Practical work Evaluation Summary: LLM-Based Movie Recommendation Systems

Objective

Evaluate how different LLM prompting strategies affect movie recommendation accuracy, diversity, and novelty. The system uses a decoder-based language model to infer top movie suggestions from a user's historical preferences.

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)

Dataset

- MovieLens 1M
- Interaction Data: `test_data_ml1m_fullInteraction`
- Metadata: `df_users_ml`, `df_items_ml`
- Evaluation: 10 users × 120 movie candidates (3 inputs, rest for ground truth)

Model Variants

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

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@3/5	NDCG@3/5
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

Summary of Experimental Results

Accuracy

- **S1 (Simple)** remains the most accurate, with perfect Hit Rate and leading Recall@3 and NDCG@3 — but lacks diversity.
- **S2 (Genre-Focused)** and **S3 (Diversify + xLSTM)** maintain decent hit rates but underperform in Recall and NDCG, suggesting many irrelevant top results.
- **S7 (Chain-of-Thought)** excels in **ranking relevance** (Avg. Rank = 2.10) but needs improvement in Recall to surface more relevant items in top positions.

Diversity & Novelty

- **S5 (Surprise)** and **S6 (Motivate Reasoning)** offer a good balance between novelty and explanation, reflected in higher system-level entropy and moderate Gini scores.
- **S3** suffers from **over-personalization** (Gini = 0.4427), likely reinforcing niche preferences too strongly.
- **S7 (CoT)** shows potential in both precision and diversity but needs better coverage to improve hit and recall rates.

Explanation Power

- **S6** and **S7** lead in transparency, generating **natural language rationales** that improve user trust and interpretability.
- **S5–S7** prioritize explainability, making them well-suited for **human-centered or trust-aware applications**, even at the expense of pure accuracy.