Do LLMs Memorize Recommendation Datasets?

A Preliminary Study on MovieLens-1M

Dario Di Palma*, Felice Antonio Merra*, Maurizio Sfilio, Vito Walter Anelli, Fedelucio Narducci, Tommaso Di Noia

Scan for PDF

Yes, and now we know it

Table 1: Coverage of movies.dat, users.dat, and ratings.dat. Models are grouped by version and ordered by size.

Model Name	Item Coverage (3883 items)	User Coverage (6040 users)	Interaction Coverage (1M interactions)
GPT-4o	80.76%	16.52%	9.37%
GPT-4o mini	8.47%	13.34%	7.17%
GPT-3.5 turbo	60.47%	17.38%	8.92%
Llama-3.3 70B	7.65%	5.84%	2.08%
Llama-3.2 3B	2.68%	13.26%	6.22%
Llama-3.2 1B	1.93%	10.98%	6.49%
Llama-3.1 405B	15.09%	15.30%	8.32%
Llama-3.1 70B	8.01%	15.81%	6.83%
Llama-3.1 8B	3.71%	13.59%	3.82%

How much do LLMs know about MovieLens-1M? 🍿

1. **LLMs possess extensive knowledge of the MovieLens-1M**, including items, user attributes, and interaction histories.
2. **A simple prompt** allows GPT-4o to **recover nearly 80% of Item records** (movies.dat).
3. **None of the examined models are free of this knowledge**, suggesting that MovieLens-1M data is likely included in their training sets.

Table 2: Recommendation accuracy performance on standards recommended and LLM-based recommendation. LLMs are grouped by model family and sorted by size. Best performance in each family is shown in **bold**.

Model Name	HR@1	nDCG@1	HR@5	nDCG@5	HR@10	nDCG@10
Random	0.0093	0.0093	0.0442	0.0092	0.0851	0.0094
MostPop	0.0212	0.0212	0.0775	0.0222	0.1520	0.0251
UserKNN	0.0306	0.0306	0.1209	0.0306	0.2250	0.0347
ItemKNN	0.0394	0.0394	0.1217	0.0353	0.1828	0.0337
BPRMF	**0.0406**	**0.0406**	**0.1278**	**0.0350**	**0.2149**	**0.0356**
$EASE^R$	0.0295	0.0295	0.1124	0.0278	0.1975	0.0299
LightGCN	0.0358	0.0358	0.1136	0.0306	0.1882	0.0311
GPT-4o	**0.2796**	**0.2796**	**0.5889**	**0.2276**	**0.6897**	**0.1948**
GPT-4o mini	0.0316	0.0316	0.2132	0.0451	0.3091	0.0413
GPT-3.5 turbo	0.2298	0.2298	0.4217	0.1281	0.5902	0.1229
Llama-3.3 70B	0.2293	0.2293	0.4985	0.1693	0.5922	0.1359
Llama-3.2 3B	**0.0421**	**0.0421**	**0.1886**	**0.0443**	**0.2982**	**0.0432**
Llama-3.2 1B	0.0222	0.0222	0.1018	0.0234	0.1419	0.0207
Llama-3.1 405B	**0.1975**	**0.1975**	**0.4165**	**0.1294**	**0.5119**	**0.1039**
Llama-3.1 70B	0.1302	0.1302	0.3828	0.1095	0.5148	0.0969
Llama-3.1 8B	0.0687	0.0697	0.2281	0.0609	0.3500	0.0571

Is the recommendation performance of LLM-based models influenced by memorization? 📈

While the recommendation results appear strong, a comparison between Table 1 and Table 2 reveals a notable pattern.

1. **Larger models exhibit greater memorization** of the dataset.
2. **Within each model family**, the variant with **higher memorization** also achieves **better recommendation performance**:
   - *GPT-4o > GPT-4o mini*
   - *LLaMA-3.1 405B > LLaMA-3.1 70B > LLaMA-3.1 8B*
3. This trend suggests that part of the performance may be attributed to **memorization** rather than **true generalization**.

Do LLM-based recommenders already know your test set?

How LLMs Memorize Items? 🎯

- LLMs exhibit a **strong popularity bias**: the **top 20%** of most **popular items** are significantly **easier to retrieve** than the bottom 20%.
- This bias **mirrors** the uneven **distribution of data**, where popular movies are heavily overrepresented.
- LLMs tend to **memorize and prioritize popular items**, raising **concerns** about **fairness**, **diversity**, and generalization in recommendation tasks.

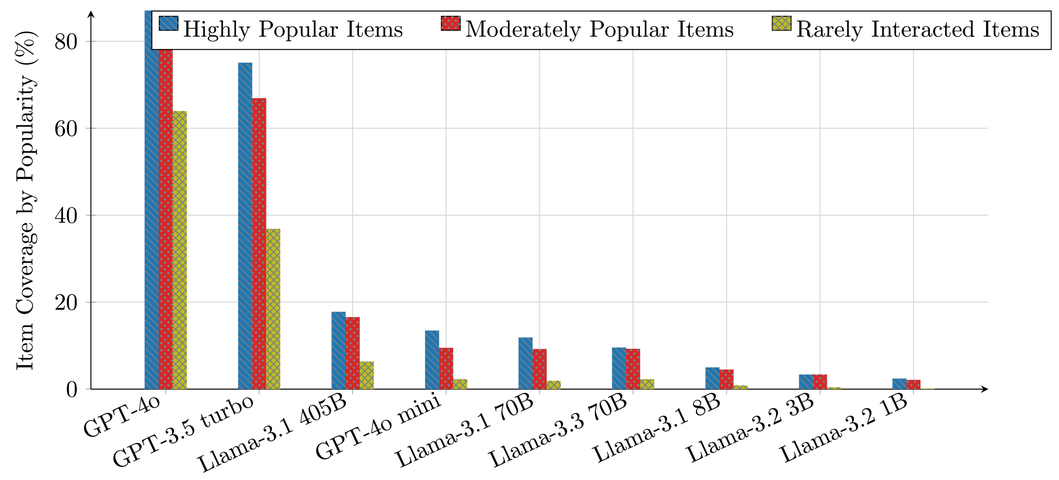

Figure 2: Comparison of item coverage across models by popularity tier. The figure shows the percentage of items covered in three categories: Highly Popular (Top 20%), Moderately Popular (Middle 20%), and Rarely Interacted (Bottom 20%).