recommendations)

Model	System-Level Entropy
S5 (Surprise)	5.2398
S6 (Motivate Reasoning)	6.0461
S7 (Chain-of-Thought)	3.5850

7. Visualizations (Plot.ipynb)

The file Plot.ipynb produces comparative figures to support evaluation, including:

- Normalized Metric Comparison Across Models (Bar)
- Recall@5 and NDCG@5 Comparison (Bar)
- Diversity Metrics: HHI, Entropy, Gini, System-Level Entropy (Bar)
- Radar Plots (Per-model & Cross-model)
- Normalized Metric Line Plots
- Model Performance Heatmap
- Accuracy vs Diversity Scatter Plot
- Horizontal Bar: System-Level Entropy

These plots reveal trade-offs between precision and novelty across models.

8. Results Summary

Accuracy

- S1 achieves the highest hit rate and recall but lacks diversity.
- S2 and S3 maintain balance between genre relevance and minor diversity.
- S7 performs best in ranking (Avg. Rank = 2.10) but still lacks top-K coverage.

Diversity & Novelty

- S5 and S6 improve system-level entropy by promoting obscure titles.
- S3 shows over-personalization (Gini = 0.44), overfitting to user niches.
- S7 demonstrates high ranking precision with moderate diversity.

Explanation Power

- S6 and S7 generate text explanations.
- These are suitable for trust-aware and human-centered applications.

9. Conclusion

This practical work demonstrates how LLMs can enhance movie recommendation beyond traditional filtering. While S1 performs best in pure accuracy, LLM-based models (S6, S7) introduce meaningful explanations and S5–S7 promote novelty. Future work could explore fine-tuning, dialog-based interaction, and hybrid methods combining ranking and retrieval.

10. References

[1] Yasmin Delbar, et al. "Benchmarking Large Language Models as Recommender Systems." In Proceedings of the ACM Web Conference 2024 (WWW '24), ACM, 2024. https://dl.acm.org/doi/pdf/10.1145/3690655

GitHub Reference: https://github.com/yasdel/Benchmark_RecLLM_Fairness