



KDD2024
BARCELONA, SPAIN



Benchmarking Recommendation Ability of Foundation Models: **Legomenders** and **RecBench**

Qijiong LIU

The Hong Kong Polytechnic University

liu@qijiong.work

Existing Recommendation Benchmarks

- **DeepCTR (2017)**
 - 29 CTR Models
- **BARS (2022)**
 - 19 Matching Models
 - 44 Ranking (CTR) Models
- **RecBole (2021)**
 - 18 Ranking (CTR) Models
 - 32 Matching Models
 - 10 KG-based Models
 - 31 Sequential Models

No Benchmark for Content-based Recommendation?

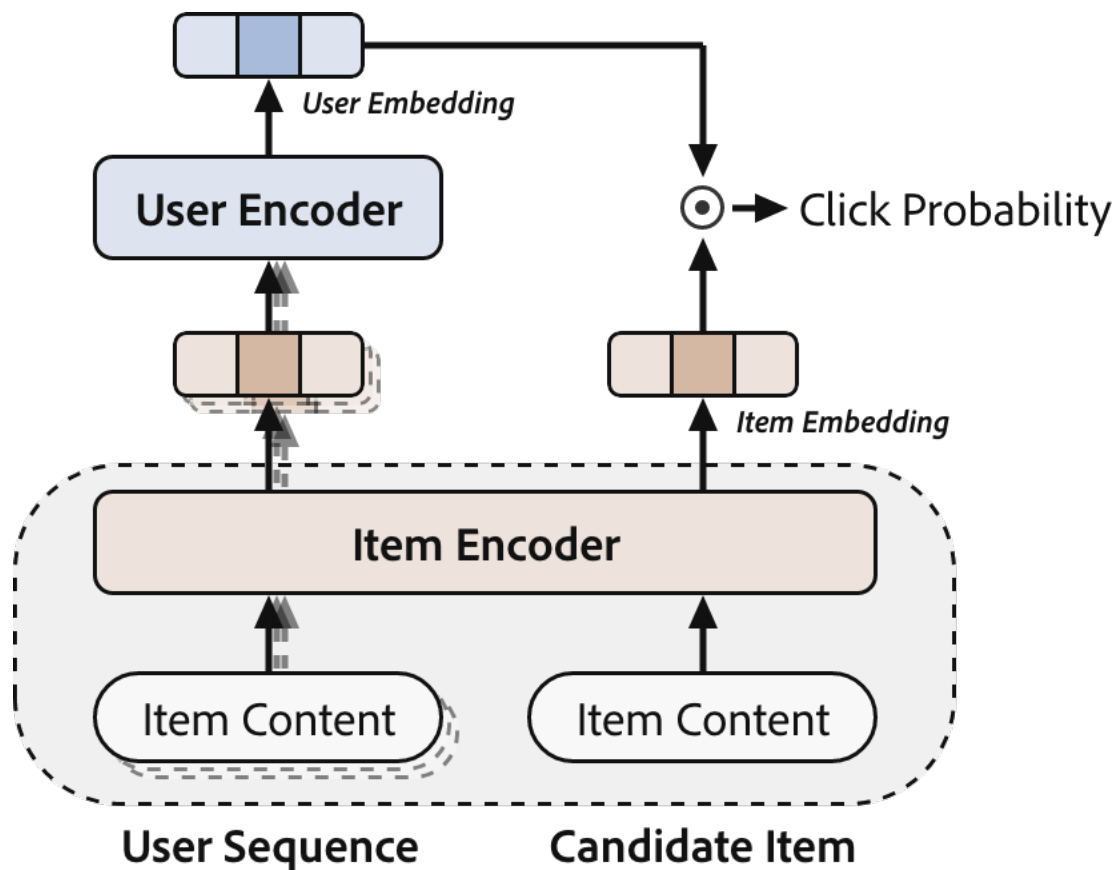
Content-based Recommendation Scenario

- Providing Informative Knowledge
- Mitigating Cold-start Problem

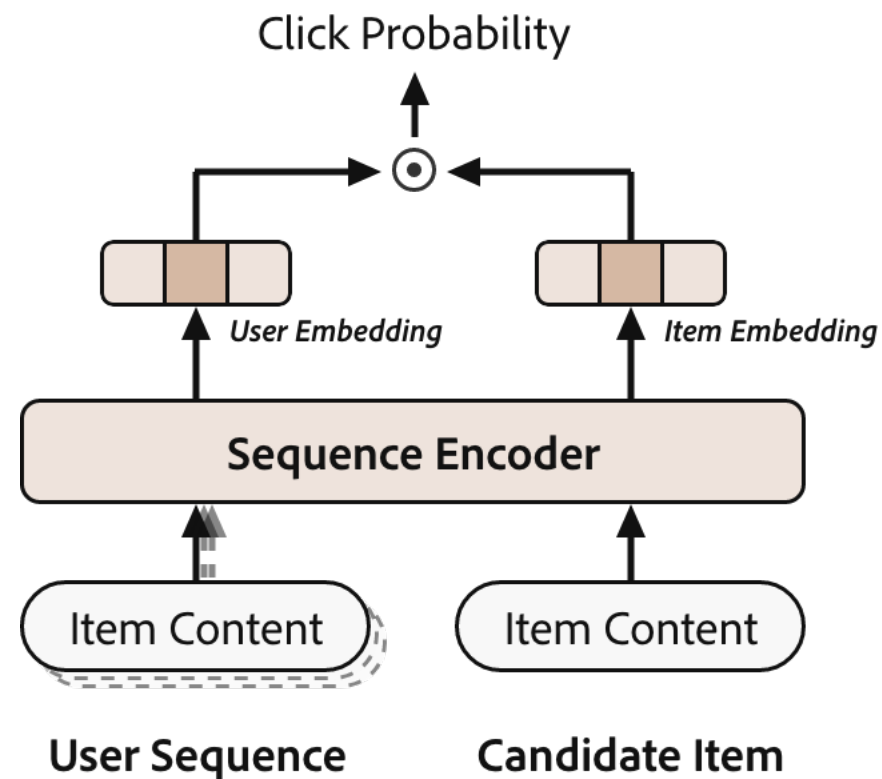
- News Recommendation
- Music Recommendation
- Book Recommendation
- Advertisement Recommendation

How to Integrate Content Features into Recommender Systems?

Paradigms of Content-based Recommenders



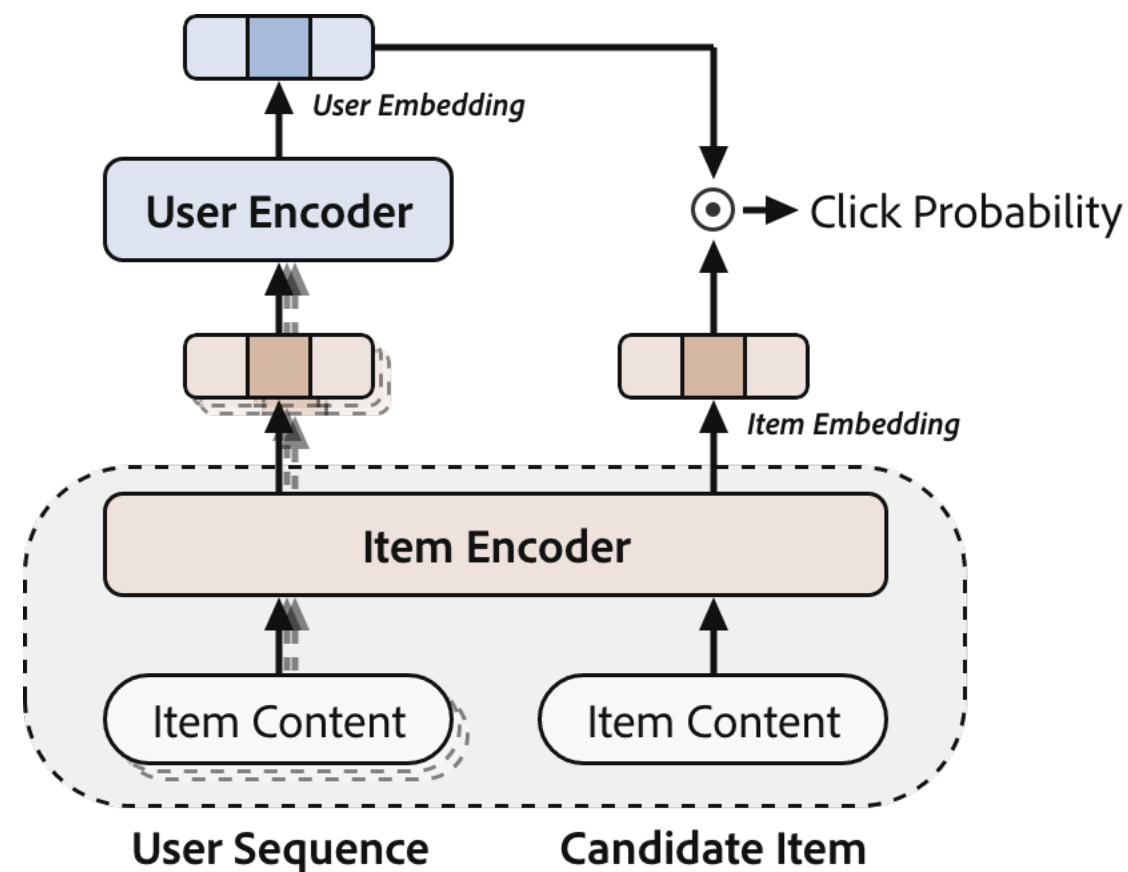
Cascaded Paradigm



Flatten Paradigm

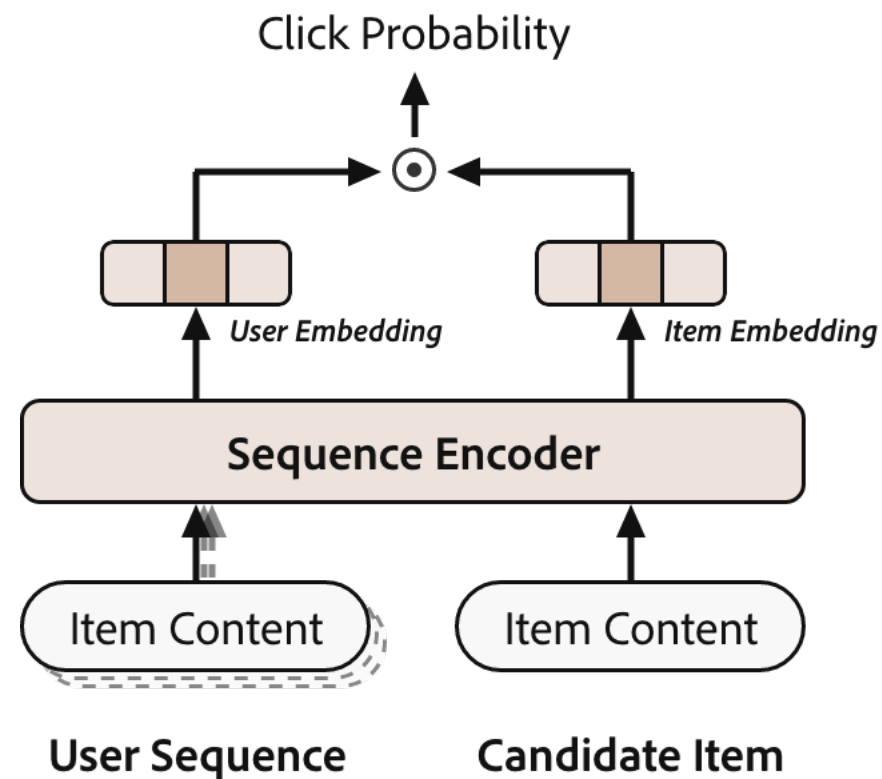
Cascaded Paradigm

- Components
 - Item Encoder
 - User Encoder
 - Interaction Module
- Item Encoder
 - Decoupled
 - End-to-end




Flatten Paradigm

- Components
 - Sequence Encoder
 - Interaction Module
- End-to-end training



Paradigm Comparison

Paradigm	Cascaded Paradigm		Flatten Paradigm
	Decoupled	End-to-end	End-to-end
User Sequence Length	Short	Short	Long
Training Efficiency	Fast	Slow	Extremely Slow
Inference Acceleration	User Caching	User and Item Caching	User and Item Caching
Function of LLMs	Feature Extractor	Item Encoder	Sequence Encoder
Representative Models	RecoBERT	PLM-NR	Recformer
DeepCTR / BARS / RecBole	✓	✗	✗
 Legommenders	✓	✓	- (Partially)



Legommenders

Legommenders

A modular framework for recommender systems



Updates

Jan. 23, 2024

- Legommenders partially supports the flatten sequential recommendation model.

Predefined Item Encoders

Network	Representative Models
Convolutional Neural Network	NAML
Attention Module	NRMS
Average Pooling	CTR Models (<i>e.g.</i> , DCN, DeepFM)
Pretrained Language Model	ONCE, RecoBERT
Pretrained Vision Model	Fusion-Q-Former, KDSR
Fastformer Module	Fastformer

Predefined User/Sequence Encoders

Network	Representative Models
Attention Module	NRMS
Transformer Module	BST, UIST
Average Pooling	CTR Models (<i>e.g.</i> , DCN, DeepFM)
Pretrained Language Model	Recformer
Fastformer Module	Fastformer
GRU Module	LSTUR

Predefined Interaction Modules

Network	Representative Models
Dot Product	NAML, NRMS, ONCE
Deep & Cross Network	DCN, GDCN, DCNv2
Deep Interest Network	DIN
Product Neural Network	PNN

Predefined Training Schemes

Schemes	Matching	Ranking
Training Objective	Cross Entropy Loss	Binary Cross Entropy Loss
Training Sample	<User, Pos Item, Neg Items>	<User, Item>
Number of Training Samples	Number of Positive Pairs	Number of Pairs
Negative Sampling	Yes	N/A
Representative Models	NAML, RecoBERT, MINER	DCN, DeepFM, PNN



Legomenders

- A Modular Framework for Recommender Systems
- Construct customized models by code-free configuration

Item Encoder	User Encoder	Interaction Module	Negative Sampling	Model Name
Average Pooling	Average Pooling	Deep & Cross	No	DCN
Attention	Attention	Dot Product	Yes	NRMS
Fastformer	Fastformer	Dot Product	Yes	Fastformer
Transformer	Transformer	Poly-DCN	No	UIST-DCN
N/A	LongFormer	Dot Product	Yes	Recformer

Other Features

- Pre-defined metrics
 - AUC, MRR, NDCG, F1, Recall, HitRatio, ...
- Pre-defined datasets
 - MIND, Goodreads, MovieLens, EB-Nerd...
- **Inference Acceleration**
 - Item Caching and User Caching (2 hours to 10 minutes)
- **Green AI Metric**
 - ApC: AUC per Carbon Emission



Used By Leading Institutions



Researches Using Legommenders

- ONCE: Boosting Content-based Recommendation with Both Open- and Closed-source Large Language Models (WSDM 2024)
- Benchmarking News Recommendation in the Era of Green AI (TheWebConf 2024)
- Discrete Semantic Tokenization for Deep CTR Prediction (TheWebConf 2024)
- SPAR: Personalized Content-Based Recommendation via Long Engagement Attention (arXiv 2024)
- EmbSum: Leveraging the Summarization Capabilities of Large Language Models for Content-Based Recommendations (RecSys 2024)

Experimental Results

$$ApC = \frac{AUC - 50}{Carbon\ Emission} \times 100$$

Backbone	NAML			DCN		
	Metric	AUC	NDCG@5	ApC	AUC	NDCG@5
ID-based	50.13	22.35	0.68	53.92	24.43	6.53
Text-based	60.14	29.33	24.14	62.63	30.52	20.05
PLM-NR	62.06	32.25	6.78	63.32	32.58	0.75
PREC	62.95	32.01	58.86	64.57	33.48	23.50

Selected from Benchmarking News Recommendation in the Era of Green AI



Legommenders

Modular Framework for Recommender Systems,
Following MIT License and supporting commercial use.

<https://github.com/Jyonn/Legommenders>

Paradigm Shift from “*LLM for RS*” to “*LLM as RS*”

RecBench

Benchmarking Recabilities for Large Language Models

Within 10 diverse recommendation scenarios

Over more than 10 large language models

Across 2 benchmarking schemes

By zero-shot as well as supervised finetuning

Benchmarking Scenarios

Dataset	Scenario	Provider
H&M	Fashion	H&M
MIND	News	Microsoft
MicroLens	Video	Westlake Uni.
Goodreads	Book	UCSD
CDs	Music	Amazon
MovieLens	Movie	UMN
Yelp	Restaurant	Yelp
Steam	Game	OPI PIB
Electronics	E-commerce	Amazon
HotelRec	Hotel	EPFL

Benchmarking Baselines

Type	Date	Model	Size	Institution
General Large Language Models	2018.10	BERT-base	110M	Google
	2022.05	OPT-base	331M	Meta
	2022.05	OPT-large	1,316M	Meta
	2022.11	GPT-3.5	-	OpenAI
	2023.02	Llama-1	6,738M	Meta
	2023.05	E5-2	109M	Microsoft
	2023.06	Llama-2	6,738M	Meta
	2023.10	Phi-2	2,780M	Microsoft
	2024.01	GLM-4	9,400M	THU
	2024.02	Mistral-2	7,248M	MistralAI
	2024.06	Llama-3	8,030M	Meta
	2024.10	Qwen-2	7,616M	Alibaba
Large Recommendation Model	2022.03	P5	223M	Rutgers Uni.
	2023.05	Recformer	148M	UCSD
	2024.05	RecGPT	6,738M	VinAI

Benchmarking Schemes

- Generative Scoring
- Discriminative Matching

RecBench: Generative Recability by Scoring

You are a recommender. Please response “YES” or “NO” to represent whether this user is interested in this item. User behavior sequence:

- (1) 20 Words and Phrases You Had No Idea Were Coined in New York City
- (2) The 50 Worst Netflix Original Movies, Ranked According to Critics
- (3) 12 Shed Storage Ideas to Organize Your Space at Last ...

Candidate Item:

Viral College Football Fan Under Fire for Past Offensive Tweets.

Answer (Yes/No): No

RecBench: Discriminative Recability by Matching

User behavior sequence:

- (1) 20 Words and Phrases You Had No Idea Were Coined in New York City
- (2) The 50 Worst Netflix Original Movies, Ranked According to Critics
- (3) 12 Shed Storage Ideas to Organize Your Space at Last ...

Candidate Item:

Viral College Football Fan Under Fire for Past Offensive Tweets.

Benchmarking Pipelines

- Zero-shot Recommendation
- Recommendation with Supervised Finetuning

Benchmarking Results (Zero-shot, Generative)

	MIND	MovieLens	MicroLens	Goodreads	Yelp	Steam	CDs	H&M	Electronics	HotelRec	Overall
BERT _{base}	0.4963	0.4934	0.4992	0.4958	0.4914	0.5002	0.5059	0.5204	0.5037	0.4955	0.5002
OPT _{base}	0.5490	0.5104	0.4773	0.5015	0.5158	0.4257	0.5093	0.4555	0.5054	0.5028	0.4953
OPT _{large}	0.5338	0.5174	0.5236	0.5042	0.5026	0.3825	0.4994	0.5650	0.5205	0.5026	0.5052
Llama-1	0.4583	0.6216	0.4572	0.4994	0.5206	0.5817	0.4995	0.4035	0.4892	0.5457	0.5077
Llama-2	0.4945	0.6030	0.4877	0.5273	0.5378	0.5622	0.5191	0.4519	0.4962	0.5305	0.5210
Llama-3	0.4904	0.6412	0.5577	0.5191	0.5267	0.7690	0.5136	0.5454	0.5223	0.5342	0.5620
Llama-3.1	0.5002	0.6444	0.5403	0.5271	0.5317	0.7650	0.5088	0.5462	0.5284	0.5268	0.5619
Mistral	0.6300	0.6705	0.6579	0.5718	0.5187	0.8619	0.5230	0.7166	0.5205	0.4851	0.6156
GLM-4	0.6304	0.6568	0.6647	0.5671	0.5292	0.8556	0.5213	0.7319	0.5174	0.4961	0.6171
Qwen-2	0.5862	0.6321	0.6640	0.5494	0.5317	0.8327	0.5256	0.7124	0.5202	0.5317	0.6086
Phi-2	0.4851	0.5296	0.5078	0.5049	0.5186	0.6061	0.4991	0.5447	0.5138	0.4986	0.5208
GPT-3.5	0.5057	0.5170	0.5110	0.5122	0.5039	0.6184	0.5046	0.5801	0.5134	0.5076	0.5274
RecGPT	0.5078	0.5069	0.4703	0.5083	0.5140	0.4924	0.5019	0.4875	0.5107	0.4937	0.4993
P5	0.4911	0.5138	0.5017	0.5027	0.5080	0.5296	0.5447	0.4845	0.5263	0.4905	0.5093
Llama-3	0.6732	0.7203	0.6223	0.5864	0.5764	0.7828	0.5626	0.7295	0.6296	0.5806	0.6464

Benchmarking Results (Zero-shot, Discriminative)

	MIND	MovieLens	MicroLens	Goodreads	Yelp	Steam	CDs	H&M	Electronics	HotelRec	Overall
BERT _{base}	0.5263	0.4978	0.5305	0.5160	0.5123	0.5609	0.5139	0.5167	0.5257	0.5137	0.5214
OPT _{base}	0.5416	0.4914	0.5567	0.5290	0.5288	0.7462	0.5197	0.5334	0.5125	0.5078	0.5467
OPT _{large}	0.5510	0.5303	0.5447	0.5258	0.5289	0.7989	0.5137	0.6370	0.5130	0.5127	0.5656
Llama-1	0.5267	0.5352	0.5234	0.5063	0.4946	0.6298	0.5072	0.5832	0.5192	0.5044	0.5330
Llama-2	0.5291	0.4876	0.5251	0.5117	0.4997	0.5369	0.5165	0.4970	0.5228	0.5064	0.5133
Llama-3	0.5666	0.4985	0.5218	0.5150	0.5212	0.6900	0.5162	0.6487	0.5189	0.5061	0.5503
Llama-3.1	0.5682	0.4622	0.5247	0.5177	0.5182	0.6851	0.5109	0.6437	0.5274	0.5058	0.5464
Mistral	0.5607	0.4792	0.5329	0.5240	0.5142	0.7685	0.5198	0.6051	0.5177	0.5123	0.5534
GLM-4	0.5179	0.5030	0.5362	0.5173	0.5211	0.7049	0.5173	0.6442	0.5180	0.5071	0.5487
Qwen-2	0.5347	0.4754	0.5391	0.5190	0.5106	0.6790	0.5212	0.6201	0.5213	0.5164	0.5437
Phi-2	0.5381	0.5333	0.5364	0.5224	0.5093	0.7294	0.5167	0.5928	0.5208	0.5101	0.5509
E5-2	0.6010	0.5062	0.6315	0.5217	0.4806	0.8221	0.5030	0.6569	0.5237	0.4788	0.5725
P5	0.5328	0.4791	0.5383	0.5046	0.4876	0.5802	0.5010	0.6049	0.5078	0.4931	0.5229
Recformer	0.5948	0.5421	0.6423	0.5186	0.5088	0.7329	0.5218	0.6781	0.5147	0.5262	0.5780
RecGPT	0.5129	0.5700	0.5160	0.5106	0.4949	0.6416	0.5209	0.5656	0.5066	0.5198	0.5359

Benchmarking Results (SFT, Generative)

	MIND	MovieLens	MicroLens	Goodreads	Yelp	Steam	CDs	H&M	Electronics	HotelRec	Overall
BERT _{base-1}	0.5393	0.5052	0.5052	0.5218	0.5100	0.6883	0.5053	0.6758	0.5145	0.4854	0.5451
BERT _{base-2}	0.6014	0.5073	0.5096	0.5165	0.5059	0.7125	0.5055	0.6995	0.5162	0.4815	0.5556
BERT _{base-3}	0.5524	0.5827	0.5817	0.5019	0.5049	0.6769	0.5163	0.6979	0.5261	0.4934	0.5634
BERT _{base-4}	0.5221	0.5437	0.5475	0.5222	0.5169	0.5553	0.5050	0.6525	0.5306	0.4978	0.5394
BERT _{base-5}	0.5438	0.5421	0.5475	0.5203	0.4822	0.6511	0.5096	0.6641	0.5130	0.4876	0.5461
OPT _{large-1}	0.5326	0.5142	0.5142	0.5027	0.5054	0.4195	0.5023	0.5530	0.5302	0.4988	0.5073
OPT _{large-2}	0.6311	0.5457	0.5472	0.5330	0.4703	0.8305	0.5072	0.6894	0.5280	0.4841	0.5767
OPT _{large-3}	0.5996	0.6156	0.6165	0.5189	0.4853	0.7665	0.5181	0.7002	0.5446	0.5037	0.5869
OPT _{large-4}	0.5492	0.6054	0.6064	0.5141	0.5167	0.6299	0.5211	0.7029	0.5286	0.5043	0.5679
OPT _{large-5}	0.5612	0.5982	0.5979	0.5394	0.4676	0.8405	0.5016	0.7380	0.5317	0.4772	0.5853
Llama-3	0.6732	0.7203	0.6223	0.5864	0.5764	0.7828	0.5626	0.7295	0.6296	0.5806	0.6464

Features of RecBench

- First large-scale recability benchmarking platform
- Easy to use
 - Python toolkit: `pip install recbench`
 - Online leaderboard: <https://github.com/RecBench>
 - Providing SFT codes and data for reproducibility

Future of RecBench

- Benchmarking on private recommendation data
- More benchmarking schemes
 - Explainable recommendation
 - Conversational recommendation
- Supporting multimodal datasets and LLMs

<https://github.com/RecBench>



KDD2024
BARCELONA, SPAIN



THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學

THANKS FOR YOUR TIME!