LLM4UM

大模型如何为用户建模赋能？

2024-8-30 董彦

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

目录

1 概况

2 实现方法

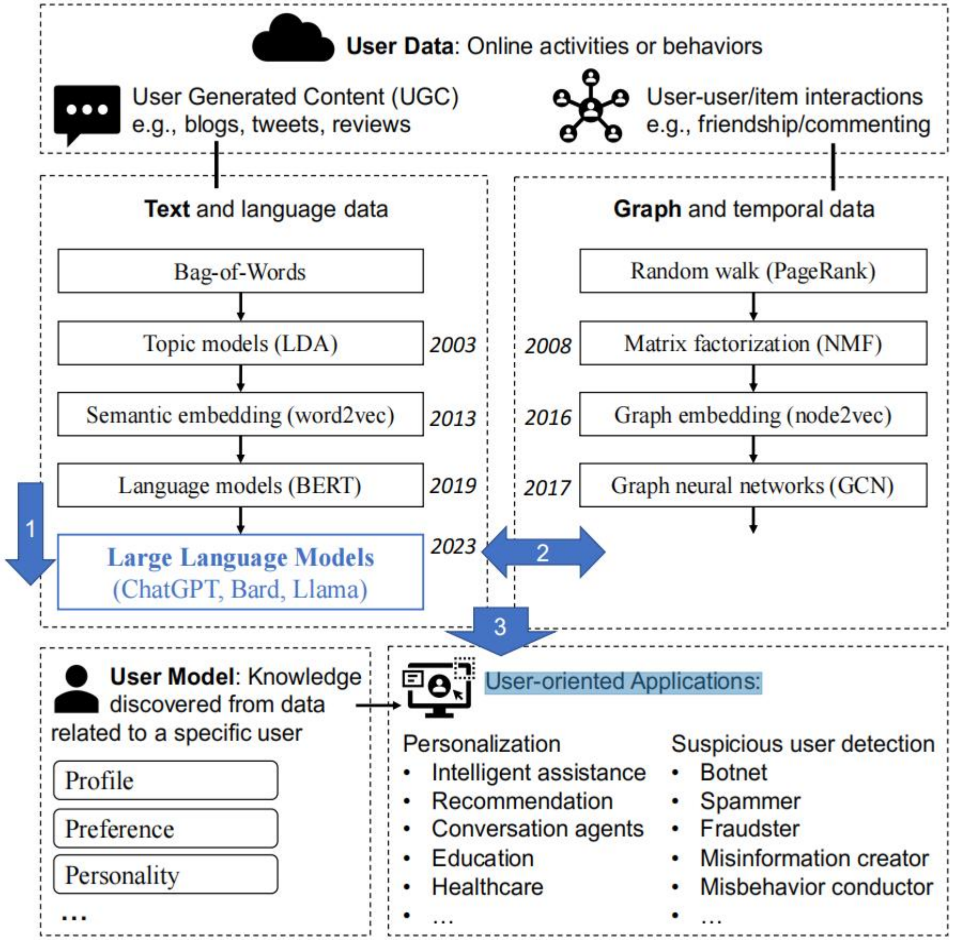
3 总结展望

概况

用户建模(UM)旨在从**用户数据**中**提取有价值的见解和模式**，使系统能够定制和适应特定用户的需求。

用户数据：

- 内容数据：推文、评论、博客等
- 行为数据：用户-用户，用户-物品关系等



User Modeling in the Era of Large Language Models: Current Research and Future Direction（arxiv2312）

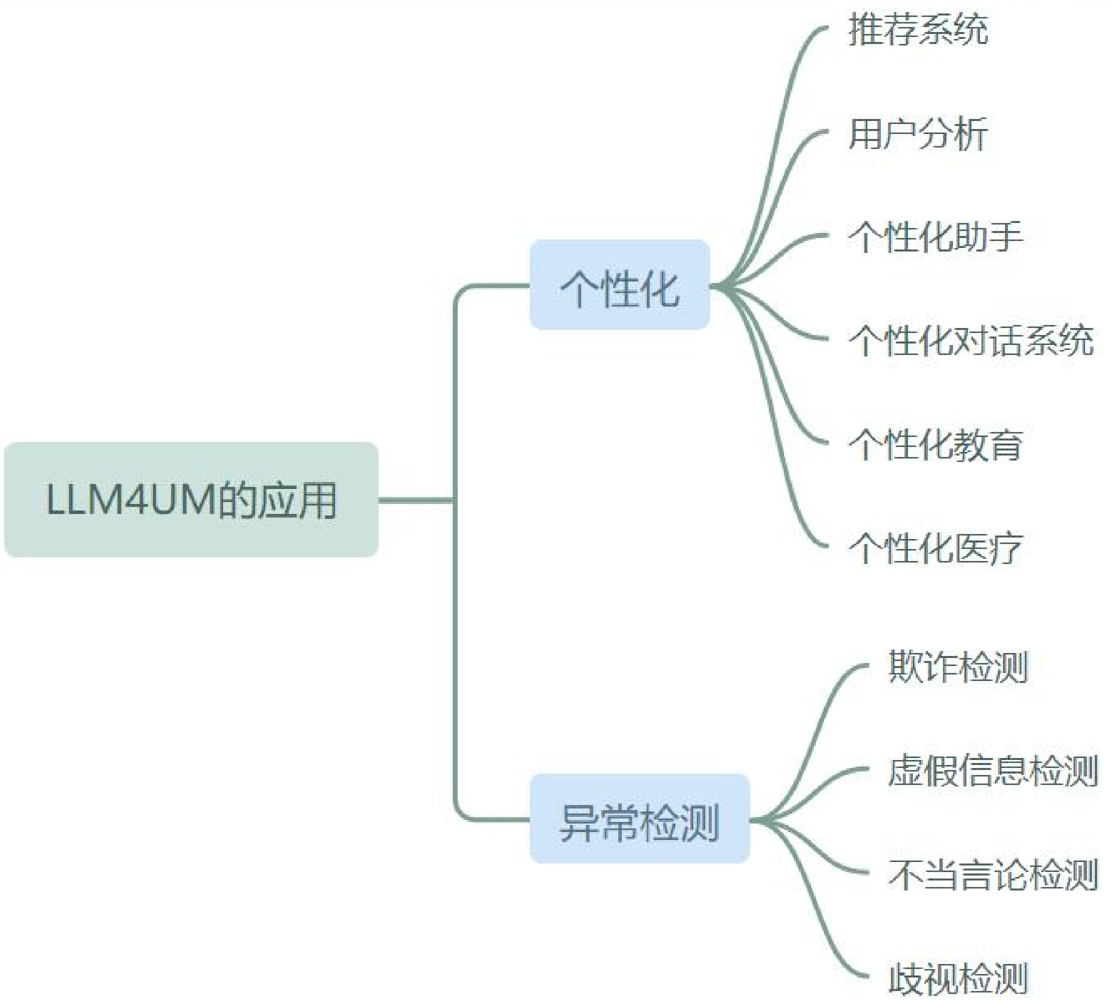
概况

LLM4UM的优势

- 泛化能力
- 生成能力

LLM4UM的分类

- LLM 作为增强器
- LLM 作为预测器
- LLM 作为控制器

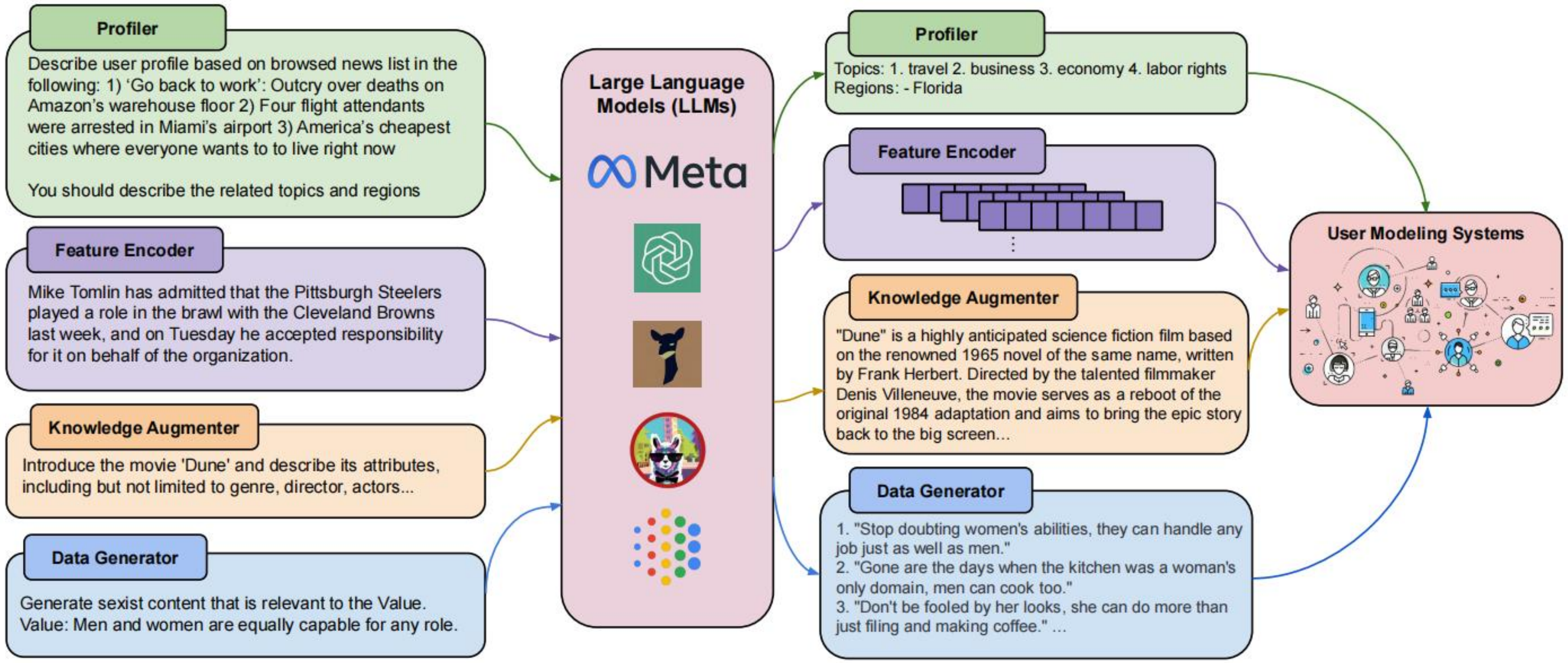


When large language models meet personalization: perspectives of challenges and opportunities（WWW 2024）

概况

LLM 作为增强器

- 画像工具
- 特征编码器
- 知识增强器
- 数据生成器

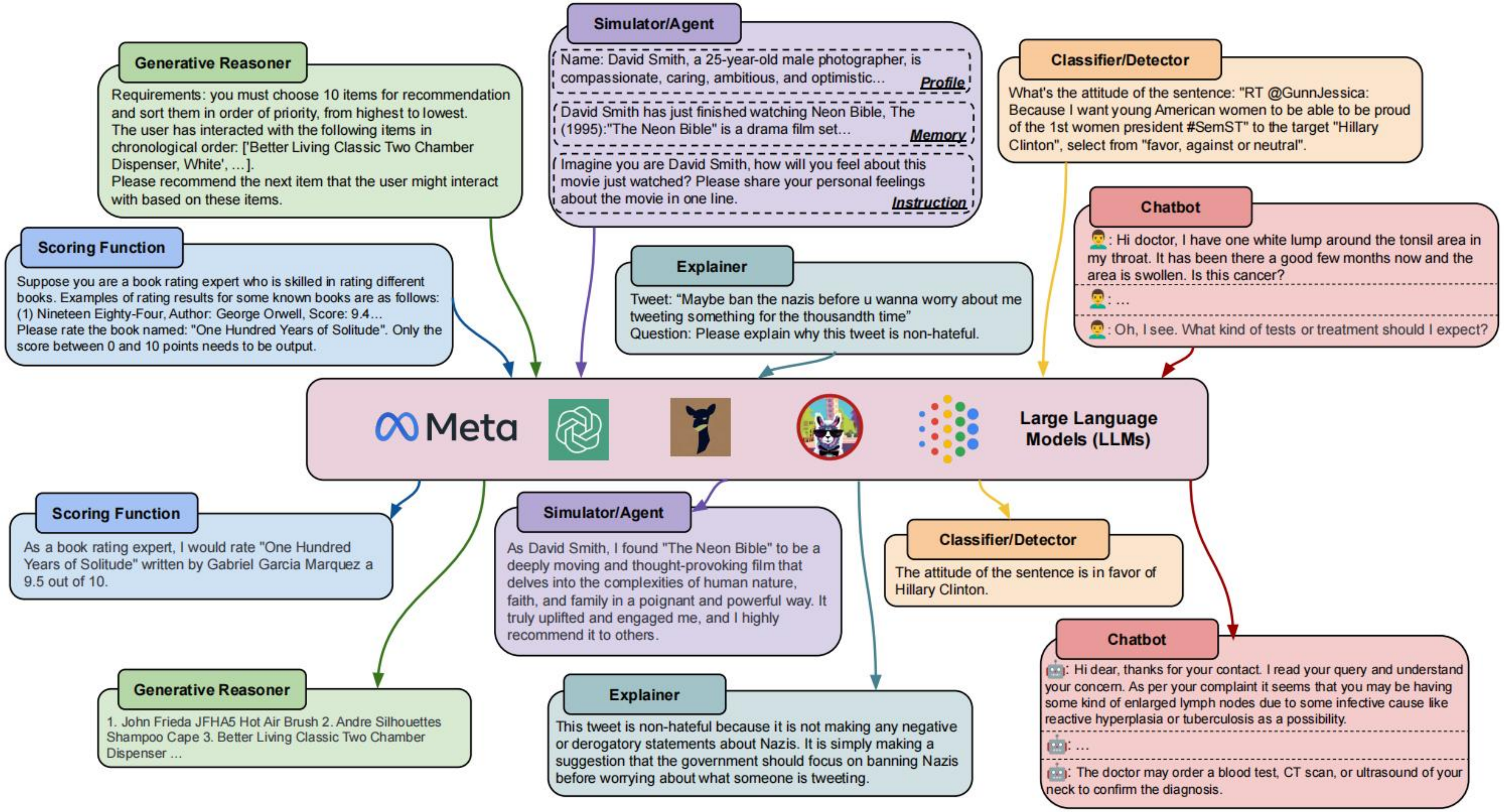


User Modeling in the Era of Large Language Models: Current Research and Future Direction（arxiv2312）

概况

LLM 作为预测器

- 生成推理器
- 模拟器/代理
- 分类器/检测器
- 评分函数
- 解释器
- 聊天机器人

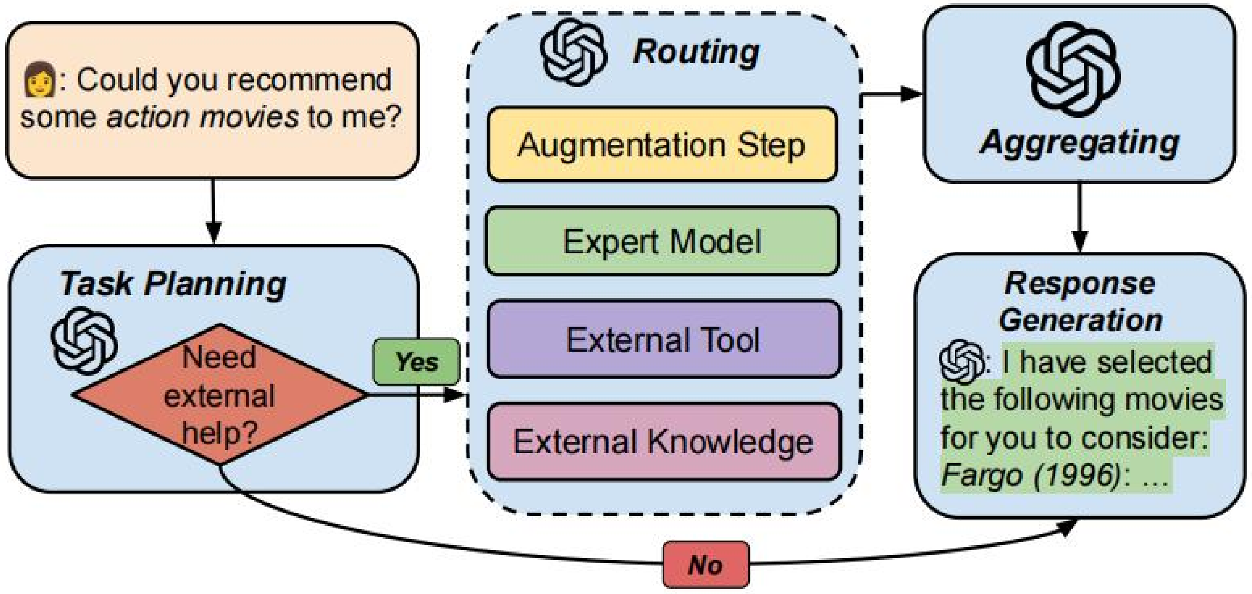


User Modeling in the Era of Large Language Models: Current Research and Future Direction（arxiv2312）

概况

LLM 作为控制器

- 管理和组织专家模型
- 调用外部工具
- 引入外部知识



LLMs-as-Controllers

User Modeling in the Era of Large Language Models: Current Research and Future Direction（arxiv2312）

实现方法

- **LLM 作为增强器**
  - Sequential Recommendation with Latent Relations based on Large Language Model（SIGIR 2024）
  - Representation Learning with Large Language Models for Recommendation（WWW 2024）
- **LLM 作为预测器**
  - CoRAL: Collaborative Retrieval-Augmented Large Language Models Improve Long-tail Recommendation（KDD2024）
  - A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems（arxiv2312）
- **LLM 作为控制器**
  - On Generative Agents in Recommendation（arxiv2405）

实现方法

- **LLM 作为增强器（Encoder）**

Sequential Recommendation with Latent Relations based on Large Language Model

Shenghao Yang
DCST, Tsinghua University
Beijing, China
ysh21@mails.tsinghua.edu.cn

Weizhi Ma[*]
AIR, Tsinghua University
Beijing, China
mawz@tsinghua.edu.cn

Peijie Sun
DCST, Tsinghua University
Beijing, China
sun.hfut@gmail.com

Qingyao Ai
DCST, Tsinghua University
Beijing, China
aiqy@tsinghua.edu.cn

Yiqun Liu
DCST, Tsinghua University
Beijing, China
yiqunliu@tsinghua.edu.cn

Mingchen Cai
Meituan
Beijing, China
caimingchen@meituan.com

Min Zhang[*]
DCST, Tsinghua University
Beijing, China
z-m@tsinghua.edu.cn

Sequential Recommendation with Latent Relations based on Large Language Model **（SIGIR 2024）**

LLMs-as-Enhancers（Encoder)

- **Motivation**

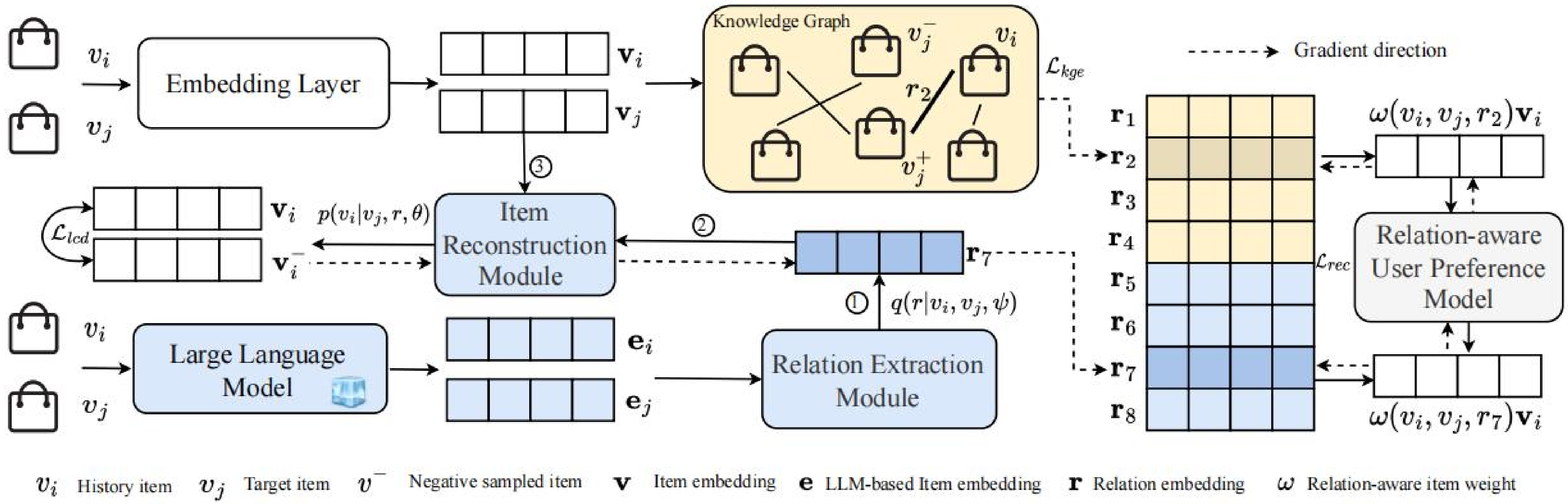
现有的基于关系感知的用户建模方法**通常依赖于手动预定义的物品关系，**存在稀疏性问题

- 在现实世界中，项目之间的关系是不同的，手工定义的关系比所有潜在的关系是稀疏的
- 依赖于一组有限的预定义关系，限制了模型在不同的推荐场景中有效地泛化的能力

Sequential Recommendation with Latent Relations based on Large Language Model（**SIGIR 2024**）

LLMs-as-Enhancers (Encoder)

- **Method**



Sequential Recommendation with Latent Relations based on Large Language Model （**SIGIR 2024**）

LLMs-as-Enhancers（Encoder)

- **关系提取模块**

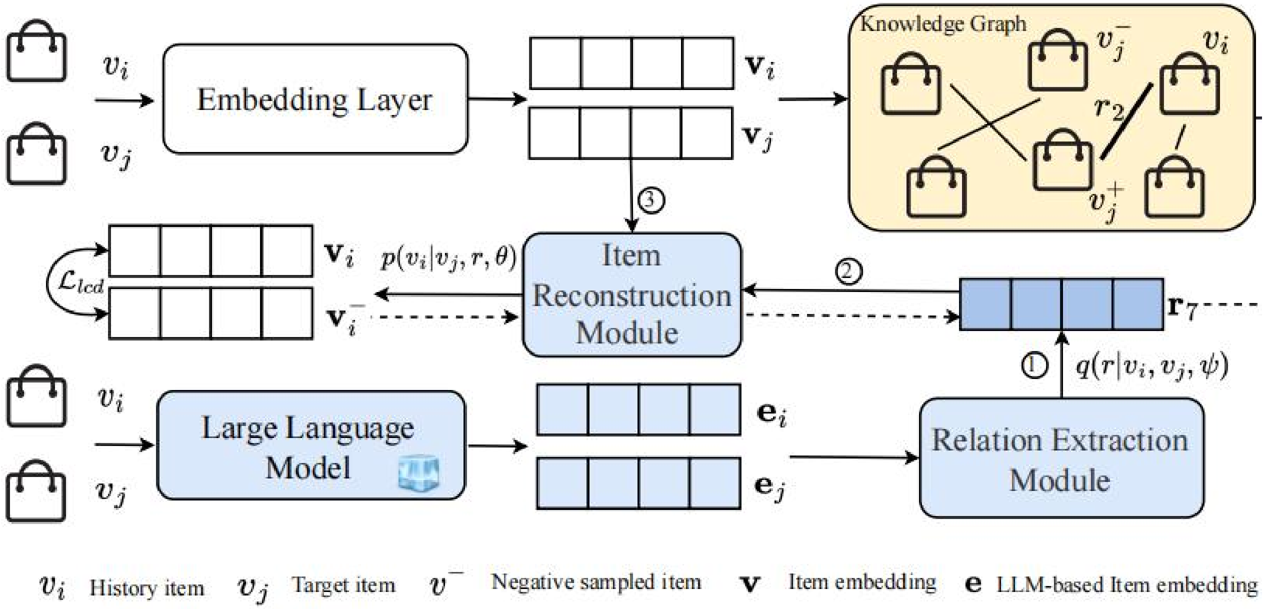
$$\mathbf{e} = W_1(LLM([w_1, w_2, w_3, ..., w_{N_i}])) + b_1,$$

$$q(r|v_i, v_{-i}, \psi) = SoftMax(W_2[\mathbf{e}_i; \mathbf{e}_{-i}] + b_2),$$

- **物品重构模块**

$$p(v_i|v_{-i}, r, \theta) = \frac{\exp(\phi(v_i, v_{-i}, r))}{\sum_{v_i' \in \mathcal{V}} \exp(\phi(v_i', v_{-i}, r))},$$

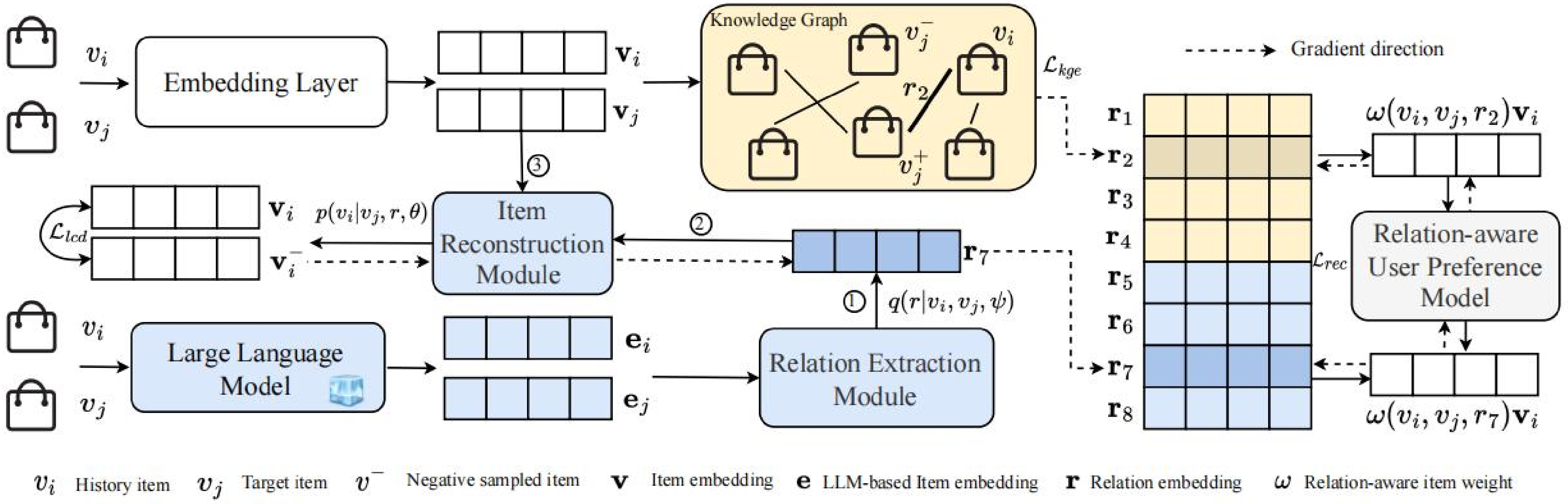
$$\mathcal{L}(\theta, \psi) = \sum_{i=1}^{2} \sum_{r \in \mathcal{R}} q(r|v_i, v_{-i}, \psi)[\log \sigma(\phi(v_i, v_{-i}, r, \theta)) + \log \sigma(-\phi(v_i^-, v_{-i}, r, \theta)))] + \alpha H[q(r|v_i, v_{-i}, \psi)].$$



Sequential Recommendation with Latent Relations based on Large Language Model（**SIGIR 2024**）

LLMs-as-Enhancers（Encoder)

利用提取关系进行推荐



$$y_{u,j} = (\mathbf{u} + \mathbf{m}_{u,j})\mathbf{v}_j^T + b_j, \Leftarrow \mathbf{m}_{u,j} = \mathrm{AGG}([\mathbf{s}_{u_j,r_1}; \mathbf{s}_{u_j,r_2}; \ldots; \mathbf{s}_{u_j,r_{|\mathcal{R}|}}]), \Leftarrow \mathbf{s}_{u_j,r} = \sum_{v_i \in S_u} \omega(v_i, v_j, r)\mathbf{v}_i,$$

Sequential Recommendation with Latent Relations based on Large Language Model（**SIGIR 2024**）

LLMs-as-Enhancers (Encoder)

- **Experiments**

Datasets	MovieLens				Office				Electronics			
Metrics	H@5	H@10	N@5	N@10	H@5	H@10	N@5	N@10	H@5	H@10	N@5	N@10
Caser	0.5217	0.6872	0.3571	0.4107	0.3095	0.4762	0.1993	0.2530	0.4620	0.5865	0.3435	0.3838
GRU4Rec	0.5101	0.6723	0.3451	0.3976	0.3295	0.4856	0.2164	0.2670	0.4699	0.5994	0.3487	0.3906
SASRec	0.5186	0.6829	0.3712	0.4242	0.4027	0.5439	0.2751	0.3210	0.4805	0.6083	0.3587	0.4000
TiSASRec	0.5313	0.6882	0.3812	0.4322	0.4014	0.5433	0.2745	0.3209	0.5114	0.6329	0.3860	0.4253
RCF	0.5101	0.6660	0.3635	0.4137	0.4145	0.5696	0.2911	0.3413	0.5790	0.7004	0.4475	0.4868
RCF_{LRD}	0.5398‡	0.6882‡	0.3886‡	0.4365‡	0.4381‡	0.5761‡	0.3127‡	0.3573‡	0.5828†	0.7035†	0.4510†	0.4901†
Impro.	+5.82%	+3.33%	+6.91%	+5.51%	+5.69%	+1.14%	+7.42%	+4.69%	+0.66%	+0.44%	+0.78%	+0.68%
KDA	0.5748	0.7381	0.4182	0.4711	0.4453	0.6145	0.3127	0.3676	0.6008	0.7194	0.4665	0.5049
KDA_{LRD}	**0.6066**‡	**0.7434**‡	**0.4420**‡	**0.4867**‡	**0.4826**‡	**0.6302**‡	**0.3403**‡	**0.3881**‡	**0.6111**‡	**0.7295**‡	**0.4760**‡	**0.5143**‡
Impro.	+5.53%	+0.72%	+5.69%	+3.31%	+8.38%	+2.55%	+8.83%	+5.58%	+1.71%	+1.40%	+2.04%	+1.86%

	Datasets	MovieLens	Offices	Electronics
User-Item Interactions	#user	943	4,905	192,403
	#item	1,349	2,420	63,001
	#inter.	99,287	53,258	1,682,498
	density	7.805%	0.448%	0.014%
Item Relations	#relation	2	4	4
	#triplets	886K	778K	2,148M

Sequential Recommendation with Latent Relations based on Large Language Model **（SIGIR 2024）**

实现方法

- **LLM 作为增强器（profile）**

Representation Learning with Large Language Models for Recommendation

Xubin Ren, University of Hong Kong, xubinrencs@gmail.com
Wei Wei, University of Hong Kong, weiweics@connect.hku.hk
Lianghao Xia, University of Hong Kong, aka_xia@foxmail.com
Lixin Su, Baidu Inc., sulixinict@gmail.com
Suqi Cheng, Baidu Inc., chengsuqi@gmail.com
Junfeng Wang, Baidu Inc., wangjunfeng@baidu.com
Dawei Yin, Baidu Inc., yindawei@acm.org
Chao Huang*, University of Hong Kong, chaohuang75@gmail.com

Representation Learning with Large Language Models for Recommendation **（WWW 2024）**

LLMs-as-Enhancers (profile)

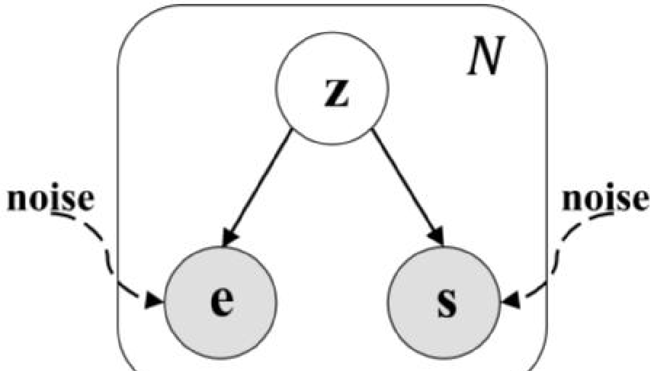
- **Motivation**

现有基于图神经网络（GNNs）的用户建模算法普遍仅依赖于ID数据构造的**结构化拓扑信息**，导致其忽略了大量存在于数据集中与用户和物品相关的原始文本数据，因此，其学习到的用户表征不够信息丰富。协同过滤的数据**存在有潜在的噪声和偏差**，也影响了对用户的建模。

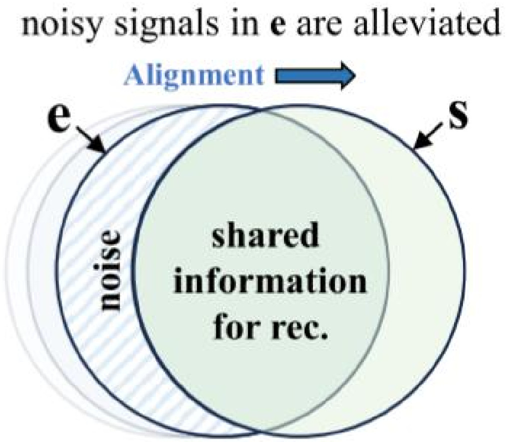
Representation Learning with Large Language Models for Recommendation（**WWW 2024**）

LLMs-as-Enhancers (profile)

- 文本特征和协同过滤特征之间的共性信息



e - CF-side rational representation
s – LLMs-enhanced semantic representation
z - hidden prior benefit for recommendation



Learning Mechanism

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} \mathbb{E}_{p(\mathbf{e},\mathbf{s})}[p(\mathbf{z},\mathbf{s}|\mathbf{e})].$$

最大化协同表征与文本表征以及潜在先验之间的一致性

$$\mathbb{E}\log[\frac{f(\mathbf{s}_i,\mathbf{e}_i)}{\sum_{\mathbf{s}_j\in\mathbf{S}} f(\mathbf{s}_j,\mathbf{e}_i)}].$$

最大化协同表征和文本表征之间的互信息

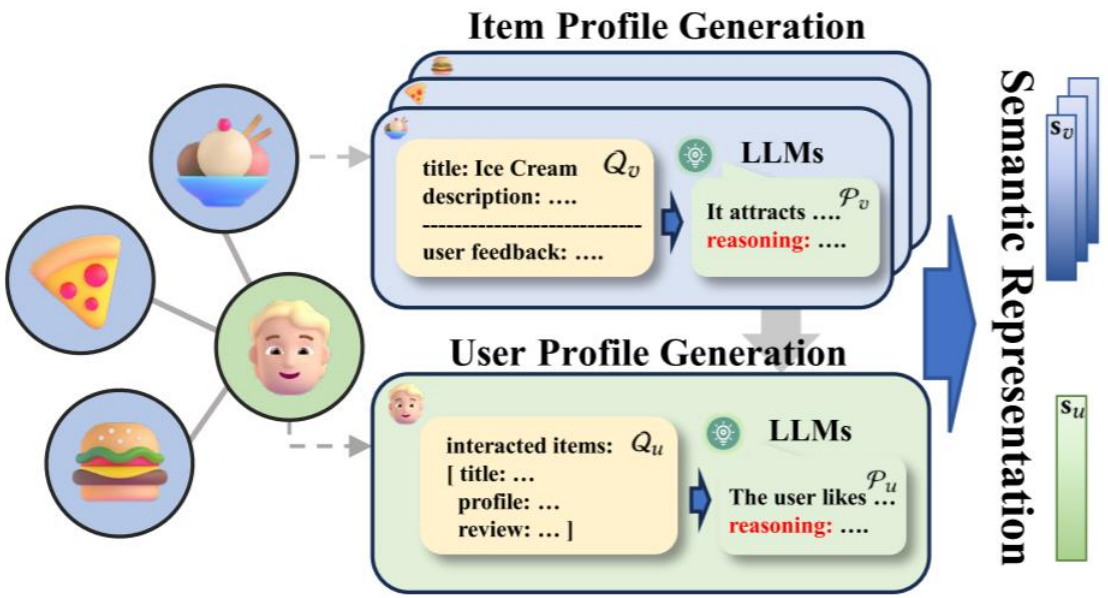
Representation Learning with Large Language Models for Recommendation（WWW 2024）

LLMs-as-Enhancers (profile)

- 如何获得高质量的文本语义表征

用户的画像：关于他们喜欢哪些类别的商品

商品的画像：能够吸引哪些目标用户群体



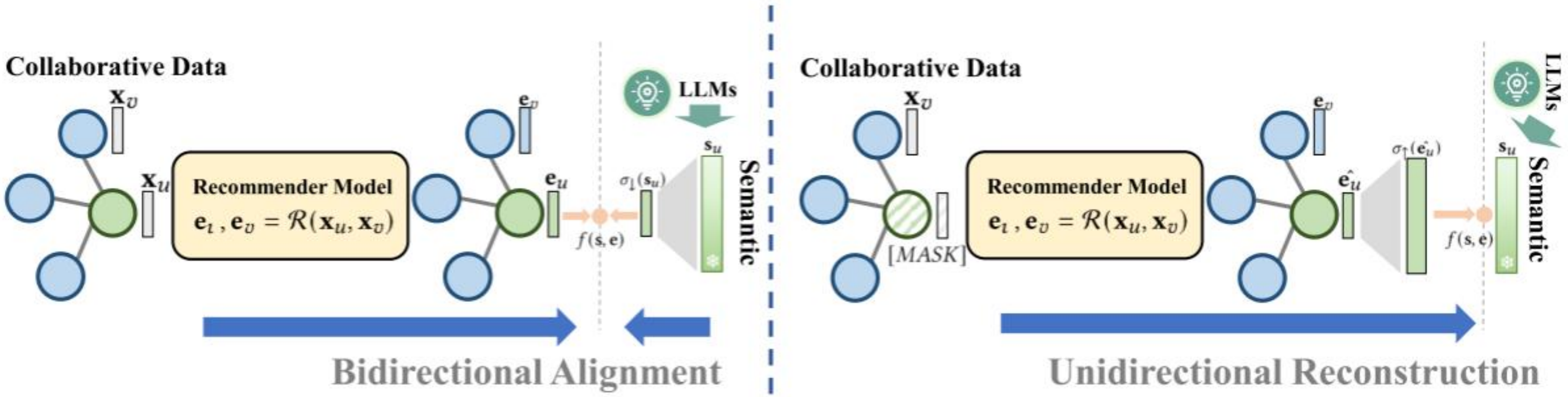
(a) Profile Generation via Reasoning

Representation Learning with Large Language Models for Recommendation（WWW 2024）

LLMs-as-Enhancers (profile)

- 文本特征和协同过滤特征之间的对齐

协同表征和文本表征的对齐 ➡ 更好的实现互信息最大化 ➡ 获得更好的用户 \ 物品表征



(b) Contrastive Alignment
(c) Generative Alignment

Representation Learning with Large Language Models for Recommendation（WWW 2024）

LLMs-as-Enhancers (profile)

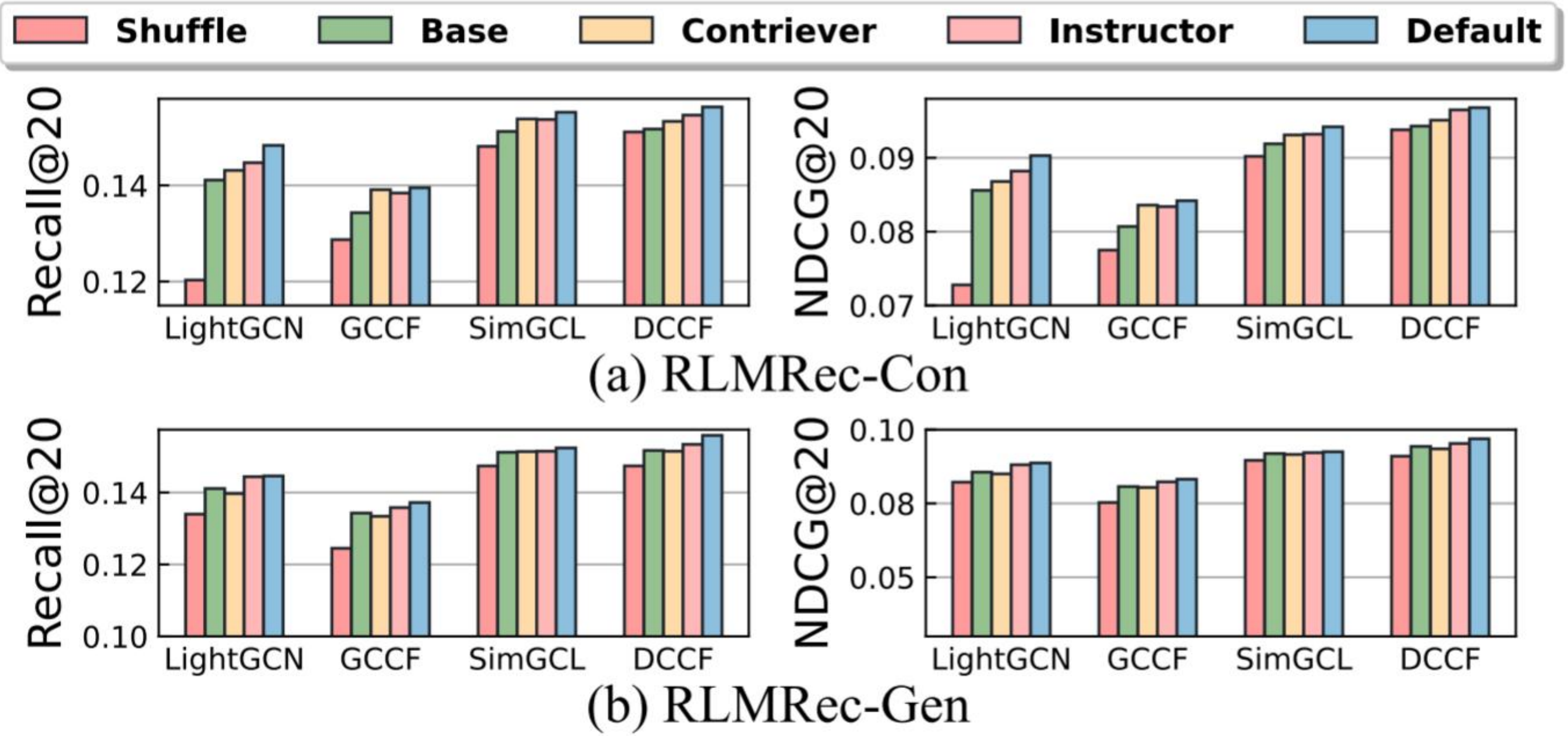
- **Experiments**

Data		Amazon-book						Yelp						Steam					
Backbone	Variants	R@5	R@10	R@20	N@5	N@10	N@20	R@5	R@10	R@20	N@5	N@10	N@20	R@5	R@10	R@20	N@5	N@10	N@20
Semantic Embeddings Only		0.0081	0.0125	0.0199	0.0072	0.0088	0.0112	0.0013	0.0022	0.0047	0.0014	0.0018	0.0026	0.0033	0.0062	0.0120	0.0031	0.0043	0.0064
GCCF	Base	0.0537	0.0872	0.1343	0.0537	0.0653	0.0807	0.0390	0.0652	0.1084	0.0451	0.0534	0.0680	0.0500	0.0826	0.1313	0.0556	0.0665	0.0830
	RLMRec-Con	**0.0561***	**0.0899***	**0.1395***	**0.0562***	**0.0679***	**0.0842***	**0.0409***	**0.0685***	**0.1144***	**0.0474***	**0.0562***	**0.0719***	**0.0538***	**0.0883***	**0.1398***	**0.0597***	**0.0713***	**0.0888***
	RLMRec-Gen	0.0551*	0.0891*	0.1372*	0.0559*	0.0675*	0.0832*	0.0393	0.0654	0.1074	0.0454	0.0535	0.0678	0.0532*	0.0874*	0.1385*	0.0588*	0.0702*	0.0875*
	Best Imprv.	↑4.28%	↑3.10%	↑3.87%	↑4.66%	↑3.98%	↑4.34%	↑4.87%	↑5.06%	↑5.54%	↑5.10%	↑5.24%	↑5.74%	↑7.60%	↑6.90%	↑6.47%	↑7.37%	↑7.22%	↑6.99%
LightGCN	Base	0.0570	0.0915	0.1411	0.0574	0.0694	0.0856	0.0421	0.0706	0.1157	0.0491	0.0580	0.0733	0.0518	0.0852	0.1348	0.0575	0.0687	0.0855
	RLMRec-Con	**0.0608***	**0.0969***	**0.1483***	**0.0606***	**0.0734***	**0.0903***	**0.0445***	**0.0754***	**0.1230***	**0.0518***	**0.0614***	**0.0776***	0.0548*	0.0895*	0.1421*	**0.0608***	0.0724*	0.0902*
	RLMRec-Gen	0.0596*	0.0948*	0.1446*	0.0605*	0.0724*	0.0887*	0.0435*	0.0734*	0.1209*	0.0505	0.0600*	0.0761*	**0.0550***	**0.0907***	**0.1433***	0.0607*	**0.0729***	**0.0907***
	Best Imprv.	↑6.67%	↑5.90%	↑5.10%	↑5.57%	↑5.76%	↑5.49%	↑5.70%	↑6.80%	↑6.31%	↑5.50%	↑5.86%	↑5.87%	↑6.18%	↑6.46%	↑6.31%	↑5.74%	↑6.11%	↑6.08%
SGL	Base	0.0637	0.0994	0.1473	0.0632	0.0756	0.0913	0.0432	0.0722	0.1197	0.0501	0.0592	0.0753	0.0565	0.0919	0.1444	0.0618	0.0738	0.0917
	RLMRec-Con	**0.0655***	**0.1017***	0.1528*	**0.0652***	**0.0778***	0.0945*	0.0452*	0.0763*	0.1248*	0.0530*	0.0626*	0.0790*	**0.0589***	**0.0956***	**0.1489***	**0.0645***	**0.0768***	**0.0950***
	RLMRec-Gen	0.0644	0.1015	**0.1537***	0.0648*	0.0777*	**0.0947***	**0.0467***	**0.0771***	**0.1263***	**0.0537***	**0.0631***	**0.0798***	0.0574*	0.0940*	0.1476*	0.0629*	0.0752*	0.0934*
	Best Imprv.	↑2.83%	↑2.31%	↑4.34%	↑3.16%	↑2.91%	↑3.72%	↑8.10%	↑6.79%	↑5.51%	↑7.19%	↑6.59%	↑5.98%	↑5.20%	↑4.03%	↑3.12%	↑4.37%	↑4.07%	↑3.60%
SimGCL	Base	0.0618	0.0992	0.1512	0.0619	0.0749	0.0919	0.0467	0.0772	0.1254	0.0546	0.0638	0.0801	0.0564	0.0918	0.1436	0.0618	0.0738	0.0915
	RLMRec-Con	**0.0633***	**0.1011***	**0.1552***	**0.0633***	**0.0765***	**0.0942***	**0.0470**	**0.0784***	**0.1292***	**0.0546**	**0.0642**	**0.0814***	**0.0582***	**0.0945***	**0.1482***	**0.0638***	**0.0760***	**0.0942***
	RLMRec-Gen	0.0617	0.0991	0.1524*	0.0622	0.0752	0.0925*	0.0464	0.0767	0.1267	0.0541	0.0634	0.0803	0.0572	0.0929	0.1456*	0.0627*	0.0747*	0.0926*
	Best Imprv.	↑2.43%	↑1.92%	↑2.65%	↑2.26%	↑2.14%	↑2.50%	↑0.64%	↑1.55%	↑3.03%	–	↑0.63%	↑1.62%	↑3.19%	↑2.94%	↑1.53%	↑3.24%	↑2.98%	↑2.95%
DCCF	Base	0.0662	0.1019	0.1517	0.0658	0.0780	0.0943	0.0468	0.0778	0.1249	0.0543	0.0640	0.0800	0.0561	0.0915	0.1437	0.0618	0.0736	0.0914
	RLMRec-Con	0.0665	0.1040*	**0.1563***	0.0668	0.0798*	0.0968*	**0.0486***	**0.0813***	**0.1321***	**0.0561***	**0.0663***	**0.0836***	**0.0572***	**0.0929***	**0.1459***	**0.0627***	**0.0747***	**0.0927***
	RLMRec-Gen	**0.0666**	**0.1046***	0.1559*	**0.0670***	**0.0801***	**0.0969***	0.0475	0.0785	0.1281*	0.0549	0.0646	0.0815	0.0570*	0.0918	0.1430	0.0625	0.0741	0.0915
	Best Imprv.	↑0.60%	↑2.65%	↑3.03%	↑1.82%	↑2.69%	↑2.76%	↑3.85%	↑4.50%	↑5.76%	↑3.31%	↑3.59%	↑4.50%	↑2.14%	↑1.53%	↑1.53%	↑1.46%	↑1.49%	↑1.42%
AutoCF	Base	0.0689	0.1055	0.1536	**0.0705**	0.0828	0.0984	0.0469	0.0789	0.1280	0.0547	0.0647	0.0813	0.0519	0.0853	0.1358	0.0572	0.0684	0.0855
	RLMRec-Con	**0.0695**	**0.1083***	**0.1586***	0.0704	**0.0837**	**0.1001***	0.0488*	0.0814*	0.1319*	0.0562*	0.0663*	0.0835*	**0.0540***	0.0876*	0.1372*	0.0593*	0.0704*	0.0872*
	RLMRec-Gen	0.0693	0.1069*	0.1581*	0.0701	0.0830	0.0996	**0.0493***	**0.0828***	**0.1330***	**0.0572***	**0.0677***	**0.0848***	0.0539*	**0.0888***	**0.1410***	**0.0593***	**0.0710***	**0.0886***
	Best Imprv.	↑0.87%	↑2.65%	↑3.26%	↓0.14%	↑1.87%	↑1.73%	↑5.12%	↑4.94%	↑3.91%	↑4.57%	↑4.64%	↑4.31%	↑4.05%	↑4.10%	↑3.83%	↑3.67%	↑3.80%	↑3.63%

Representation Learning with Large Language Models for Recommendation（**WWW 2024**）

LLMs-as-Enhancers (profile)

- **Experiments**



Representation Learning with Large Language Models for Recommendation (**WWW 2024**)

实现方法

- **LLM 作为预测器（判别式生成推理）**

CoRAL: Collaborative Retrieval-Augmented Large Language Models Improve Long-tail Recommendation

Junda Wu
juw069@ucsd.edu
University of California San Diego
La Jolla, California, USA

Cheng-Chun Chang
cc4900@columbia.edu
Columbia University
New York, New York, USA

Tong Yu
tyu@adobe.com
Adobe Research
San Jose, California, USA

Zhankui He
zhh004@eng.ucsd.edu
University of California San Diego
La Jolla, California, USA

Jianing Wang
lygwjn@gmail.com
University of California San Diego
La Jolla, California, USA

Yupeng Hou
yphou@ucsd.edu
University of California San Diego
La Jolla, California, USA

Julian McAuley
jmcauley@ucsd.edu
University of California San Diego
La Jolla, California, USA

CoRAL: Collaborative Retrieval-Augmented Large Language Models Improve Long-tail Recommendation（**KDD2024**）

LLMs-as-Predictors(判别式)

- **Motivation**
  - 大多数**基于LLM直接生成用户偏好**的用户建模系统依赖于**项目的语义**作为推理的唯一证据，忽略了**用户-项目交互的协同信息**

- Contribution
  - 发现了**LLM的推理过程**和**用户真实的行为偏好之间**的隔阂是由于**缺乏协同信息**造成的。
  - 检索额外的用户-项目交互作为**协同提示**的协同信息
  - 将检索过程描述为一个顺序决策任务，并提出了一个RL框架，在该框架中，**检索策略学习找到特定于推荐任务的最小足够协同信息**。

CoRAL: Collaborative Retrieval-Augmented Large Language Models Improve Long-tail Recommendation（**KDD2024**）

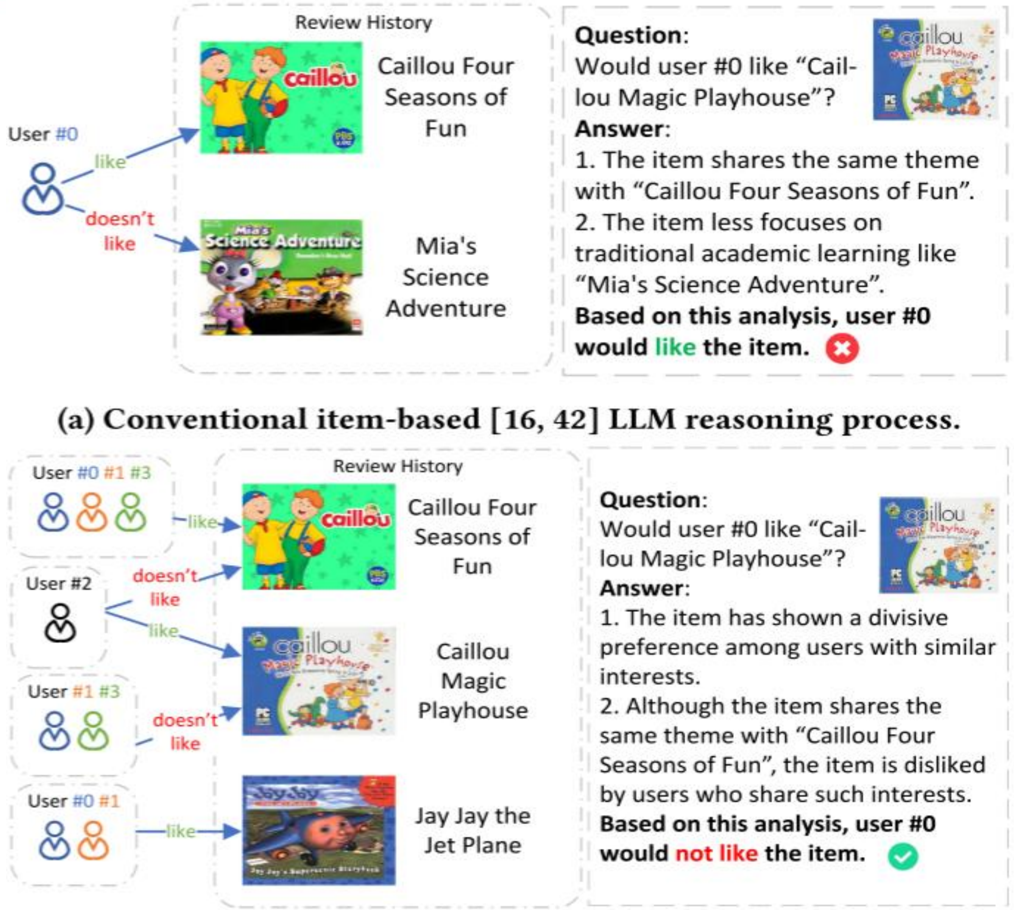
LLMs-as-Predictors(判别式)

- **Method**

We formulate the sequential retrieval process as a Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \rho, \gamma)$,

At each time step t, the retrieval policy π_θ needs to retrieve the next user-item pair (u^z_{t+1}, i^z_{t+1}) to augment current supporting evidence. In this work, we focus on how to obtain a minimal-sufficient information support for the LLM to deduce the accurate rating of z.

$$r_t\left(s_t, (u^z_t, i^z_t)\right) = \underbrace{\left|p_{t-1} - y^{gt}\right|}_{\text{discrepancy at } t-1} - \underbrace{\left|p_t - y^{gt}\right|}_{\text{discrepancy at } t},$$



(a) Conventional item-based [16, 42] LLM reasoning process.

(b) Collaborative Retrieval Augmented LLM reasoning process.

CoRAL: Collaborative Retrieval-Augmented Large Language Models Improve Long-tail Recommendation（**KDD2024**）

LLMs-as-Predictors(判别式)

- **Experiments**

	Software		Prime Pantry		Gift Cards		Appliances		Average	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
AFM [61]	75.12	58.39	69.47	52.51	46.93	61.56	76.86	65.52	67.10	59.49
DCN [50]	76.75	66.20	73.30	49.99	55.59	67.07	80.70	71.15	71.59	63.60
DFM [14]	76.04	66.63	72.92	57.86	66.76	60.01	81.83	77.37	74.39	65.47
WDL [10]	78.20	69.25	73.77	56.43	60.81	57.66	73.82	74.56	71.65	64.48
IPS [43]	78.24	71.32	72.24	61.65	64.79	63.95	82.28	75.65	74.39	66.23
CausE [6]	77.78	70.84	73.69	59.80	70.51	65.39	76.86	72.04	74.71	67.02
LLM-Language [42]	73.10	66.32	51.48	41.47	83.52	74.85	74.36	70.52	70.61	63.29
CoRAL-random	77.56	58.60	64.07	50.15	91.30	59.66	77.51	61.35	77.61	57.44
CoRAL-DFM	95.25	88.68	**93.32**	**86.73**	96.52	67.51	90.87	86.76	**93.99**	82.42
CoRAL-WDL	93.97	**91.18**	87.08	80.52	92.22	70.74	92.55	**89.22**	91.45	82.92
CoRAL-AFM	**93.99**	88.41	89.10	86.17	**98.99**	**76.17**	**92.66**	84.55	93.69	**83.83**
CoRAL-DCN	91.74	87.20	85.75	77.59	97.16	70.63	91.73	86.28	91.59	80.43

Table 1: Experimental results (AUC and F1) on four Amazon Product datasets.

CoRAL: Collaborative Retrieval-Augmented Large Language Models Improve Long-tail Recommendation **（KDD2024）**

实现方法

- **LLM 作为预测器（生成式推理）**

A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems

KEQIN BAO*, University of Science and Technology of China, China
JIZHI ZHANG*, University of Science and Technology of China, China
WENJIE WANG, National University of Singapore, Singapore
YANG ZHANG, University of Science and Technology of China, China
ZHENGYI YANG, University of Science and Technology of China, China
YANCHENG LUO, University of Science and Technology of China, China
CHONG CHEN, Huawei Inc., China
FULI FENG, University of Science and Technology of China, China
QI TIAN, Huawei Inc., China

A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems（arxiv2312）

LLMs-as-Predictors(生成式)

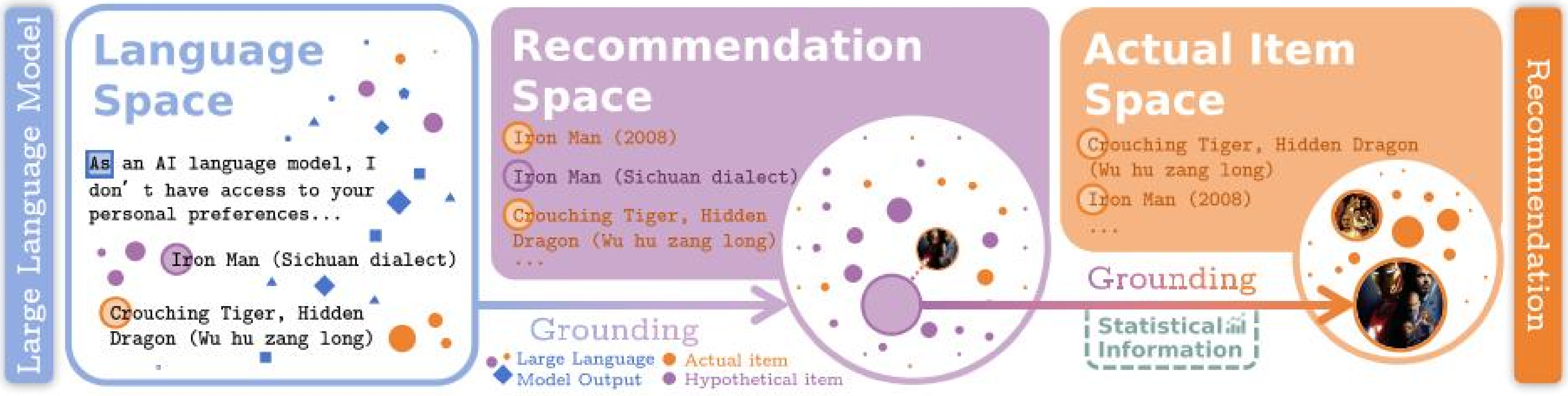
- **Motivation**
  - 现有的微调LLM的工作只能针对候选集有限的场景。

- **Contribution**
  - 研究了在全排名设置中的LLM4Rec，并引入了一个两步Grounding范式
  - 将流行度整合到BIGRec中，并揭示了这一信息在LLM4Rec中的益处

A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems（arxiv2312）

LLMs-as-Predictors(生成式)

- **Method**
  - **Grounding Language Space to Recommendation Space.**
  - **Grounding Recommendation Space to Actual ItemsSpace.**



A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems（arxiv2312）

LLMs-as-Predictors(生成式)

- **Method**
  - **Grounding Language Space to Recommendation Space.**

Table 1. Example of the instruction-tuning data for the step of grounding to the space.

Instruction Input	
Instruction:	Given ten movies that the user watched recently, please recommend a new movie that the user likes to the user.
Input:	The user has watched the following movies before: "Traffic (2000)", "Ocean's Eleven (2001)", ... "Fargo (1996)"
Instruction Output	
Output:	"Crouching Tiger, Hidden Dragon (Wu hu zang long) (2000)"

A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems（arxiv2312）

LLMs-as-Predictors(生成式)

- **Method**
  - **Grounding Recommendation Space to Actual ItemsSpace.**

$$D_i = ||\mathbf{emb}_i - \mathbf{oracle}||_2,$$

$$\begin{cases} C_i = \dfrac{N^i}{\sum_{j \in \mathcal{I}} N^j}, \\ P_i = \dfrac{C_i - \min_{j \in \mathcal{I}}\{C_j\}}{\max_{j \in \mathcal{I}}\{C_j\} - \min_{j \in \mathcal{I}}\{C_j\}}, \end{cases}$$

$$\begin{cases} \hat{D}_i = \dfrac{D_i - \min_{j \in \mathcal{I}}\{D_j\}}{\max_{j \in \mathcal{I}}\{D_j\} - \min_{j \in \mathcal{I}}\{D_j\}}, \\ \widetilde{D}_i = \dfrac{\hat{D}_i}{(1 + P_i)^{\gamma}}, \end{cases}$$

A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems（arxiv2312）

LLMs-as-Predictors(生成式)

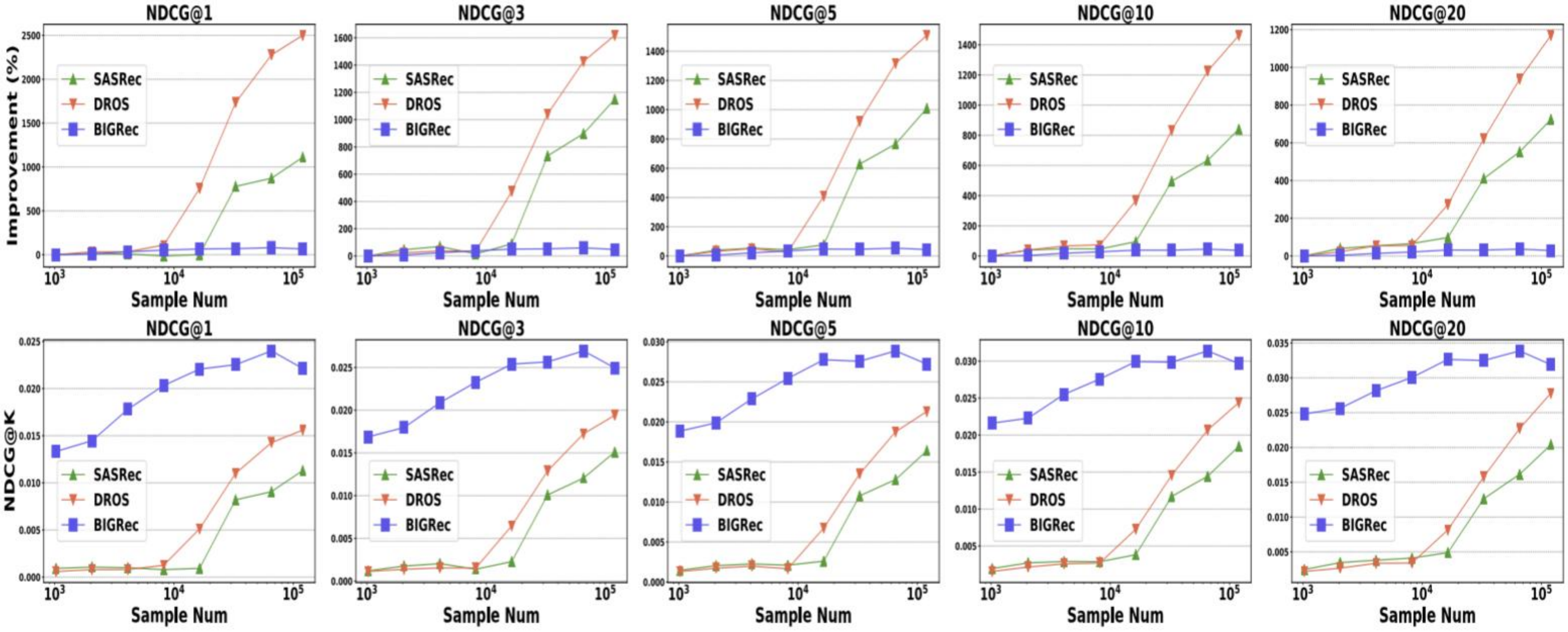
- **Experiment**

Dataset	Model	NG@1	NG@3	NG@5	NG@10	NG@20	HR@1	HR@3	HR@5	HR@10	HR@20
Movie	GRU4Rec	0.0015	0.0034	0.0047	0.0070	0.0104	0.0015	0.0047	0.0079	0.0147	0.0281
	Caser	0.0020	0.0035	0.0052	0.0078	0.0109	0.0020	0.0046	0.0088	0.0171	0.0293
	SASRec	0.0023	0.0051	0.0062	0.0082	0.0117	0.0023	0.0070	0.0097	0.0161	0.0301
	P5	0.0014	0.0026	0.0036	0.0051	0.0069	0.0014	0.0035	0.0059	0.0107	0.0176
	DROS	0.0022	0.0040	0.0052	0.0081	0.0112	0.0022	0.0051	0.0081	0.0173	0.0297
	GPT4Rec-LLaMA	0.0016	0.0022	0.0024	0.0028	0.0035	0.0016	0.0026	0.0030	0.0044	0.0074
	BIGRec (1024)	**0.0176**	**0.0214**	**0.0230**	**0.0257**	**0.0283**	**0.0176**	**0.0241**	**0.0281**	**0.0366**	**0.0471**
	Improve	**654.29%**	**323.31%**	**273.70%**	**213.71%**	**142.55%**	**654.29%**	**244.71%**	**188.39%**	**111.97%**	**56.55%**
Game	GRU4Rec	0.0013	0.0016	0.0018	0.0024	0.0030	0.0013	0.0018	0.0024	0.0041	0.0069
	Caser	0.0007	0.0012	0.0019	0.0024	0.0035	0.0007	0.0016	0.0032	0.0048	0.0092
	SASRec	0.0009	0.0012	0.0015	0.0020	0.0025	0.0009	0.0015	0.0021	0.0037	0.0057
	P5	0.0002	0.0005	0.0007	0.0010	0.0017	0.0002	0.0007	0.0012	0.0023	0.0049
	DROS	0.0006	0.0011	0.0013	0.0016	0.0022	0.0006	0.0015	0.0019	0.0027	0.0052
	GPT4Rec-LLaMA	0.0000	0.0000	0.0000	0.0001	0.0001	0.0000	0.0000	0.0000	0.0002	0.0002
	BIGRec (1024)	**0.0133**	**0.0169**	**0.0189**	**0.0216**	**0.0248**	**0.0133**	**0.0195**	**0.0243**	**0.0329**	**0.0457**
	Improve	**952.63%**	**976.26%**	**888.19%**	**799.64%**	**613.76%**	**952.63%**	**985.19%**	**660.42%**	**586.11%**	**397.10%**

A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems（arxiv2312）

LLMs-as-Predictors(生成式)

- **Experiment**



A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems（arxiv2312）

LLMs-as-Controllers

On Generative Agents in Recommendation

An Zhang*
National University of Singapore
Singapore
anzhang@u.nus.edu

Yuxin Chen*
National University of Singapore
Singapore
e1143404@u.nus.edu

Leheng Sheng*
Tsinghua University
Beijing, China
chenglh22@mails.tsinghua.edu.cn

Xiang Wang†
University of Science and Technology of China
Hefei, China
xiangwang1223@gmail.com

Tat-Seng Chua
National University of Singapore
Singapore
dcscts@nus.edu.s

On Generative Agents in Recommendation（arxiv2405）

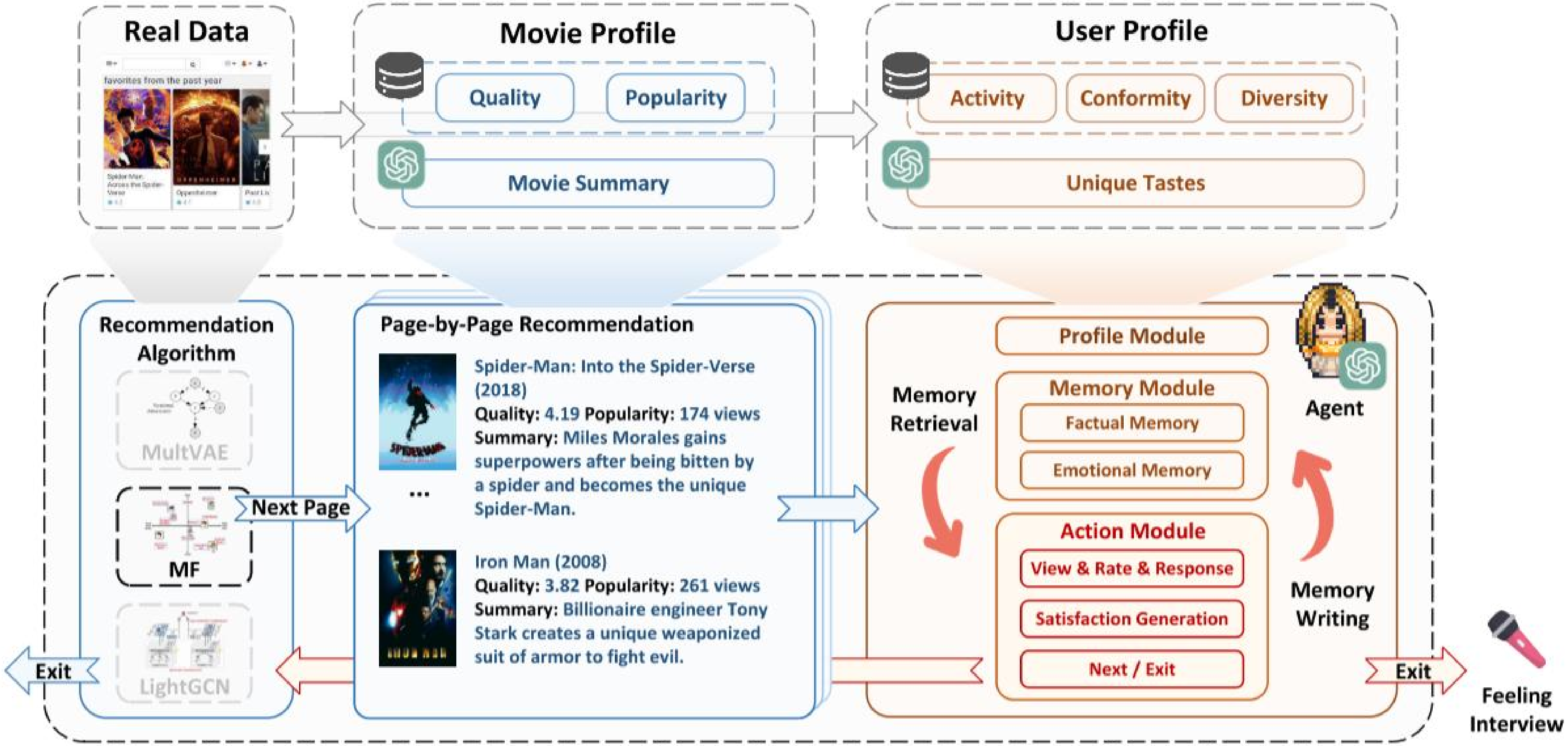
LLMs-as-Controllers

- Motivation
  - 现有推荐系统领域中，离线指标和在线表现之前存在巨大的脱节，阻碍了推荐系统的发展

- Contribution
  - 开发Agent4Rec框架，利用LLM来生成Agent模拟用户的个性化偏好和行为模式
  - 提出一种考虑离线性能和仿真反馈的双重评估方法

On Generative Agents in Recommendation（arxiv2405）

LLMs-as-Controllers



On Generative Agents in Recommendation（arxiv2405）

LLMs-as-Controllers

Experiments

Offline	MF Recall	MF NDCG	MultVAE Recall	MultVAE NDCG	LightGCN Recall	LightGCN NDCG
Origin	0.1506	0.3561	0.1609	0.3512	0.1757	0.3937
+ Unviewed	0.1523	0.3557	0.1598	0.3487	0.1729	0.3849
+ Viewed	**0.1570***	**0.3604***	**0.1613***	**0.3540***	**0.1765***	**0.3943***

Simulation	MF $\overline{N}_{exit}$	MF $\overline{S}_{sat}$	MultVAE $\overline{N}_{exit}$	MultVAE $\overline{S}_{sat}$	LightGCN $\overline{N}_{exit}$	LightGCN $\overline{S}_{sat}$
Origin	3.17	3.80	3.10	3.75	3.02	3.85
+ Unviewed	3.03	3.77	3.01	3.77	3.06	3.81
+ Viewed	**3.27***	**3.83***	**3.18***	**3.87***	**3.10***	**3.92***

On Generative Agents in Recommendation（arxiv2405）

挑战和方向

- 挑战和方向
  - 可疑性检测
  - 幻觉缓解
  - 隐私与安全
  - 复杂数据结构的理解
  - 评估标准
  - 可信的用户建模
  - 用户建模的公平性
  - 效率问题
- 补充
  - LLM+RAG for UM
  - LLM的公平性与用户的个性化之间的平衡
  - LLM 赋能 跨领域用户行为建模
  - 实时性与可扩展性

谢 谢

2024-8-30　董彦

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS