Benchmarking LLMs in Recommendation Tasks : A Comparative Evaluation with Conventional Recommenders

https://arxiv.org/html/2503.05493v1

0. Introduction

- 추천 시스템은 개인화 정보 제공에서 핵심 역할을 수행.
- 기존 추천 시스템은 정확도와 효율성 간 균형 문제 존재.
- LLM을 추천 시스템에 활용하려는 연구가 증가했지만, 벤치마크와 평가가 제한적임.
- 본 논문 기여: **RecBench** 플랫폼 제안, 다양한 아이템 표현과 17개 LLM 평가, CTR과 SeqRec 두 가지 추천 태스크에서 성능 비교.

1. Overview

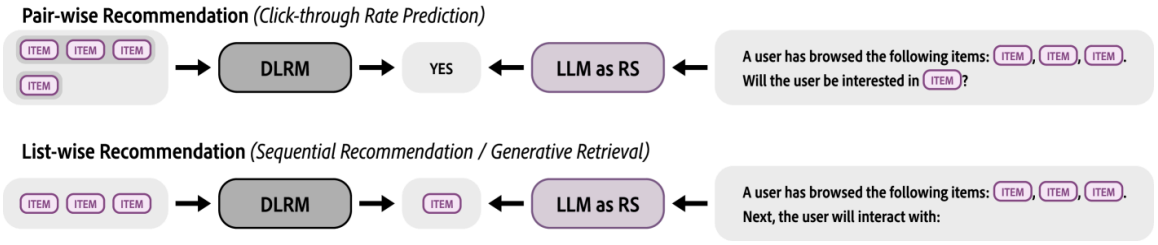
- **RecBench**: LLM 기반 추천 시스템 성능을 종합적으로 측정할 수 있는 벤치마크.
- 아이템 표현 방식 다양화: 고유 ID, 텍스트, 의미 임베딩, 의미 ID.
- 태스크: CTR 예측(클릭 여부), SeqRec(순차적 추천).
- 데이터셋 범위: 패션, 전자상거래 등 실제 산업 데이터를 기반으로 평가.
- 평가 목표: LLM 추천 성능의 정확도와 효율성을 동시에 측정.

2. Challenges

- LLM 모델 크기와 추론 비용: 대규모 LLM은 실시간 추천 구현에 부담.
- 아이템 표현 다양화로 인한 성능 편차: 단순 ID 기반 추천보다 텍스트·임베딩 활용 필요.

- 벤치마크 부족: 기존 연구는 제한된 모델과 데이터셋만 평가.
- CTR vs SeqRec 등 태스크별 특성 차이로 모델 일반화 어려움.

3. Method



- LLM을 추천 시스템에 적용하는 방식 명확히 정의.
- 아이템 표현:
  - ID: 고유 식별자 기반
  - Text: 아이템 설명 활용
  - Embedding: 의미 기반 벡터
  - Semantic ID: 의미를 반영한 ID
- 태스크별 접근:
  - CTR: LLM 입력으로 유저·아이템 정보를 넣고 클릭 확률 예측
  - SeqRec: 시퀀스 정보를 입력으로 다음 아이템 예측
- 비교 대상: 기존 추천 모델과 LLM 성능 비교
- 평가 지표: 정확도, 추천 품질, 추론 시간 등.

4. Experiments

Dataset		H&M	MIND	Micro.	Good.	CDs
Type		Fashion	News	Video	Book	Music
Text Attribute		desc	title	title	name	name
Pair-wise Test set	#Sample	20,000	20,006	20,000	20,009	20,003
	#Item	26,270	3,088	15,166	26,664	36,765
	#User	5,000	1,514	5,000	1,736	4,930
Pair-wise Finetune set	#Sample	100,000	100,000	100,000	100,005	100,003
	#Item	60,589	17,356	19,111	74,112	113,671
	#User	25,000	8,706	25,000	8,604	24,618
List-wise Test set	#Seq	5,000	5,000	5,000	5,000	5,000
	#Item	15,889	10,634	12,273	38,868	19,684
List-wise Finetune set	#Seq	40,000	40,000	40,000	40,000	40,000
	#Item	35,344	24,451	18,841	136,296	95,409

- 데이터셋:
  - 전자상거래, 패션 등 6개 산업 도메인
  - 유저-아이템 상호작용 기록 포함
- 실험 설계:
  - CTR 예측과 SeqRec 두 태스크 모두 수행
  - 아이템 표현 방식별 성능 비교
- 평가 지표:
  - Accuracy, Hit Rate, NDCG 등
  - 추론 속도 및 비용도 측정

5. Results

Recommender	MIND	Micro.	Good.	CDs	H&M	Overall	Latency
$SASRec_{3L}$	0.0090	0.0000	0.0165	0.0016	0.0209	0.0096	23.30ms
$SASRec_{6L}$	0.0097	0.0006	0.0224	0.0012	0.0297	0.0127	38.43ms
$SASRec_{12L}$	0.0241	0.0297	0.0548	0.1041	0.1235	0.0672	51.77ms
$SASRec_{24L}$	0.0119	0.0312	0.0601	0.1267	0.1191	0.0698	103.41ms
$BERT_{base}$	0.0430	**0.1867**	**0.0557**	0.1198	0.1075	0.1025	41.54ms
$QWen\text{-}2_{0.5B}$	0.0549	0.0201	0.0322	0.0128	0.0234	0.0287	556.95ms
$QWen\text{-}2_{1.5B}$	0.0506	0.0254	0.0316	0.0015	0.0217	0.0262	1.12s
$Llama\text{-}3_{7B}$	0.0550	0.0178	0.0134	0.0072	0.0353	0.0257	28.06s
$SID\text{-}SASRec_{3L}$	0.0266	0.0028	0.0029	0.0000	0.0084	0.0081	36.12ms
$SID\text{-}SASRec_{3L\text{-}CBS}$	0.0849	0.0123	0.0127	0.0007	0.0422	0.0306	66.67ms
$SID\text{-}SASRec_{6L}$	0.0225	0.0047	0.0038	0.0140	0.0097	0.0109	59.08ms
$SID\text{-}SASRec_{6L\text{-}CBS}$	0.0647	0.0179	0.0141	0.0331	0.0406	0.0341	90.41ms
$SID\text{-}SASRec_{12L}$	0.0201	0.0044	0.0039	0.0136	0.0165	0.0117	1.31s
$SID\text{-}SASRec_{12L\text{-}CBS}$	0.0695	0.0234	0.0140	0.0324	0.0598	0.0398	1.34s
$SID\text{-}BERT_{base}$	0.0654	0.0022	0.0025	0.3539	0.0467	0.0941	1.83s
$SID\text{-}BERT_{base\text{-}CBS}$	**0.1682**	0.1195	0.0059	0.4616	**0.1834**	**0.1877**	1.90s
$SID\text{-}Llama\text{-}3_{7B}$	0.0456	0.0255	0.0221	0.2443	0.0337	0.0742	167.25s
$SID\text{-}Llama\text{-}3_{7B\text{-}CBS}$	0.1677	0.0827	0.0508	0.3898	0.1125	0.1607	177.54s

- **CTR 태스크**:
  - LLM은 텍스트/임베딩 기반 표현에서 기존 모델 대비 성능 향상
  - 단순 ID 기반은 기존 모델과 유사하거나 낮음
- **SeqRec 태스크**:
  - LLM은 시퀀스 패턴 이해에서 강점
  - 모델 크기에 따라 성능-속도 트레이드오프 존재
- Ablation Study:
  - 표현 방식 변화, 모델 크기, 입력 길이 조절 시 성능 차이 관찰
- 실무 적용 가능성:
  - LLM은 정확도 향상 가능하지만 실시간 추천 비용 부담 존재

6. Insight

- LLM 기반 추천 시스템은 아이템 표현과 태스크 유형에 따라 성능 차이가 큼.
- 텍스트 및 임베딩 활용 시 기존 모델보다 CTR/SeqRec에서 유리함.
- 대규모 모델은 정확도는 높지만 추론 비용이 높아 산업 적용 시 최적화 필요.
- RecBench는 LLM 추천 성능 비교를 위한 표준화된 평가 플랫폼으로 활용 가능.
- 향후 연구 방향: 경량화 LLM, 효율적 추론 전략, 다양한 도메인 확장.