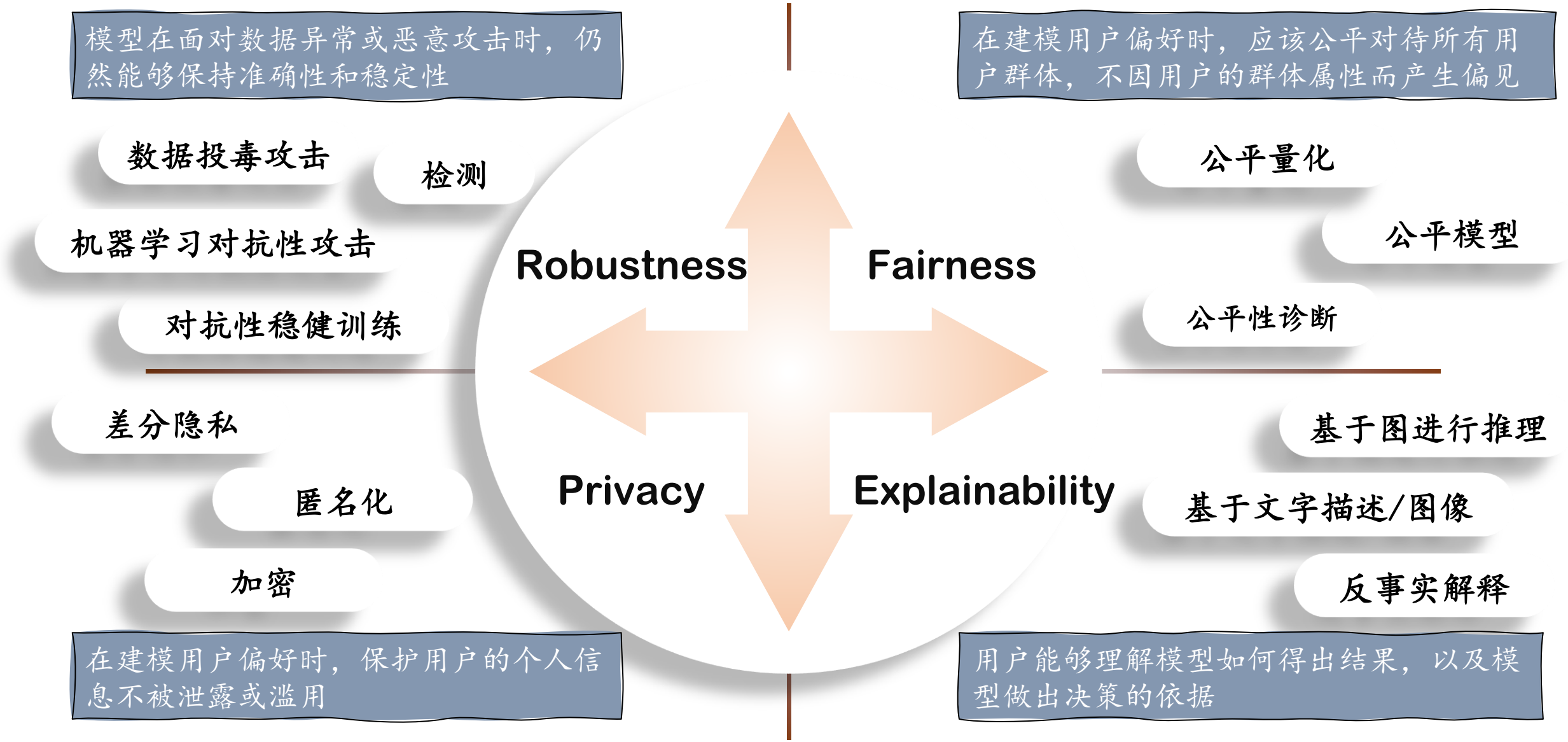
中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

Trustworthy User Preference Modeling

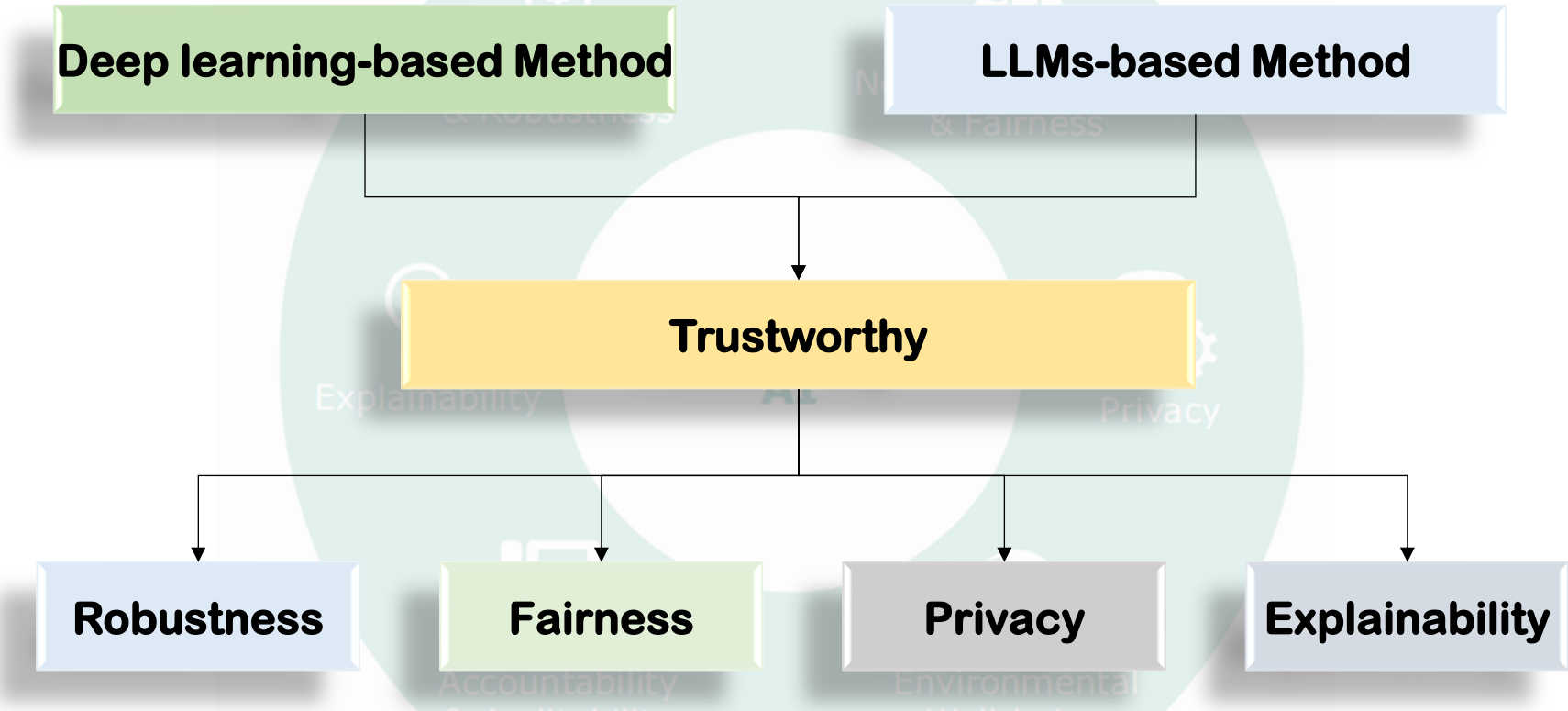
2024.8.30 吴咏萱

What is Trustworthy

What is Trustworthy

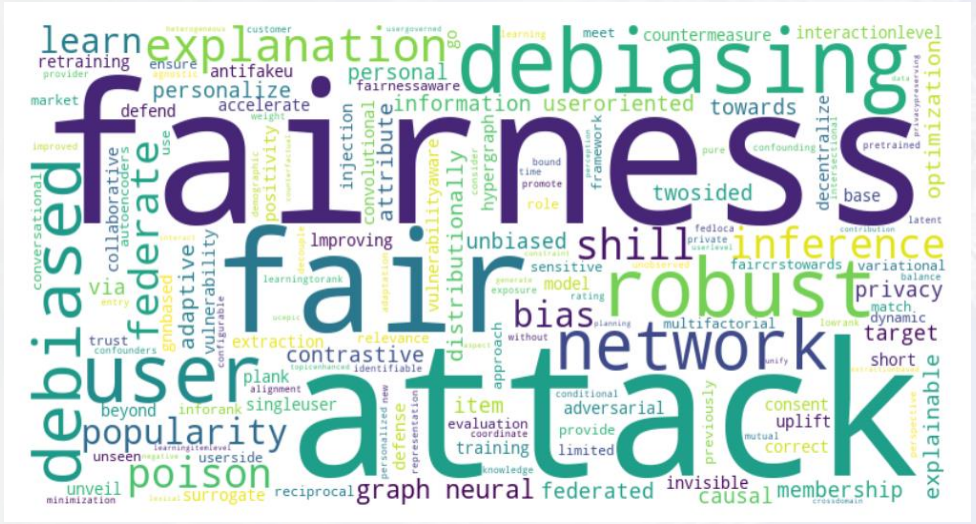


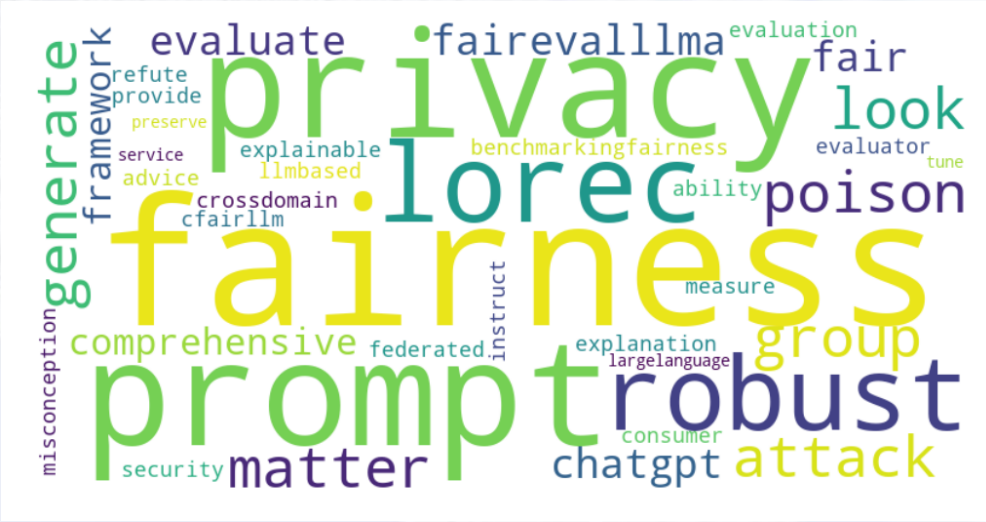
Outline


Deep learning-based Method
LLMs-based Method
Trustworthy
Robustness
Fairness
Privacy
Explainability

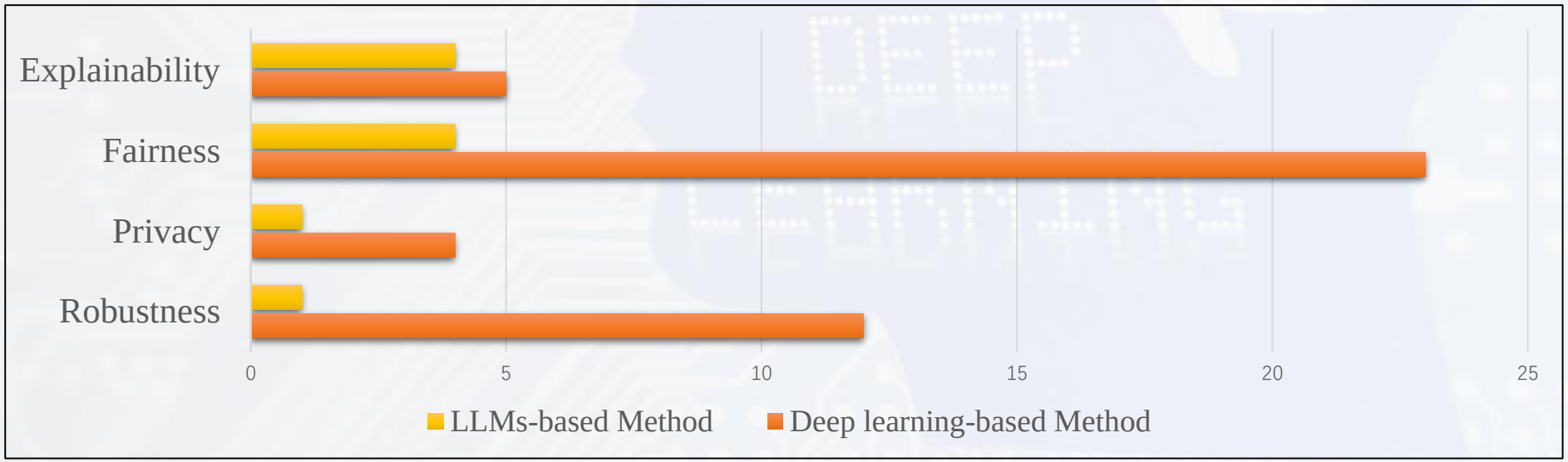
Development Vein

Development Vein





Deep learning-based Method **LLMs-based Method**

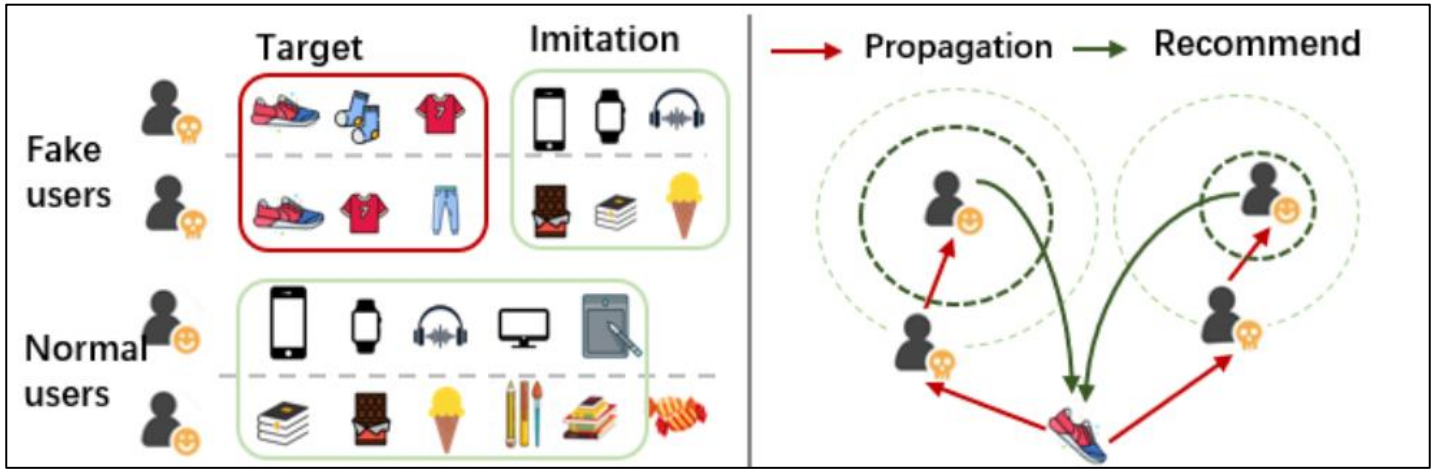


Deep learning-based Method

Robustness--😈Attack

😈数据投毒攻击

攻击者向训练数据中注入恶意或伪造的用户数据，导致模型对普通用户偏好的建模不准确，进而影响后续任务。



- 生成式攻击者：通过**生成恶意用户**进行攻击
- 基于优化的攻击者：通过代理推荐模型**迭代优化恶意用户**，以最大化精心设计的攻击目标
- 神经网络攻击者：优化神经网络以**生成有影响力的假用户**，以更大程度的影响更多用户
- 提升模型攻击者：以确定最佳攻击中**恶意用户预算分配**为目标，最大限度地提高整体攻击性能

Robustness-- Defense

数据投毒攻击防御

Anti-FakeU: Defending Shilling Attacks on Graph Neural Network based Recommender Model

Xiaoyu You
Chi Li
17212010047@fudan.edu.cn
20210240215@fudan.edu.cn
School of Computer Science, Fudan University
China

Daizong Ding
Mi Zhang*
17110240010@fudan.edu.cn
mi_zhang@fudan.edu.cn
School of Computer Science, Fudan University
China

Fuli Feng
fulifeng93@gmail.com
University of Science and Technology of China
China

Xudong Pan
Min Yang*
mi_zhang@fudan.edu.cn
m_yang@fudan.edu.cn
School of Computer Science, Fudan University
China

Citations:4

Anti-FakeU: Defending Shilling Attacks on Graph Neural Network based Recommender Model(WWW 2023)

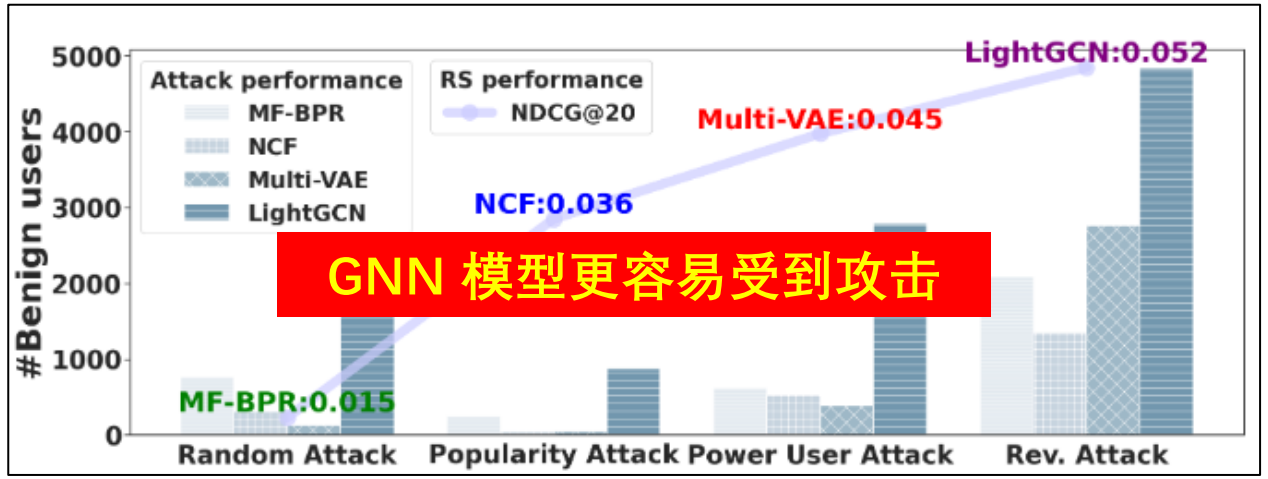
Robustness-- Defense

数据投毒攻击防御

对抗性训练方法

基于检测的方法

如何提高GNN推荐模型的对于数据投毒攻击的鲁棒性



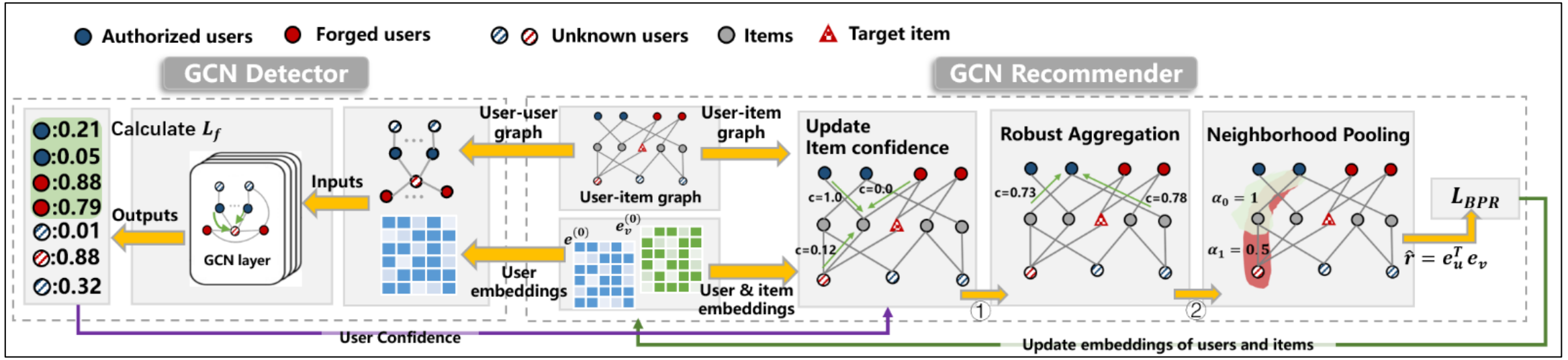
当聚合度较高时，节点更有可能依赖其邻居的信息，这为恶意用户打开了新的攻击窗口。

Anti-FakeU: Defending Shilling Attacks on Graph Neural Network based Recommender Model(WWW 2023)

Robustness-- Defense

数据投毒攻击防御

通过检测恶意用户和阻断恶意用户影响力传播的方法来实现防御



数据增强 → 恶意用户检测 → 检测器的置信度分数 → 邻域聚合机制修改

调用具有不同配置的各种攻击策略来主动生成一组恶意用户

恶意用户通常有相同的目标，构建**用户-用户图**

$$h_u^{(\ell+1)} = \sigma\left(\tilde{D}_U^{-\frac{1}{2}}\tilde{A}_U\tilde{D}_U^{-\frac{1}{2}}h_u^{(\ell)}W^{\ell}\right)$$

$$z_u = \text{sigmoid}([W^o]^T \cdot h^{(L)} + b^o)$$

依据置信度分数调整邻域聚合权重

- 邻居节点的聚合
- 多跳的聚合传播

Anti-FakeU: Defending Shilling Attacks on Graph Neural Network based Recommender Model(WWW 2023)

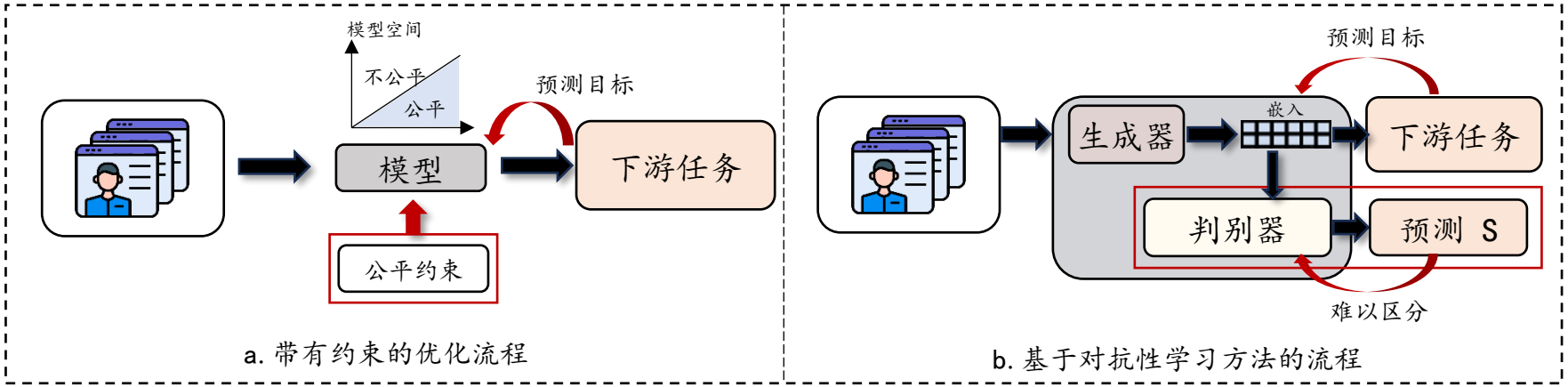
Fairness

概述

- 基于敏感属性(性别、种族等)
- 基于活跃性

用户公平：

- 确保不同用户群组获得公平的机会
  - 随着时间推移确保不同用户群体不会因模型更新或用户行为变化导致偏见(用户动态公平)
  - 不同的用户对不同的属性敏感，并且具有个性化的公平性要求(**用户个性化公平**)



解决用户公平性问题常用方法

Fairness

基于敏感属性的用户个性化公平问题

基于敏感属性组合的过滤方法

基于提示的偏差消除方法

基于解耦表示学习的方法

Adaptive Fair Representation Learning for Personalized Fairness in Recommendations via Information Alignment

Xinyu Zhu*
School of Computer Science
Sichuan University
Chengdu, China
zhuxinyu@stu.scu.edu.cn

Lilin Zhang*
School of Computer Science
Sichuan University
Chengdu, China
zhanglilin@stu.scu.edu.cn

Ning Yang†
School of Computer Science
Sichuan University
Chengdu, China
yangning@scu.edu.cn

属性的指数组合导致训练成本激增；

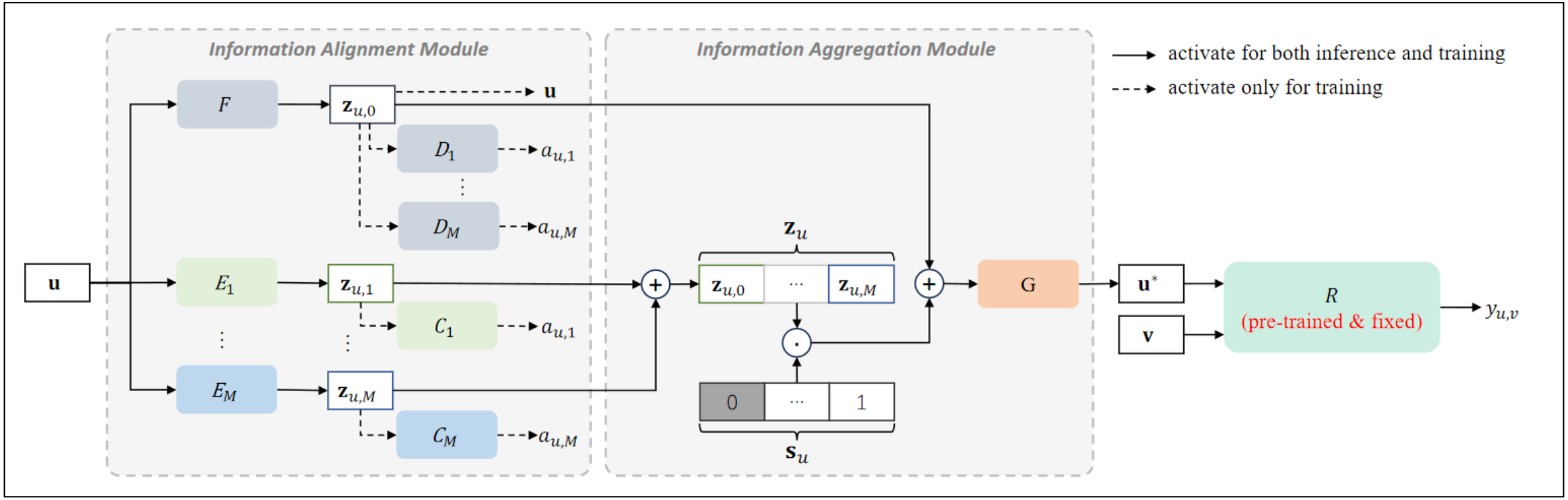
从嵌入中完全删除敏感属性信息的同时也会删除与非敏感属性重叠的信息，导致用户偏好建模效果不必要的降低

Adaptive Fair Representation Learning for Personalized Fairness in Recommendations via Information Alignment(SIGIR 2024)

Fairness

用户个性化公平问题

- ✓ 将用户公平性要求视为输入，自适应地为不同用户生成公平嵌入
- ✓ 通过精确的信息对齐和去偏协同嵌入实现准确性与公平性的权衡

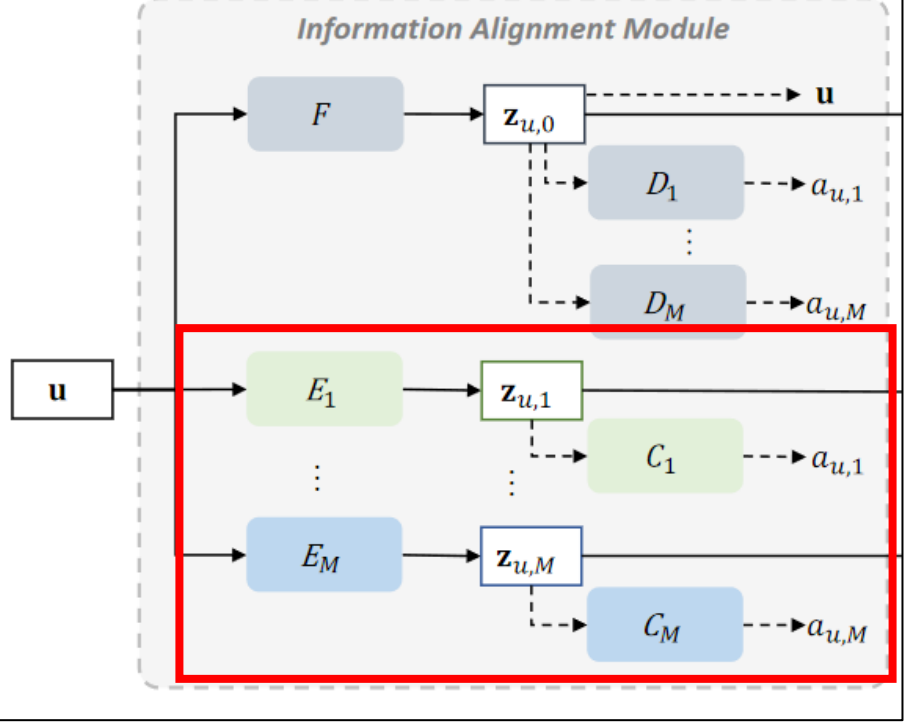


Adaptive Fair Representation Learning for Personalized Fairness in Recommendations via Information Alignment(SIGIR 2024)

Fairness

用户个性化公平问题

信息对齐模块从用户的原始嵌入中为每个属性学习属性特定的嵌入，确保嵌入和相应属性之间的信息精确对齐



最小化 Z_i 和 U 之间的互信息

$$I(Z_i;U)=\mathbb{E}_{(\mathrm{z}_{u,i},\mathrm{u})\sim p(Z_i,U)}\left[\log\frac{p(\mathrm{z}_{u,i}|\mathrm{u})}{p(\mathrm{z}_{u,i})}\right]$$

$$\underset{E_i}{\operatorname{argmin}} I(Z_i;U)=\mathbb{E}_{\mathbf{u}\sim p(U)}\frac{1}{2}\parallel\mu_u\parallel_2^2 \quad (p(\mathbf{z}_{u,i}|\mathbf{u})\sim\mathcal{N}(\mu_u,0))$$

防止 Z_i 中存在与第i个属性无关的噪声或其他信息

最大化 Z_i 和 A_i 之间的互信息

$$I(Z_i;A_i)=\mathbb{E}_{(z_{u,i},a_{u,i})\sim p(Z_i,A_i)}\left[\log\frac{p(a_{u,i}|z_{u,i})}{p(a_{u,i})}\right]$$

$$I(Z_i;A_i)=\mathbb{E}_{(z_{u,i},a_{u,i})\sim p(Z_i,A_i)}[\log\frac{p(a_{u,i}|z_{u,i})q(a_{u,i}|z_{u,i};C_i)}{p(a_{u,i})q(a_{u,i}|z_{u,i};C_i)}]$$

$$=\Phi(Z_i,C_i)+\Delta(Z_i,C_i)\quad(\textbf{EM}\text{算法迭代优化})$$

确保 Z_i 能够准确地表达用户的第 i 个属性

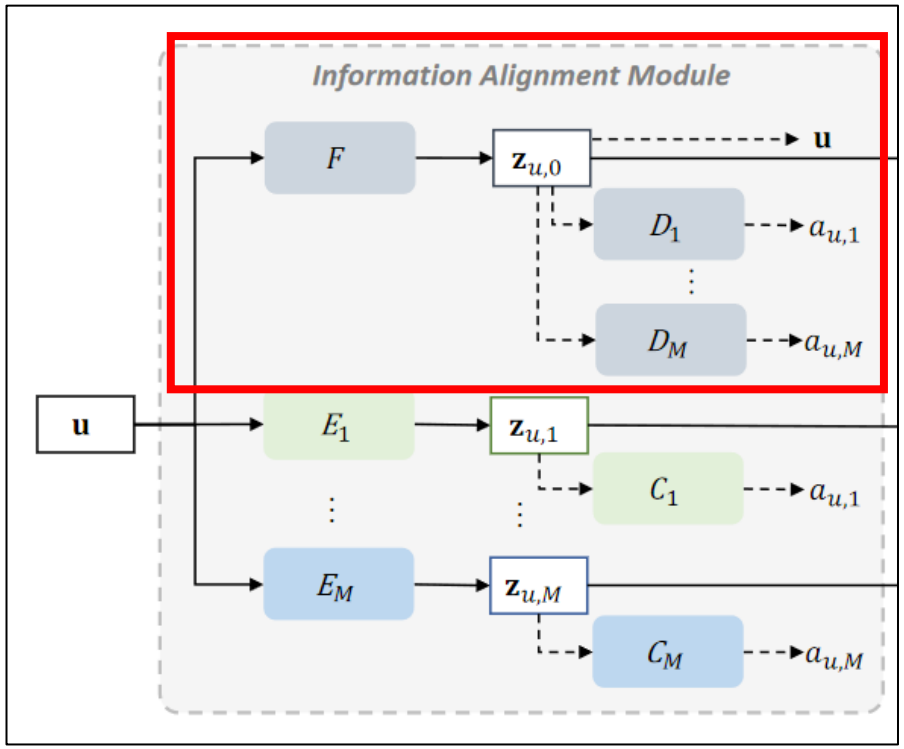
$$\min_{E_i}\min_{C_i}\mathbb{E}_{\mathbf{u}\sim p(U)}\left[\frac{1}{2}\parallel\mathbf{z}_{u,i}\parallel_2^2-\beta\log q\,(a_{u,i}|z_{u,i};C_i)\right]$$

Adaptive Fair Representation Learning for Personalized Fairness in Recommendations via Information Alignment(SIGIR 2024)

Fairness

用户个性化公平问题

通过学习去偏协同嵌入来进一步提高推荐的准确性



最小化 $Z_{u,0}$对用户属性的可区分性

$$\min_{F}\max_{D_i} -\mathbb{E}_{\mathbf{u}\sim p(U)}\sum_{i=1}^{M}\left[-\log q(a_{u,i}|\mathbf{z}_{u,0};D_i)\right]$$

（通过判别器进行对抗学习）

确保$Z_{u,0}$中不包含属性信息以实现公平

最大化 $Z_{u,0}$中编码的 u 信息

$$\min_{F}\mathbb{E}_{\mathbf{u}\sim p(U)}\parallel \mathbf{z}_{u,0}-\mathbf{u}\parallel_2^2$$

（最小化重构损失）

尽可能保留协作信息以保证推荐的准确性

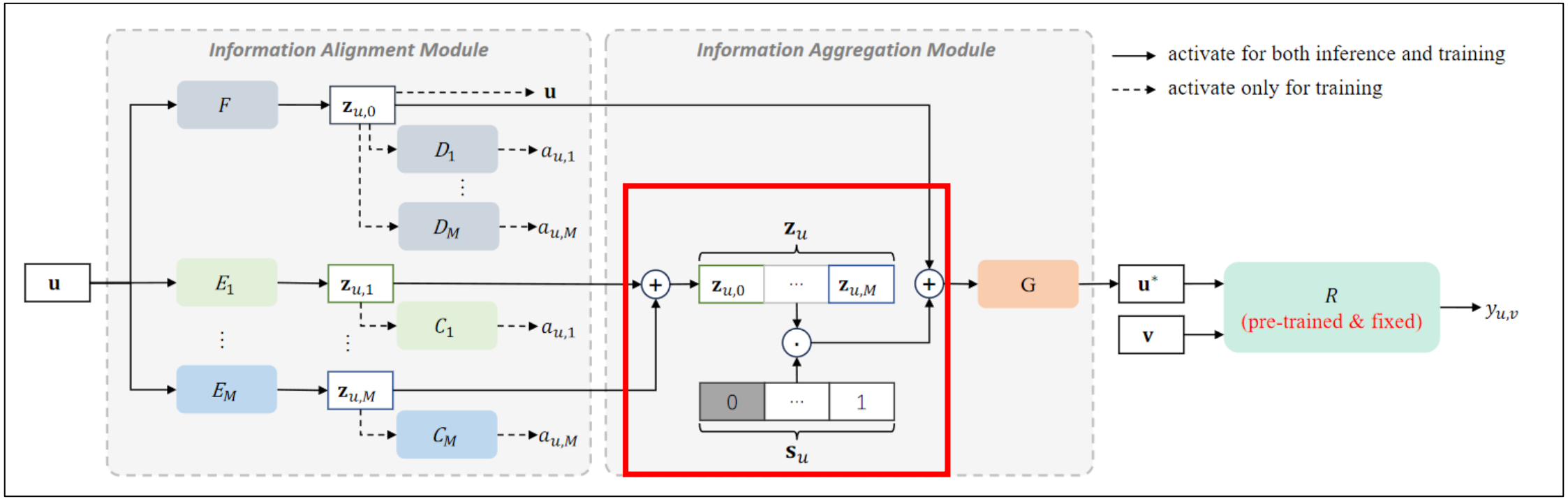
$$\min_{F}\max_{D_i}\mathbb{E}_{\mathbf{u}\sim p(U)}\parallel \mathbf{z}_{\mathbf{u},0}-\mathbf{u}\parallel_2^2-\lambda\sum_{i=1}^{M}\left[-\log q(a_{u,i}|\mathbf{z}_{\mathbf{u},0};D_i)\right]$$

Adaptive Fair Representation Learning for Personalized Fairness in Recommendations via Information Alignment(SIGIR 2024)

Fairness

⚖ 用户个性化公平问题

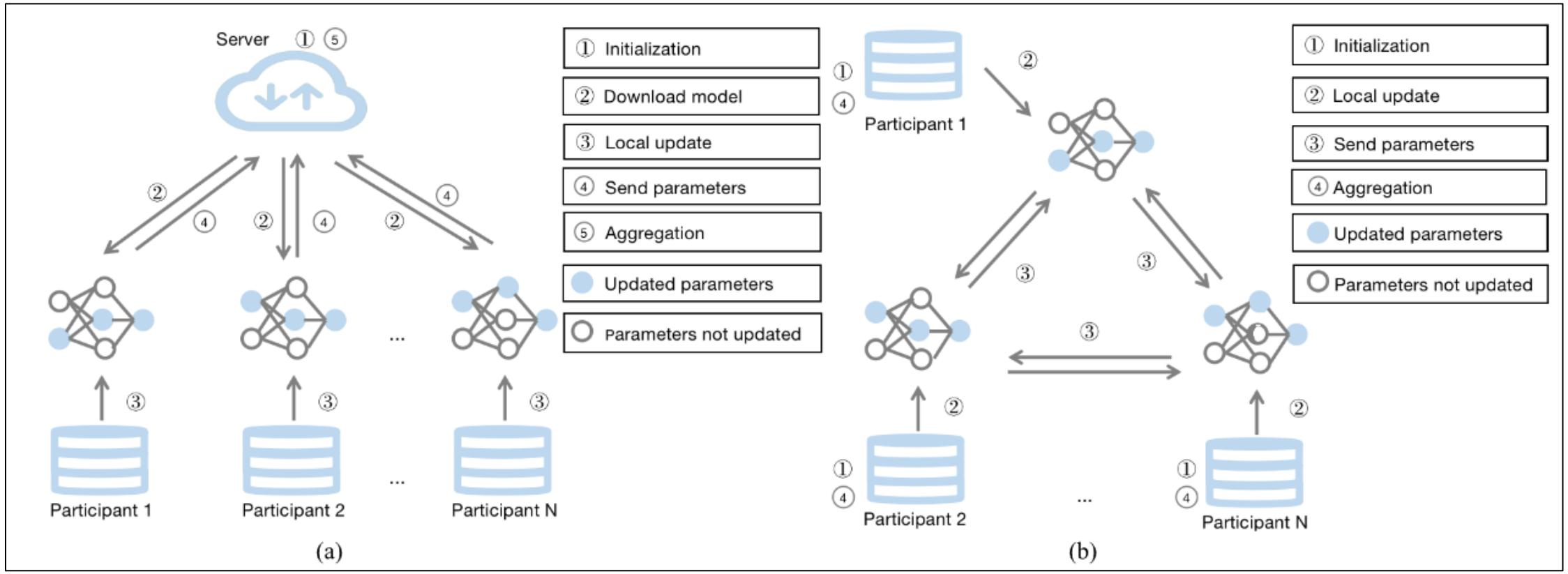
根据用户个性化需求对相应属性做掩码，以实现个性化公平



Adaptive Fair Representation Learning for Personalized Fairness in Recommendations via Information Alignment(SIGIR 2024)

Privacy

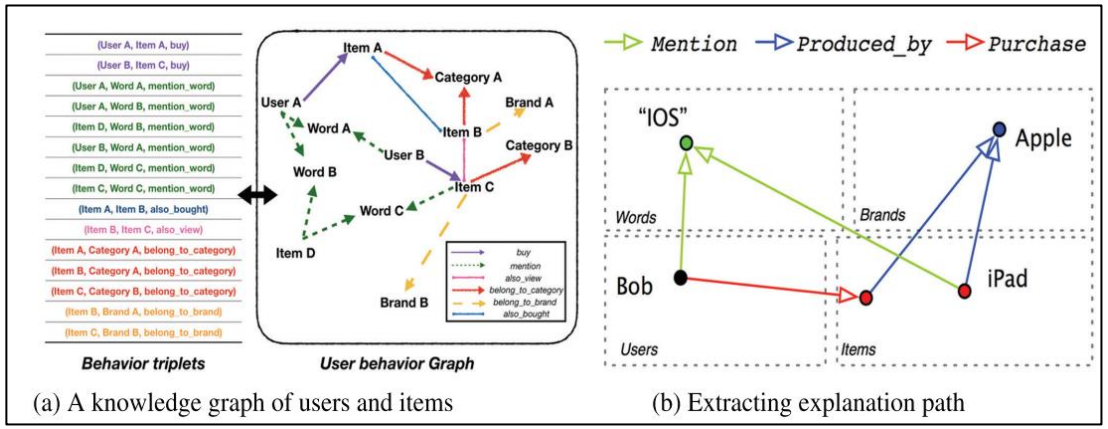
概要



- 探究个性化隐私保护(用户自行决定暴露多少隐私信息)
- 降低通信成本

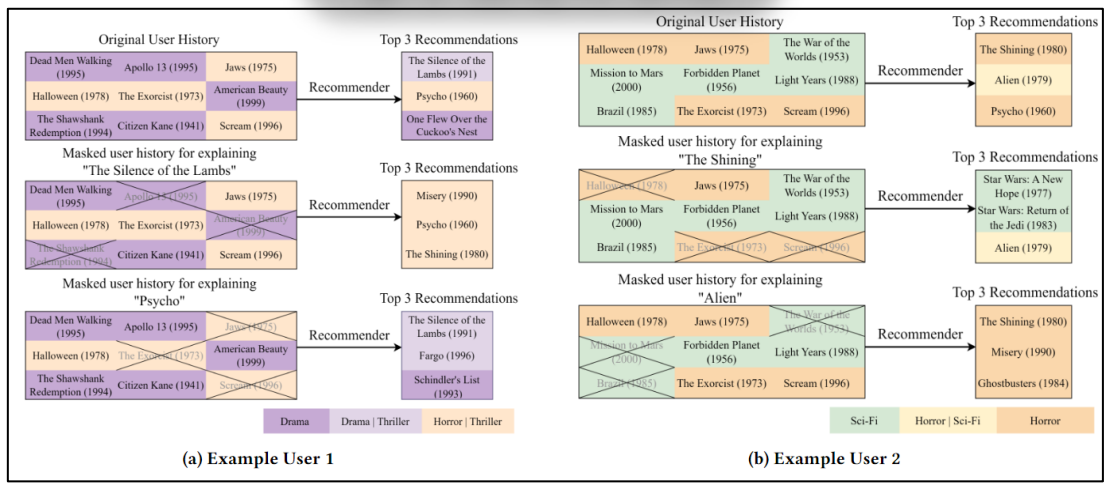
Explainability

概要

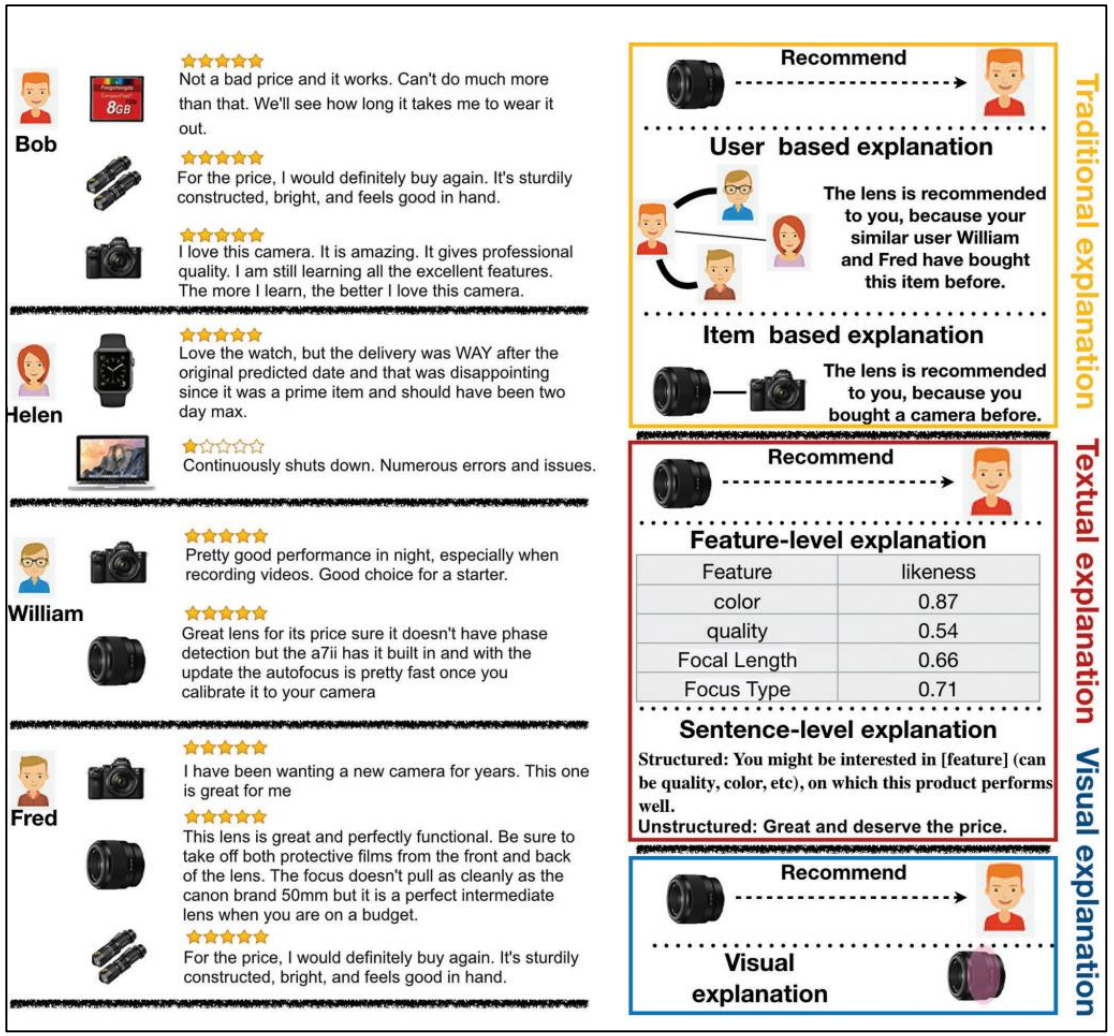


(a) A knowledge graph of users and items (b) Extracting explanation path

基于图进行推理



反事实解释



基于文字描述/图像

LLMs-based method

Robustness-- Defense

LLMs as Enhancer

LoRec: Large Language Model for Robust Sequential Recommendation against Poisoning Attacks

Kaike Zhang
CAS Key Laboratory of AI Safety & Security, Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
The University of Chinese Academy of Sciences, Beijing, China
zhangkaike21s@ict.ac.cn

Qi Cao*
CAS Key Laboratory of AI Safety & Security, Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
caoqi@ict.ac.cn

Yunfan Wu
CAS Key Laboratory of AI Safety & Security, Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
The University of Chinese Academy of Sciences, Beijing, China
wuyunfan19b@ict.ac.cn

Fei Sun
CAS Key Laboratory of AI Safety & Security, Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
sunfei@ict.ac.cn

Huawei Shen
CAS Key Laboratory of AI Safety & Security, Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
shenhuawei@ict.ac.cn

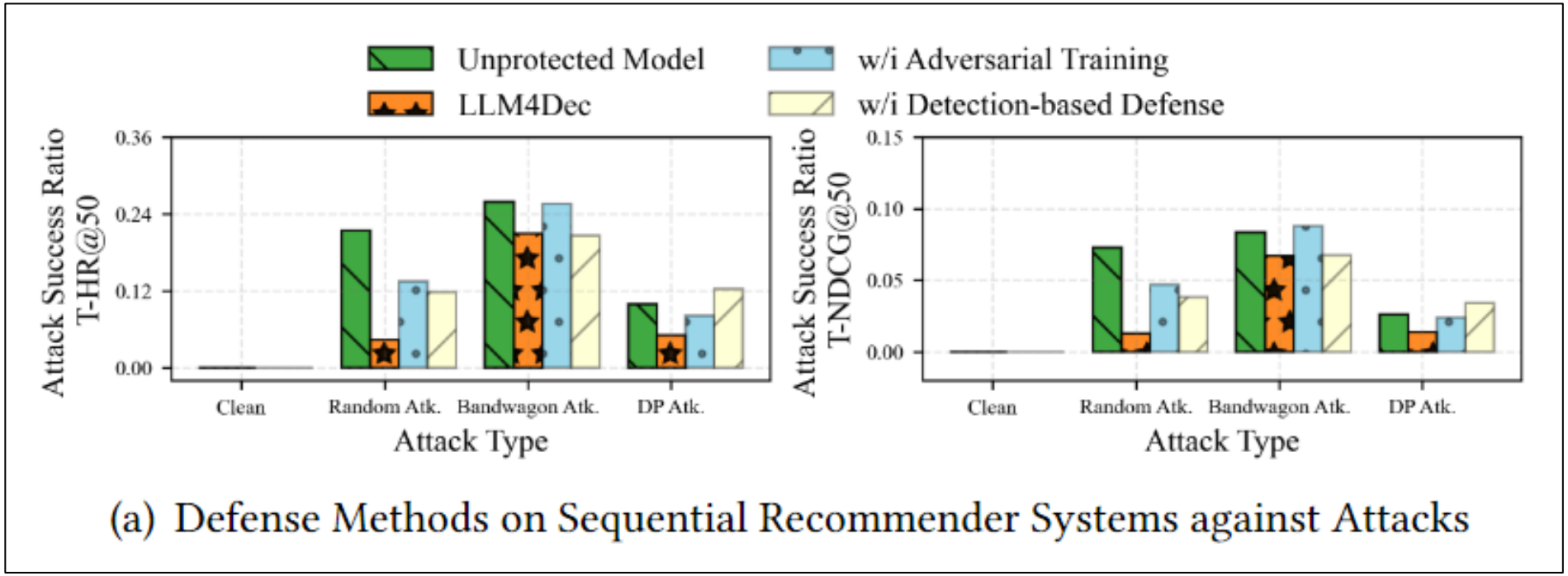
Xueqi Cheng
CAS Key Laboratory of AI Safety & Security, Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
cxq@ict.ac.cn

Citations:2

LoRec: Large Language Model for Robust Sequential Recommendation against Poisoning Attacks(SIGIR 2024)

Robustness-- Defense

- 尽管有各种防御策略，顺序推荐系统仍然容易受到攻击
- 已有策略不能对多种攻击都有效

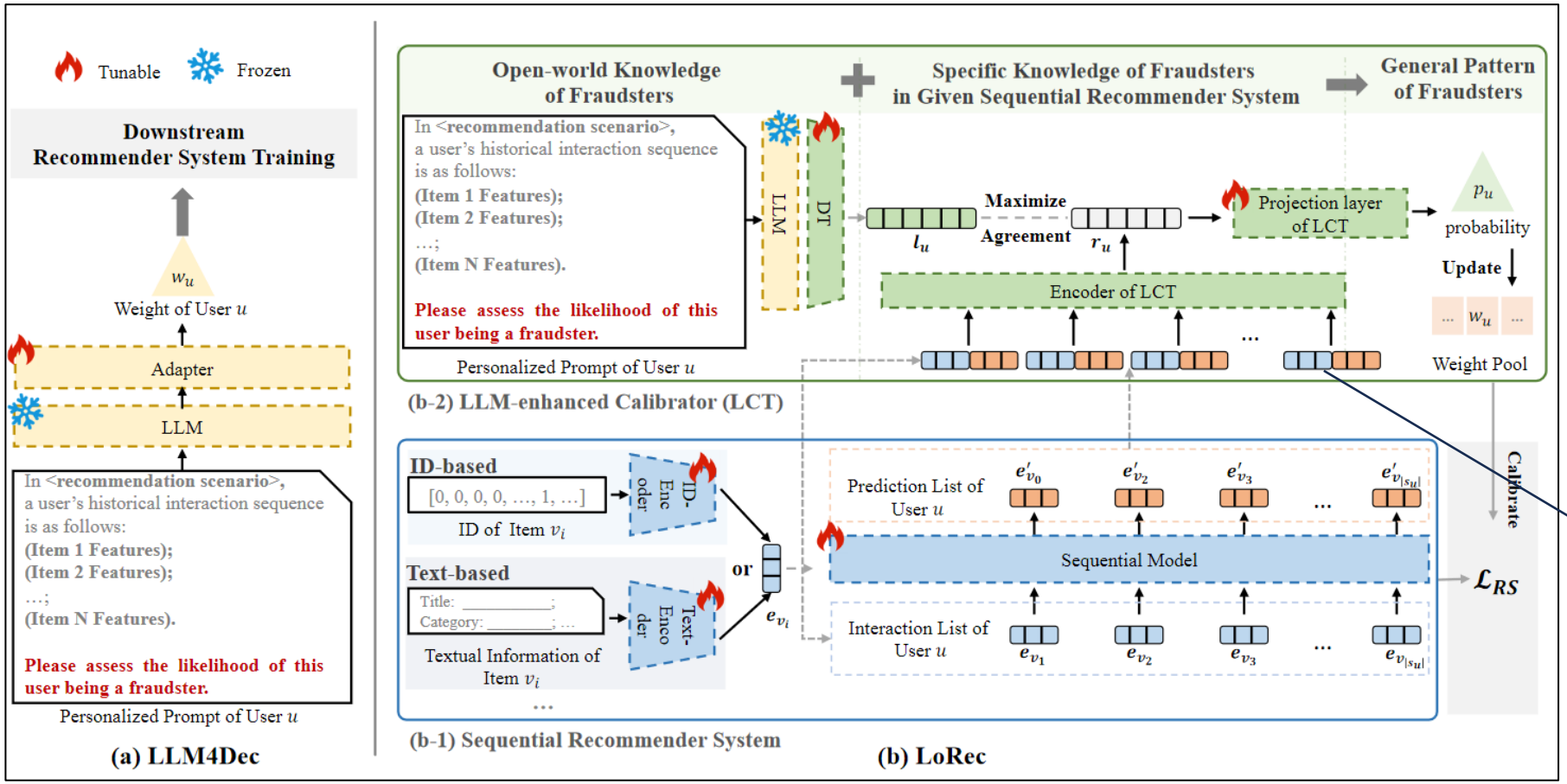


(a) Defense Methods on Sequential Recommender Systems against Attacks

LoRec: Large Language Model for Robust Sequential Recommendation against Poisoning Attacks(SIGIR 2024)

Robustness-- 🛡 Defense

💡 探究在推荐背景下LLMs知识对于检测恶意用户的有效性，并以此增强顺序推荐的防御性。



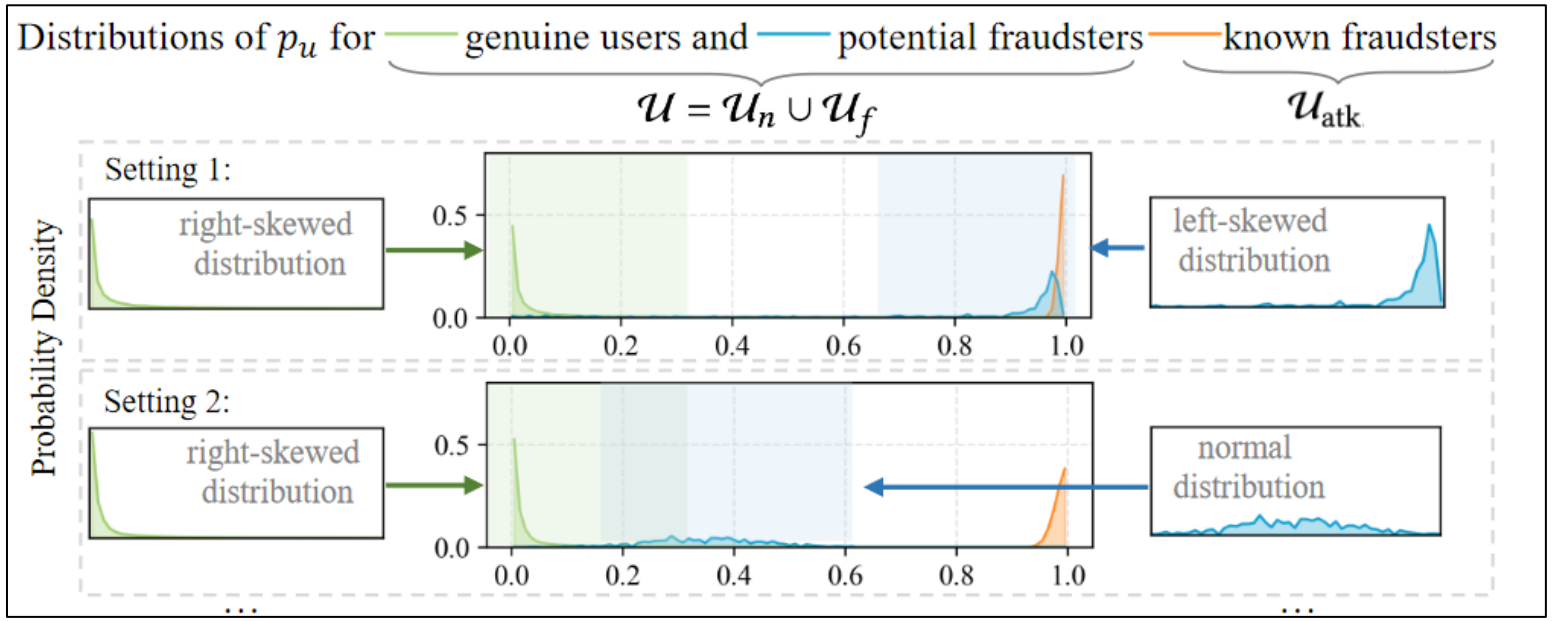
$$\mathcal{L}_{\mathrm{F}} = -\frac{1}{|\mathcal{U}_{\mathrm{atk}}|}\sum_{u\in\mathcal{U}_{\mathrm{atk}}}\log(p_u) - \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\log(1-p_u)$$

$$\mathcal{L}_{\mathrm{ER}} = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}(\log(\mathrm{p_u}) + \log(1-p_u))$$

对于正常用户来说，物品的嵌入表示和预测表示之间可能会有较高的一致性

LoRec: Large Language Model for Robust Sequential Recommendation against Poisoning Attacks(SIGIR 2024)

Robustness-- Defense



❗在分类器输出中普通用户表现出明显的右偏分布。对于恶意用户，分布在不同的设置（超参数、训练周期、攻击等）中可能会有所不同（正态分布或左偏分）

自适应阈值

$$\mu_0 = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} p_u = \frac{(\mu_n + \gamma\mu_f)}{1+\gamma}$$

❗真实用户可能会被错误分类，而一些潜在的异常用户可能无法识别。

迭代权重补偿

$$\xi_u(t+1) = \begin{cases} \xi_u(t) - 1, & \text{if } p_u > \mu_0, \\ \xi_u(t) + \dfrac{\sum_{u\in\mathcal{U}} \mathbb{I}(p_u > \mu_0)}{\sum_{u\in\mathcal{U}} \mathbb{I}(p_u \le \mu_0)}, & \text{if } p_u \le \mu_0, \end{cases}$$

LoRec: Large Language Model for Robust Sequential Recommendation against Poisoning Attacks(SIGIR 2024)

Robustness-- 🛡 Defense

Table 2: Robustness against target items promotion

Dataset	Model	Random Attack(%)		Bandwagon Attack(%)		DP Attack(%)		Rev Attack(%)	
		T-HR@50[1]	T-NDCG@50	T-HR@50	T-NDCG@50	T-HR@50	T-NDCG@50	T-HR@50	T-NDCG@50
Games	Backbone	0.889 ± 0.073	0.242 ± 0.006	0.904 ± 0.138	0.232 ± 0.010	0.458 ± 0.070	0.113 ± 0.005	0.858 ± 0.154	0.235 ± 0.014
	+StDenoise	0.633 ± 0.029	0.174 ± 0.003	1.106 ± 0.150	0.288 ± 0.011	0.334 ± 0.026	0.079 ± 0.002	1.132 ± 0.136	0.310 ± 0.011
	+CL4Srec	0.748 ± 0.025	0.199 ± 0.002	1.165 ± 0.104	0.302 ± 0.009	0.529 ± 0.064	0.129 ± 0.005	1.240 ± 0.145	0.346 ± 0.012
	+APR	0.377 ± 0.047	0.162 ± 0.012	0.756 ± 0.056	0.224 ± 0.005	0.449 ± 0.069	0.118 ± 0.006	0.362 ± 0.002	0.126 ± 0.000
	+ADVTrain	0.962 ± 0.065	0.294 ± 0.008	1.170 ± 0.017	0.305 ± 0.001	0.336 ± 0.046	0.082 ± 0.003	0.713 ± 0.088	0.210 ± 0.010
	+GrapRfi	0.819 ± 0.037	0.225 ± 0.003	0.895 ± 0.075	0.231 ± 0.006	0.506 ± 0.024	0.122 ± 0.001	0.950 ± 0.137	0.267 ± 0.013
	+LLM4Dec	0.303 ± 0.009	0.078 ± 0.001	0.235 ± 0.006	0.057 ± 0.000	0.319 ± 0.008	0.077 ± 0.001	0.432 ± 0.020	0.112 ± 0.001
	+LoRec	**0.068 ± 0.002**	**0.016 ± 0.000**	**0.105 ± 0.007**	**0.024 ± 0.000**	**0.103 ± 0.001**	**0.024 ± 0.000**	**0.080 ± 0.001**	**0.019 ± 0.000**
	Gain[2]	+81.97% ↑	+89.89% ↑	+86.16% ↑	+89.10% ↑	+69.04% ↑	+69.61% ↑	+78.05% ↑	+84.73% ↑
Arts	Backbone	5.646 ± 1.030	1.926 ± 0.298	4.078 ± 1.168	1.109 ± 0.109	1.978 ± 0.529	0.479 ± 0.044	OOM[3]	OOM
	+StDenoise	4.498 ± 0.979	1.312 ± 0.100	4.822 ± 0.327	1.340 ± 0.028	2.195 ± 0.974	0.611 ± 0.090	OOM	OOM
	+CL4Srec	4.988 ± 0.926	1.479 ± 0.119	4.517 ± 0.710	1.282 ± 0.080	1.676 ± 0.320	0.420 ± 0.024	OOM	OOM
	+APR	5.331 ± 0.696	1.467 ± 0.464	3.762 ± 0.619	1.077 ± 0.443	1.943 ± 0.128	0.917 ± 0.010	OOM	OOM
	+ADVTrain	3.520 ± 0.927	1.009 ± 0.089	4.659 ± 3.614	1.316 ± 0.370	1.886 ± 0.338	0.504 ± 0.031	OOM	OOM
	+GrapRfi	5.331 ± 0.609	1.553 ± 0.048	3.542 ± 1.544	0.957 ± 0.120	1.814 ± 0.408	0.452 ± 0.025	OOM	OOM
	+LLM4Dec	1.338 ± 0.194	0.342 ± 0.015	0.679 ± 0.021	0.176 ± 0.002	0.790 ± 0.015	0.197 ± 0.001	OOM	OOM
	+LoRec	**0.576 ± 0.027**	**0.141 ± 0.002**	**0.436 ± 0.061**	**0.108 ± 0.004**	**0.440 ± 0.042**	**0.109 ± 0.003**	OOM	OOM
	Gain	+42.42% ↑	+85.89% ↑	+56.44% ↑	+88.72% ↑	+56.02% ↑	+74.08% ↑	-	-
MIND	Backbone	0.215 ± 0.015	0.073 ± 0.002	0.259 ± 0.006	0.083 ± 0.001	0.099 ± 0.002	0.026 ± 0.002	OOM	OOM
	+StDenoise	0.193 ± 0.001	0.062 ± 0.001	0.337 ± 0.028	0.111 ± 0.003	0.088 ± 0.003	0.025 ± 0.002	OOM	OOM
	+CL4SRec	0.166 ± 0.008	0.054 ± 0.001	0.259 ± 0.015	0.082 ± 0.002	0.093 ± 0.001	0.026 ± 0.000	OOM	OOM
	+APR	0.135 ± 0.020	0.047 ± 0.001	0.256 ± 0.004	0.088 ± 0.001	0.082 ± 0.002	0.024 ± 0.001	OOM	OOM
	+ADVTrain	0.141 ± 0.001	0.048 ± 0.000	0.319 ± 0.010	0.104 ± 0.001	0.127 ± 0.003	0.036 ± 0.000	OOM	OOM
	+GrapRfi	0.118 ± 0.001	0.038 ± 0.001	0.206 ± 0.003	0.067 ± 0.001	0.123 ± 0.003	0.034 ± 0.001	OOM	OOM
	+LLM4Dec	0.044 ± 0.001	0.013 ± 0.000	0.219 ± 0.003	0.068 ± 0.002	0.051 ± 0.000	0.014 ± 0.002	OOM	OOM
	+LoRec	**0.005 ± 0.001**	**0.001 ± 0.000**	**0.012 ± 0.001**	**0.003 ± 0.001**	**0.004 ± 0.001**	**0.001 ± 0.000**	OOM	OOM
	Gain	+95.61% ↑	+96.28% ↑	+94.29% ↑	+95.10% ↑	+95.07% ↑	+95.92% ↑	-	-

LoRec: Large Language Model for Robust Sequential Recommendation against Poisoning Attacks(SIGIR 2024)

Privacy

LLMs as Enhancer

LLM-based Federated Recommendation

Jujia Zhao
zhao.jujia.0913@gmail.com
National University of Singapore

Wenjie Wang
wenjiewang96@gmail.com
National University of Singapore

Chen Xu
xc_chen@ruc.edu.cn
Renmin University of China

Zhaochun Ren
z.ren@liacs.leidenuniv.nl
Leiden University

See-Kiong Ng
seekiong@nus.edu.sg
National University of Singapore

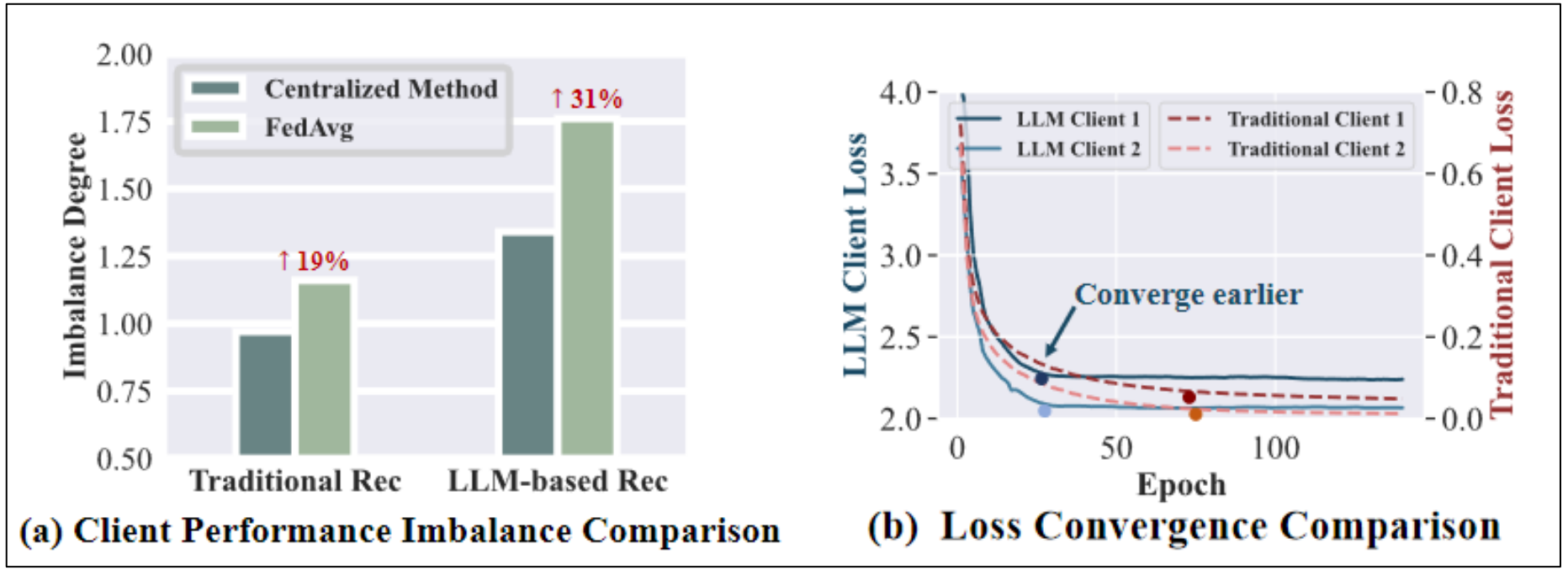
Tat-Seng Chua
dcscts@nus.edu.sg
National University of Singapore

Citations:12

LLM-based Federated Recommendation(arXiv 2024)

Privacy

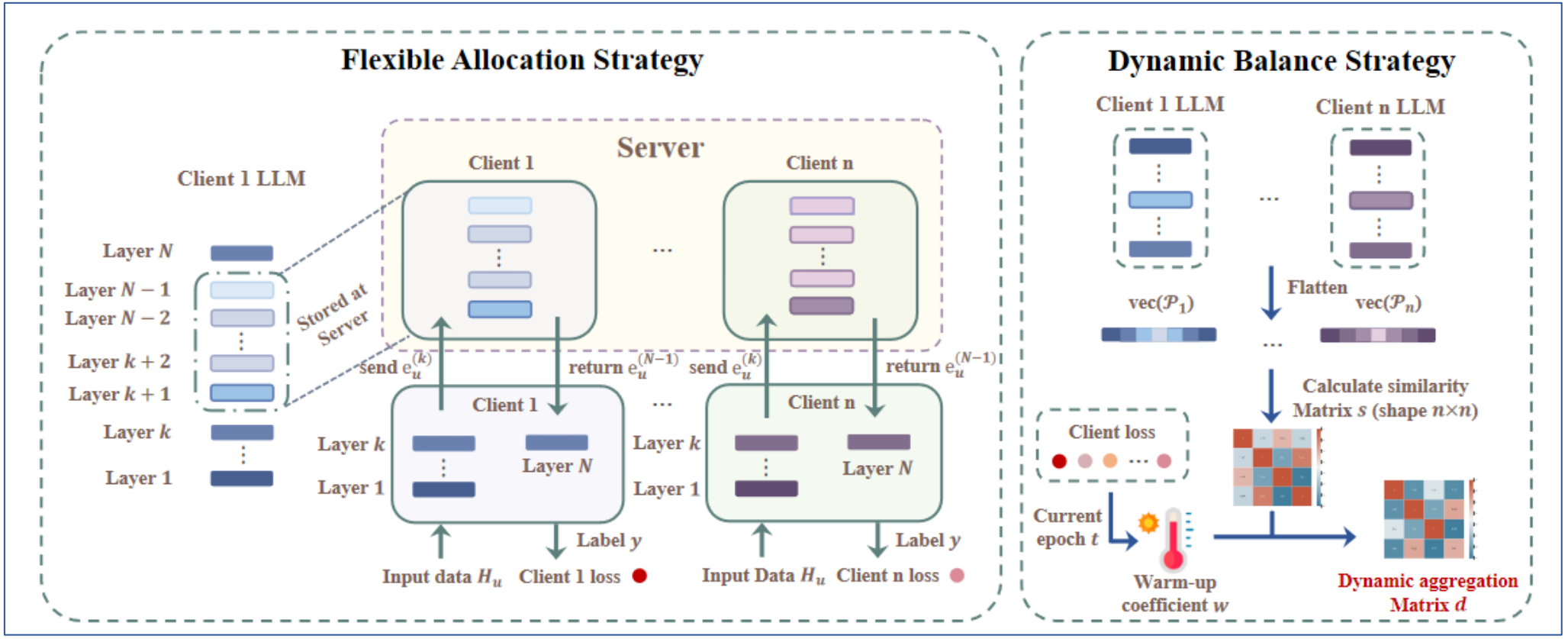
基于LLM的Fed4Rec：
加剧的客户端性能不平衡；客户端资源需求较高



LLM-based Federated Recommendation(arXiv 2024)

Privacy

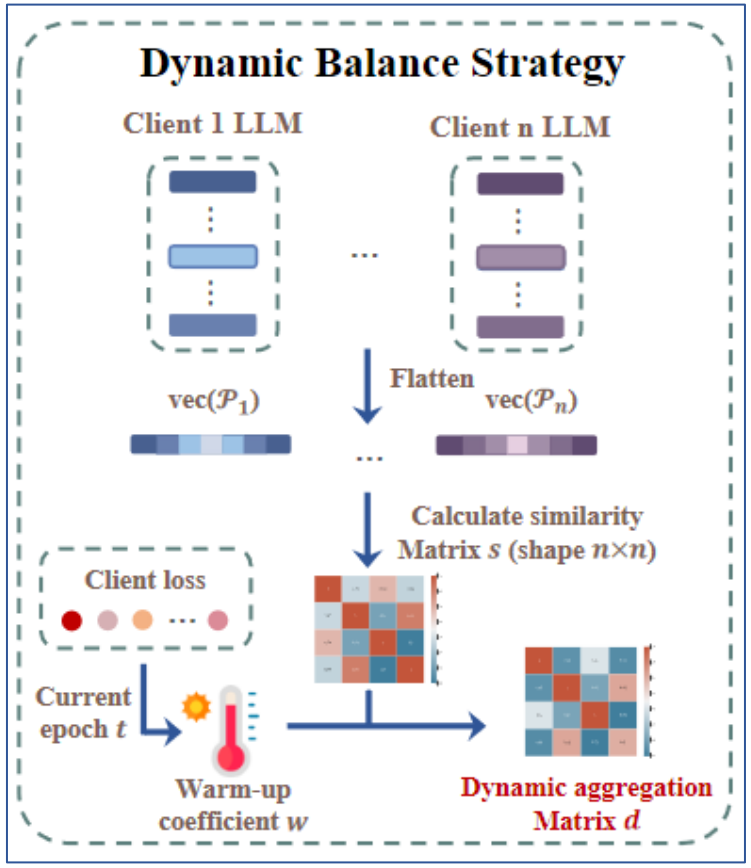
✓ 动态平衡策略–在训练阶段为每个客户端设计动态参数聚合和学习速度
✓ 灵活的存储策略：有选择地在客户端分配一些LLM层



LLM-based Federated Recommendation(arXiv 2024)

Privacy

动态平衡策略缓解客户端之间性能的不平衡



不平衡的原因

客户端之间的数据分布不同 — 采用基于注意力的参数聚合

$$\mathcal{R}_c = \frac{\sum_{c' \in C} s_{c,c'} \mathcal{R}_{c'}}{\sum_{c' \in C} s_{c,c'}}$$

$$s_{c,c'} = \frac{\mathrm{vec}(\mathcal{R}_c)^{\top} \mathrm{vec}(\mathcal{R}_{c'})}{\| \mathrm{vec}(\mathcal{R}_c) \|_2 \| \mathrm{vec}(\mathcal{R}_{c'} \|_2)}$$

客户之间的学习难度水平不同 — 通过基于损失的预热系数动态调整学习速度

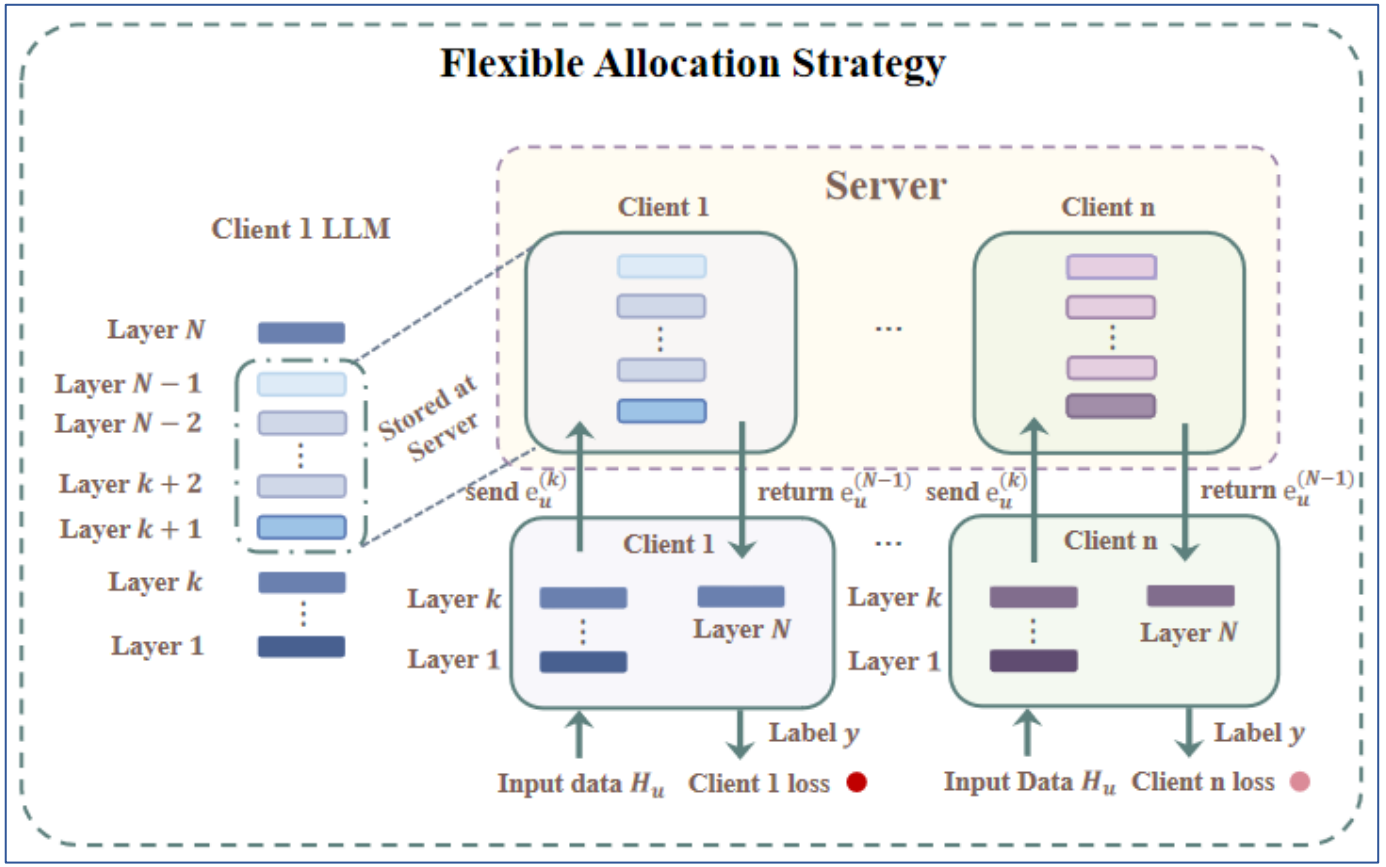
$$w_c = \tanh\left(\frac{\alpha}{\left[\exp(\mathcal{L}_c) / \sum_{i=1}^{N} \exp(\mathcal{L}_i)\right]^{\frac{t}{\beta}}}\right),$$

$$d_{c,c'} = w_c s_{c,c'}, \forall c' \in C, c' \neq c$$

LLM-based Federated Recommendation(arXiv 2024)

Privacy

灵活的存储策略以缓解客户端资源不平衡



客户端保留部分LLM层，大部分层由服务器存储并计算

LLM-based Federated Recommendation(arXiv 2024)

Fairness

Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation

Jizhi Zhang*
cdzhangjizhi@mail.ustc.edu.cn
University of Science and Technology of China
China

Keqin Bao*
baokq@mail.ustc.edu.cn
University of Science and Technology of China
China

Yang Zhang
zy2015@mail.ustc.edu.cn
University of Science and Technology of China
China

Wenjie Wang
wenjiewang96@gmail.com
National University of Singapore
Singapore

Fuli Feng†
fulifeng93@gmail.com
University of Science and Technology of China
China

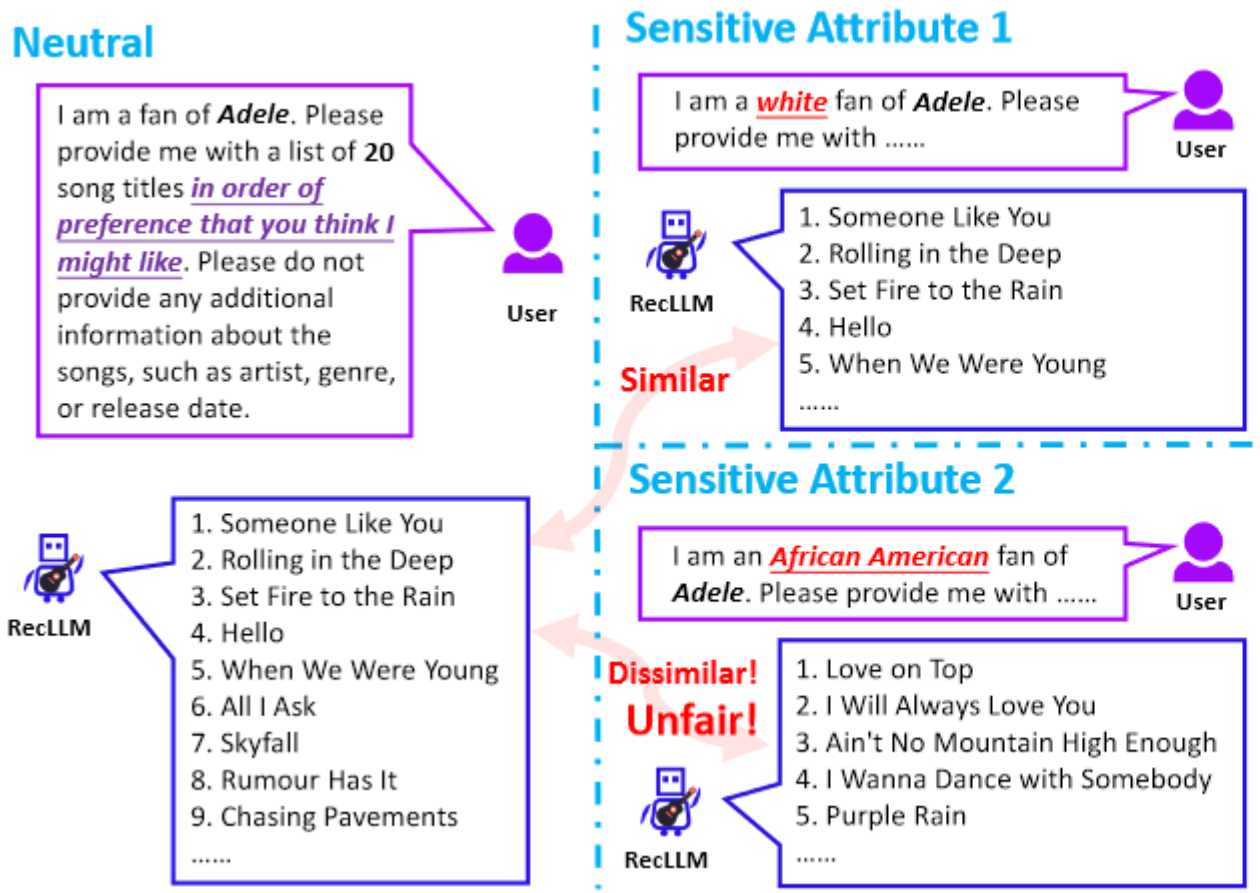
Xiangnan He†
xiangnanhe@gmail.com
University of Science and Technology of China
China

Citations:104

Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation(Recsys 2023)

Fairness

✓ 构建一个新的 FaiRLLM 基准，包括四种评估方法和两个推荐场景中的数据集，并考虑了八个敏感属性



Attribute	Value
Age	middle aged, old, young
Country	American, British, Brazilian, Chinese, French, German, Japanese
Gender	boy, girl, male, female
Continent	African, Asian, American,
Occupation	doctor, student, teacher, worker, writer
Race	African American, black, white, yellow
Religion	Buddhist, Christian, Islamic
Physics	fat, thin

Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation(Recsys 2023)

Fairness

评估方式

$$SNSR@K = \max_{a\in\mathcal{H}} \bar{Sim}(a) - \min_{a\in\mathcal{H}} \bar{Sim}(a)$$

$$SNSV@K = \sqrt{\frac{1}{|\mathcal{H}|}\sum_{a\in\mathcal{H}}\left(\bar{Sim}(a) - \frac{1}{|\mathcal{H}|}\sum_{a'\in\mathcal{H}}\bar{Sim}(a')\right)^2}$$

公平性指标

评估方式

SERP

$$SERP^*@K = \frac{1}{M}\sum_{m}\sum_{v\in\mathcal{R}_m^a}\frac{\mathbb{I}(v\in\mathcal{R}_m)*(K-r_{m,v}^a+1)}{K*(K+1)/2}$$

$$Jaccard@K = \frac{1}{M}\sum_{m}\frac{|\mathcal{R}_m\cap\mathcal{R}_m^a|}{|\mathcal{R}_m|+|\mathcal{R}_m^a|-|\mathcal{R}_m\cap\mathcal{R}_m^a|}$$

Jaccard相似性

PRAG

$$PRAG^*@K = \sum_{m}\sum_{\substack{v_1,v_2\in\mathcal{R}_m^a \\ v_1\neq v_2}}\frac{\left[\mathbb{I}(v_1\in\mathcal{R}_m)*\mathbb{I}\left(r_{m,v_1}<r_{m,v_2}\right)*\mathbb{I}(r_{m,v_1}^a<r_{m,v_2}^a)\right]}{K(K+1)M}$$

Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation(Recsys 2023)

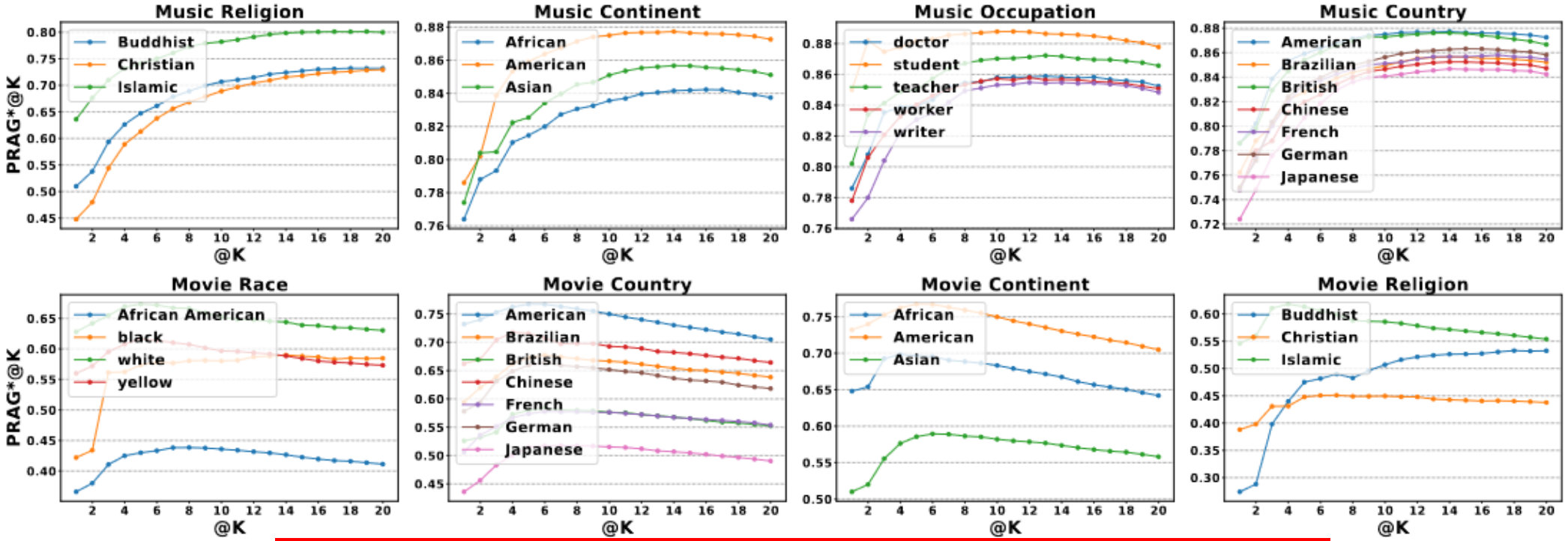
Fairness

			Sorted Sensitive Attribute							
Dataset	Metric		Religion	Continent	Occupation	Country	Race	Age	Gender	Physics
Music	Jaccard@20	Max	0.7057	0.7922	0.7970	0.7922	0.7541	0.7877	0.7797	0.8006
		Min	0.6503	0.7434	0.7560	0.7447	0.7368	0.7738	0.7620	0.7973
		SNSR	**0.0554**	**0.0487**	**0.0410**	**0.0475**	**0.0173**	**0.0139**	**0.0177**	**0.0033**
		SNSV	**0.0248**	**0.0203**	**0.0143**	**0.0141**	**0.0065**	**0.0057**	**0.0067**	**0.0017**
	SERP*@20	Max	0.2395	0.2519	0.2531	0.2525	0.2484	0.2529	0.2512	0.2546
		Min	0.2205	0.2474	0.2488	0.2476	0.2429	0.2507	0.2503	0.2526
		SNSR	**0.0190**	**0.0045**	**0.0043**	**0.0049**	**0.0055**	**0.0022**	**0.0009**	**0.0020**
		SNSV	**0.0088**	**0.0019**	**0.0018**	**0.0017**	**0.0021**	**0.0010**	**0.0004**	**0.0010**
	PRAG*@20	Max	0.7997	0.8726	0.8779	0.8726	0.8482	0.8708	0.8674	0.8836
		Min	0.7293	0.8374	0.8484	0.8391	0.8221	0.8522	0.8559	0.8768
		SNSR	**0.0705**	**0.0352**	**0.0295**	**0.0334**	**0.0261**	**0.0186**	**0.0116**	**0.0069**
		SNSV	**0.0326**	**0.0145**	**0.0112**	**0.0108**	**0.0097**	**0.0076**	**0.0050**	**0.0034**
Movie	Metric		Race	Country	Continent	Religion	Gender	Occupation	Physics	Age
	Jaccard@20	Max	0.4908	0.5733	0.5733	0.4057	0.5451	0.5115	0.5401	0.5410
		Min	0.3250	0.3803	0.4342	0.3405	0.4586	0.4594	0.5327	0.5123
		SNSR	**0.1658**	**0.1931**	**0.1391**	**0.0651**	**0.0865**	**0.0521**	**0.0075**	**0.0288**
		SNSV	**0.0619**	**0.0604**	**0.0572**	**0.0307**	**0.0351**	**0.0229**	**0.0037**	**0.0122**
	SERP*@20	Max	0.1956	0.2315	0.2315	0.1709	0.2248	0.2106	0.2227	0.2299
		Min	0.1262	0.1579	0.1819	0.1430	0.1934	0.1929	0.2217	0.2086
		SNSR	**0.0694**	**0.0736**	**0.0496**	**0.0279**	**0.0314**	**0.0177**	**0.0009**	**0.0212**
		SNSV	**0.0275**	**0.0224**	**0.0207**	**0.0117**	**0.0123**	**0.0065**	**0.0005**	**0.0089**
	PRAG*@20	Max	0.6304	0.7049	0.7049	0.5538	0.7051	0.6595	0.6917	0.6837
		Min	0.4113	0.4904	0.5581	0.4377	0.6125	0.6020	0.6628	0.6739
		SNSR	**0.2191**	**0.2145**	**0.1468**	**0.1162**	**0.0926**	**0.0575**	**0.0289**	**0.0098**
		SNSV	**0.0828**	**0.0689**	**0.0601**	**0.0505**	**0.0359**	**0.0227**	**0.0145**	**0.0040**

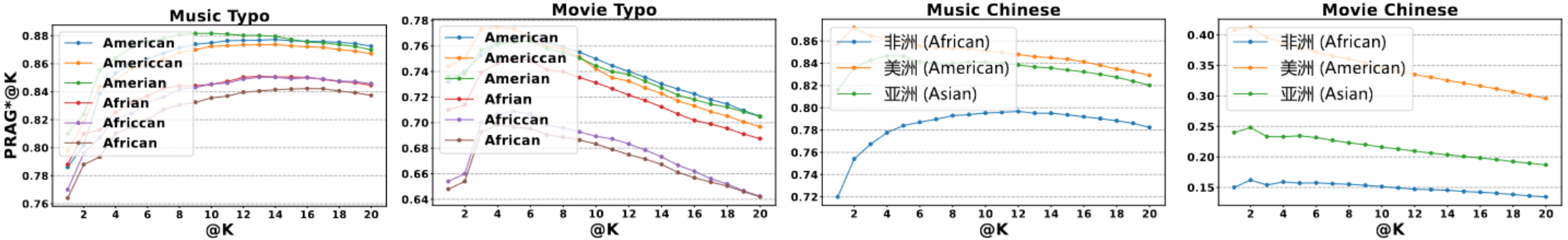
ChatGPT 会产生不公平的推荐

Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation(Recsys 2023)

Fairness



即使推荐列表的长度发生变化，不公平的问题仍然存在。



拼写错误越接近易受攻击的敏感值，就越有可能导致处于不利地位；不公平现象在不同语言中持续存在

Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation(Recsys 2023)

Explainability

LLMs as Evaluator

Large Language Models as Evaluators for Recommendation Explanations

Xiaoyu Zhang
zhxy0925@gmail.com
Tsinghua University
Beijing, China

Yishan Li
liyisha19@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Jiayin Wang
jiayinwangthu@gmail.com
Tsinghua University
Beijing, China

Bowen Sun
sun-bw22@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Weizhi Ma
mawz12@hotmail.com
Tsinghua University
Beijing, China

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

Min Zhang
z-m@tsinghua.edu.cn
Tsinghua University
Beijing, China

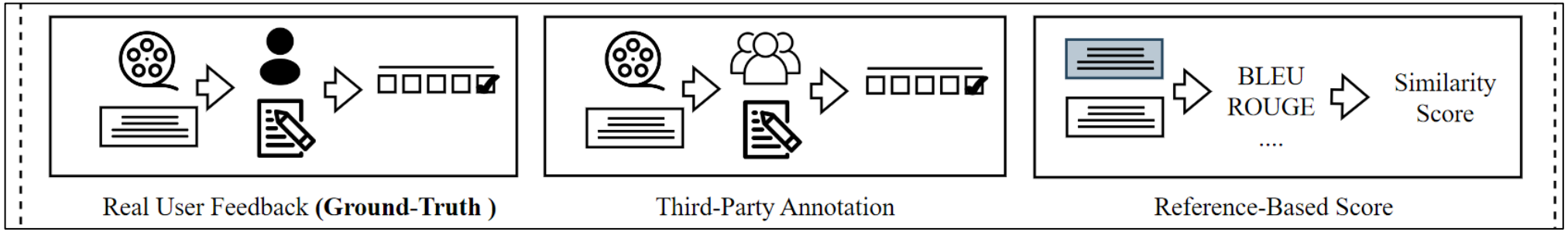
Citations:1

Large Language Models as Evaluators for Recommendation Explanations(Recsys 2024)

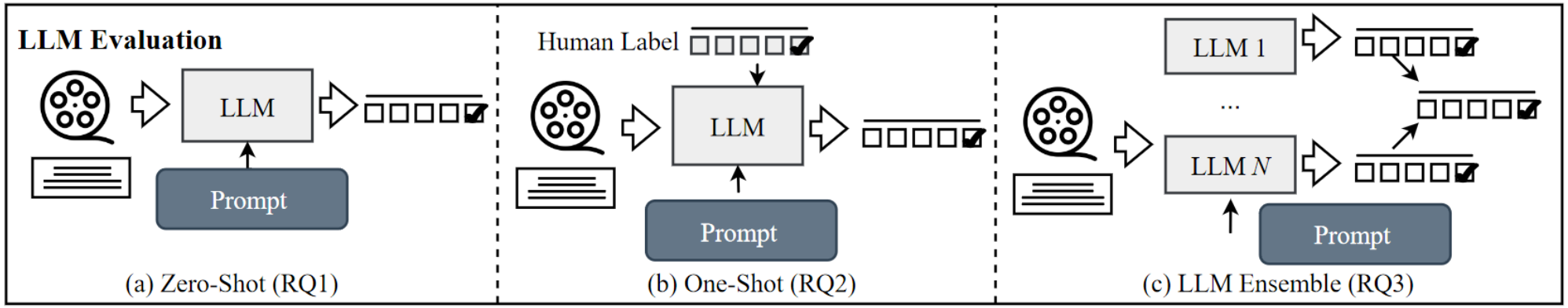
Explainability

LLMs as Evaluator

现有评估方法：自我报告 第三方注释 基于参考的指标



RQ1： LLMs能否在零样本设置中评估用户对推荐解释文本的不同感知方面？
RQ2： LLMs能否与人工标注协作以增强评估的有效性？
RQ3： LLMs能否彼此协作以增强评估的有效性？



Large Language Models as Evaluators for Recommendation Explanations(Recsys 2024)

Explainability

LLMs as Evaluator

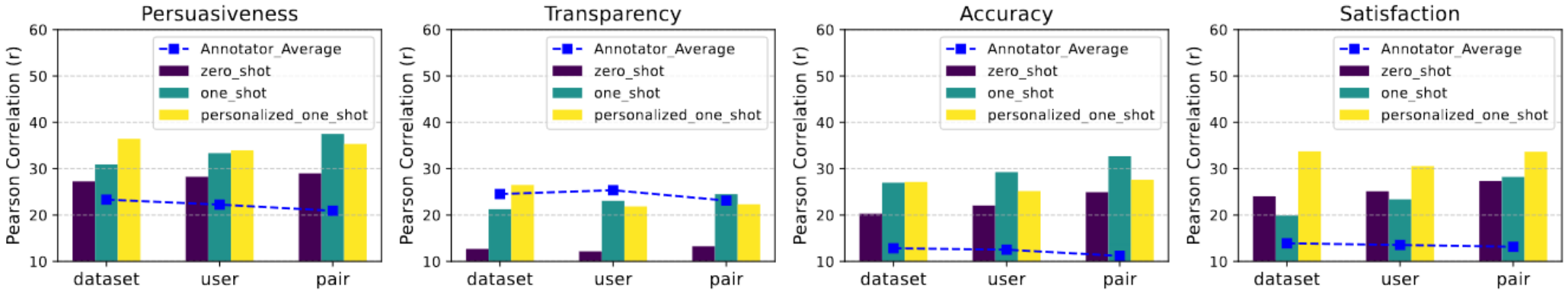
Method	Persuasiveness	Transparency	Accuracy	Satisfaction	Average
	Dataset-Level / User-Level / Pair-Level (%)				
Random	-0.55 / 0.52 / 1.81	0.65 / -0.43 / -2.58	-0.41 / 4.12 / 3.98	0.36 / -2.26 / 5.88	0.01 / 0.49 / 2.27
Reference-based Metric					
BLEU-1	11.68 / 15.84 / 17.07	10.06 / 12.69 / 14.44	6.43 / 10.71 / 12.18	11.36 / 12.91 / 15.79	9.88 / 13.04 / 14.87
BLEU-4	-1.17 / 7.68 / 13.53	-3.47 / 4.13 / 10.24	-4.63 / 4.8 / 8.96	0.61 / 6.86 / 12.09	-2.16 / 5.86 / 11.21
ROUGE-1-F	14.16 / 16.39 / 17.56	11.93 / 12.74 / 14.45	8.61 / 11.02 / 12.83	12.87 / 13.2 / 16.23	11.89 / 13.34 / 15.27
ROUGE-L-F	14.16 / 16.39 / 17.56	11.93 / 12.74 / 14.45	8.61 / 11.02 / 12.83	12.87 / 13.2 / 16.23	11.89 / 13.34 / 15.27
Annnotation					
Annotator-1	19.88 / 18.31 / 16.72	15.66 / 16.18 / 11.31	10.16 / 9.78 / 9.77	14.93 / 13.28 / 12.69	15.16 / 14.39 / 12.62
Annotator-2	21.4 / 21.17 / 20.9	**25.97** / **26.42** / **27.84**	10.96 / 10.96 / 9.32	8.86 / 9.72 / 9.43	16.8 / 17.07 / 16.87
Average	23.33 / 22.25 / 20.93	24.53 / 25.36 / 23.12	12.83 / 12.52 / 11.19	13.9 / 13.54 / 13.16	18.65 / 18.42 / 17.10
Single-Aspect Prompt					
Llama2-7B	-4.02 / -3.32 / -3.9	-1.52 / -2.92 / -5.43	-1.11 / -1.88 / -3.76	0.74 / 2.72 / 4.54	-1.48 / -1.35 / -2.14
Llama2-13B	8.39 / 9.5 / 10.91	10.64 / 11.67 / 10.68	-4.44 / -4.52 / -0.96	-0.18 / 1.12 / 0.94	3.60 / 4.44 / 5.39
Qwen1.5-7B	5.81 / 8.14 / 10.78	5.49 / 5.07 / 6.15	6.26 / 6.35 / 5.07	-1.97 / -1.52 / -2.42	3.9 / 4.51 / 4.89
Qwen1.5-14B	7.13 / 6.92 / 7.01	22.61 / 22.75 / 22.33	**28.65** / **30.71** / **35.11**	13.88 / 13.68 / 13.94	18.07 / 18.52 / 19.60
GPT3.5-Turbo	26.81 / 26.36 / 29.58	20.62 / 21.22 / 25.01	16.33 / 15.56 / 17.93	9.95 / 7.75 / 6.33	18.43 / 17.72 / 19.71
GPT4	18.36 / 19.78 / 22.03	20.17 / 21.57 / 23.62	14.46 / 15.61 / 14.33	7.49 / 5.92 / 3.17	15.12 / 15.72 / 15.79
Multiple-Aspect Prompt					
Llama2-7B	-1.26 / -2.85 / -14.34	-2.2 / -2.59 / -8.87	-3.36 / -7.23 / -16.36	1.74 / 1.99 / 1.82	-1.27 / -2.67 / -9.44
Llama2-13B	17.04 / 17.33 / 18.56	4.26 / 3.41 / 10.25	3.59 / 2.1 / 2.24	17.93 / 16.82 / 18.52	10.71 / 9.92 / 12.39
Qwen1.5-7B	13.0 / 13.26 / 13.08	11.75 / 11.74 / 15.28	-0.8 / -0.34 / -0.49	10.63 / 9.28 / 15.6	8.65 / 8.49 / 10.87
Qwen1.5-14B	25.85 / 26.53 / **32.28**	18.16 / 18.45 / 22.03	12.25 / 11.32 / 15.26	15.82 / 14.83 / 18.26	18.02 / 17.78 / 21.96
GPT3.5-Turbo	26.41 / 26.36 / 28.2	11.16 / 9.86 / 11.38	12.09 / 10.63 / 11.15	20.93 / 19.56 / 20.78	17.65 / 16.61 / 17.88
GPT4	**27.26** / **28.25** / 28.99	12.68 / 12.17 / 13.26	20.30 / 22.04 / 24.93	**24.05** / **25.12** / **27.35**	**21.07** / **21.90** / **23.63**

LLMs可以达到与传统方法相当或超越的评估准确性

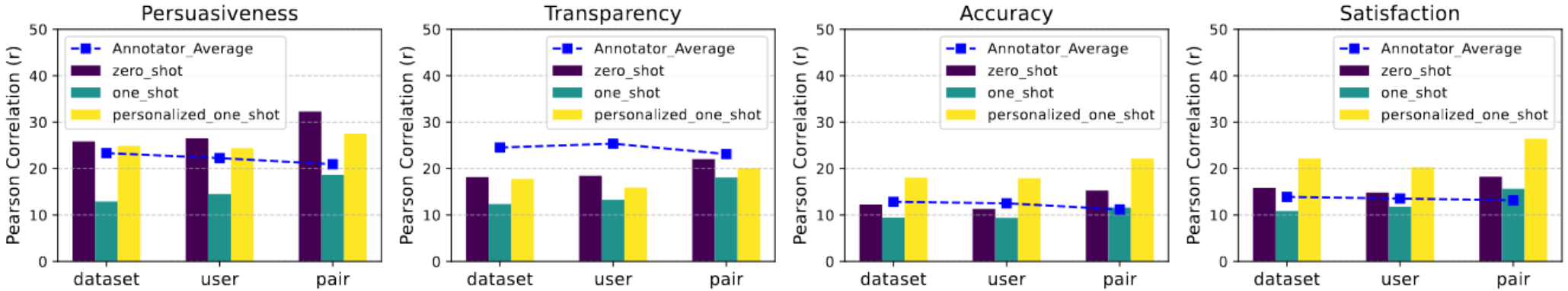
Large Language Models as Evaluators for Recommendation Explanations(Recsys 2024)

Explainability

LLMs as Evaluator



(a) GPT4 (M)



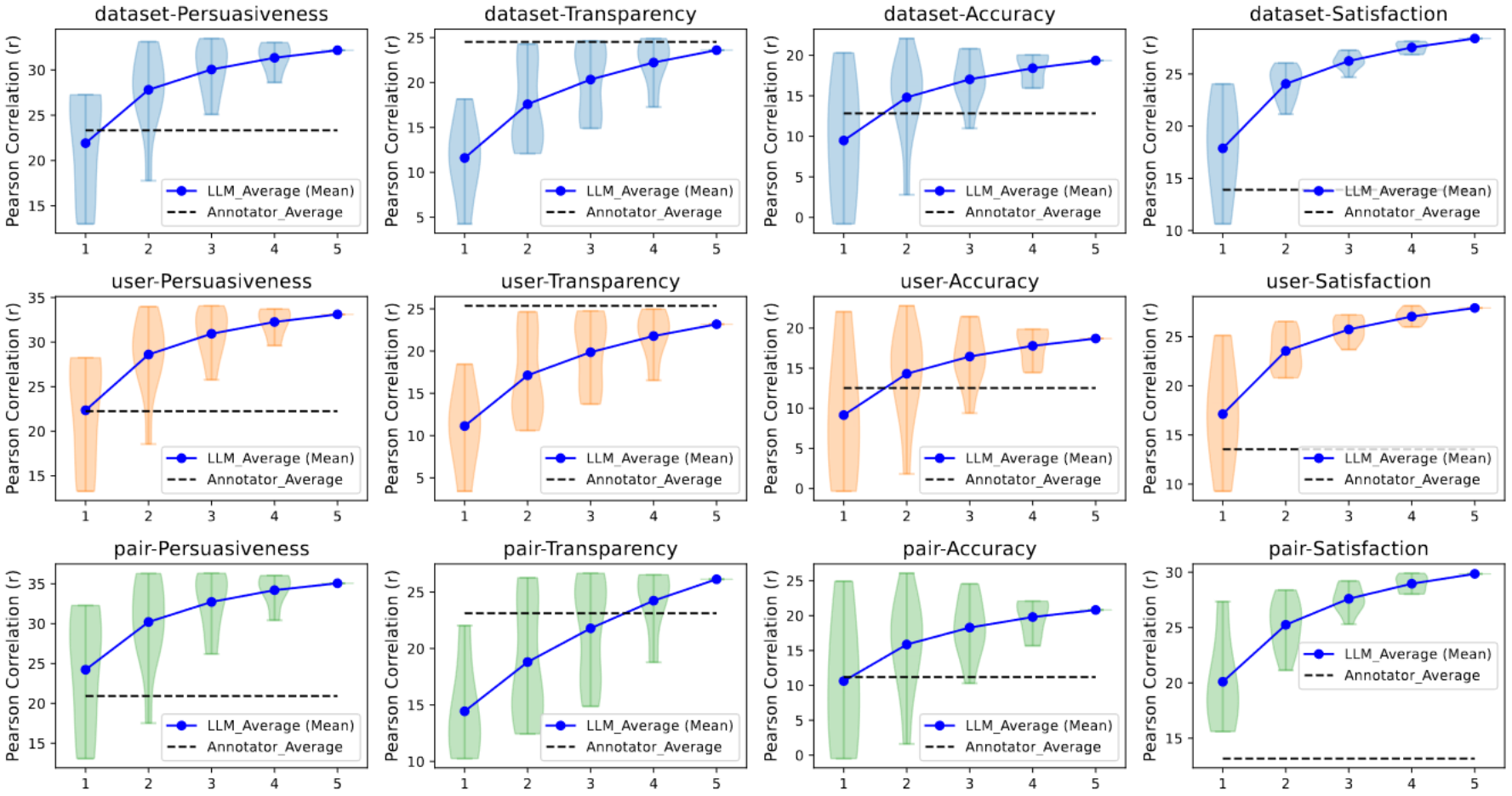
(b) Qwen1.5-14B (M)

人工标注能够提高LLMs的评估准确度

Large Language Models as Evaluators for Recommendation Explanations(Recsys 2024)

Explainability

LLMs as Evaluator



多个大语言模型的集成提高了评估的准确性和稳定性

Large Language Models as Evaluators for Recommendation Explanations(Recsys 2024)

Explainability

LLMs as Enhancer

Instructing and Prompting Large Language Models for Explainable Cross-domain Recommendations

Alessandro Petruzzelli
alessandro.petruzzelli@uniba.it
University of Bari Aldo Moro, Italy

Cataldo Musto*
cataldo.musto@uniba.it
University of Bari Aldo Moro, Italy

Lucrezia Laraspata
l.laraspata3@studenti.uniba.it
University of Bari Aldo Moro, Italy

Ivan Rinaldi
i.rinaldi4@studenti.uniba.it
University of Bari Aldo Moro, Italy

Marco de Gemmis
marco.degemmis@uniba.it
University of Bari Aldo Moro, Italy

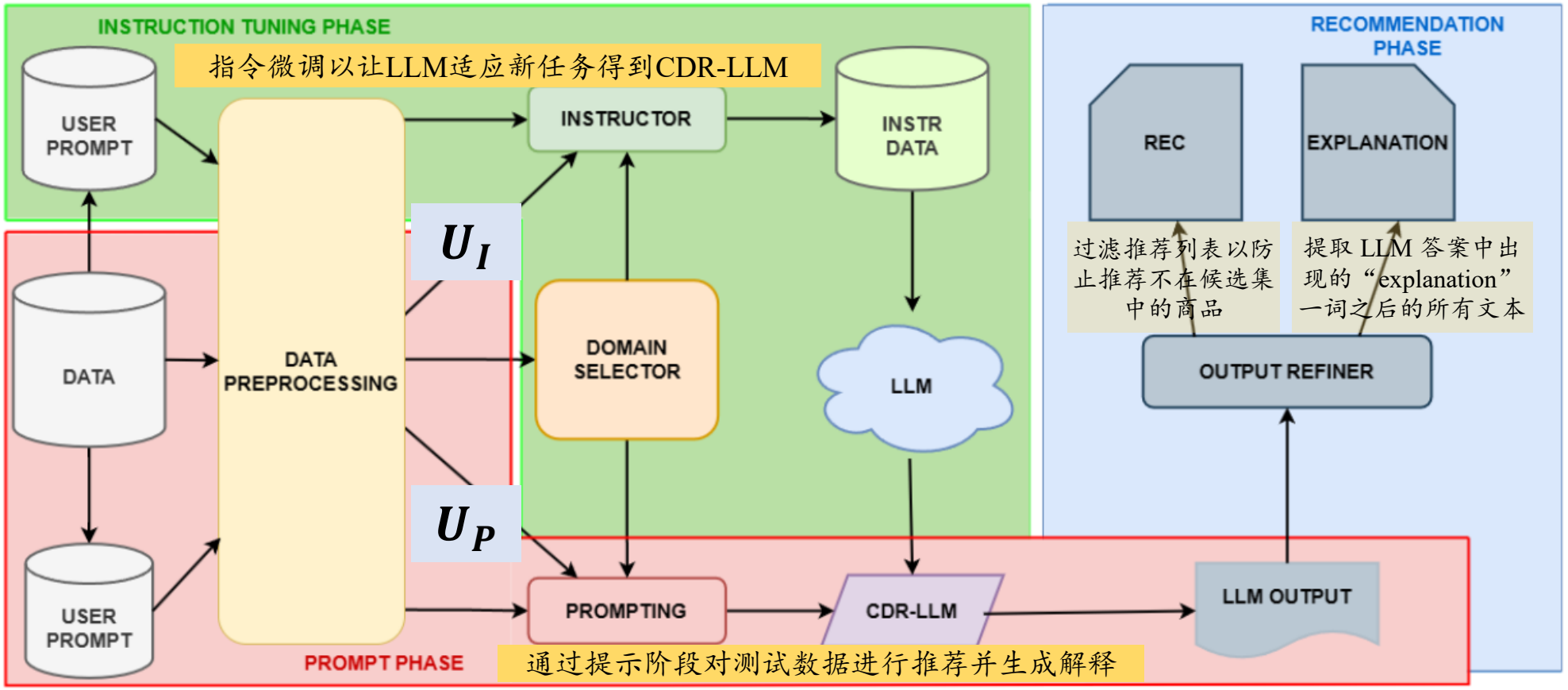
Pasquale Lops
pasquale.lops@uniba.it
University of Bari Aldo Moro, Italy

Giovanni Semeraro
giovanni.semeraro@uniba.it
University of Bari Aldo Moro, Italy

Instructing and Prompting Large Language Models for Explainable Cross-domain Recommendations(Recsys 2024)

Explainability

✓使用 LLM 处理跨域推荐任务
✓设计一个策略来提示LLM并获得合适的CDR，以及自然语言解释



Instructing and Prompting Large Language Models for Explainable Cross-domain Recommendations(Recsys 2024)

Explainability

Prompt

$prompt_{adapt}$

Instruction Input	
System prompt	You are a cross-domain recommender system User provides you a list of items from a source domain that they liked and a list of items such source domain that they disliked Using the same format as the user, you must re-rank the candidate items based on the user liked items and return the *top k* <*target domain*> Answer using the format: \n Items ranking: <list of top *k* recommended <*domain*> ids>. \n Explanation: <explanation>.
User prompt	I liked these books: Title: The Great Gatsby Brand: F. Scott Fitzgerald Categories: [Classic Literature, Fiction] Title: Harry Potter and the Philosopher's Stone Brand: J.K. Rowling Categories: [Fantasy, Young Adult] Instead, I disliked these books: Title: Fifty Shades of Grey Brand: E.L. James Categories: [Romance, Erotic Fiction] Title: Twilight Brand: Stephenie Meyer Categories: [Fantasy, Romance] Rank and return the top k movies from: Id: 1234 Title: The Great Gatsby Brand: Warner Bros. Pictures Categories: [Drama, Romance] Id: 5678 Title: Harry Potter and the Sorcerer's Stone Brand: Warner Bros. Pictures Categories: [Fantasy, Adventure] Id: 91011 Title: Gone Girl Brand: 20th Century Fox Categories: [Mystery, Thriller]
Instruction Output	
Model Output	Items ranking: 1234, 5678, 91011 Explanation: Based on the user's liked books, which include "The Great Gatsby" and "Harry Potter and the Philosopher's Stone", it's apparent that they enjoy classic books with elements of fiction and fantasy, particularly in the young adult genre.

$prompt_{test}$

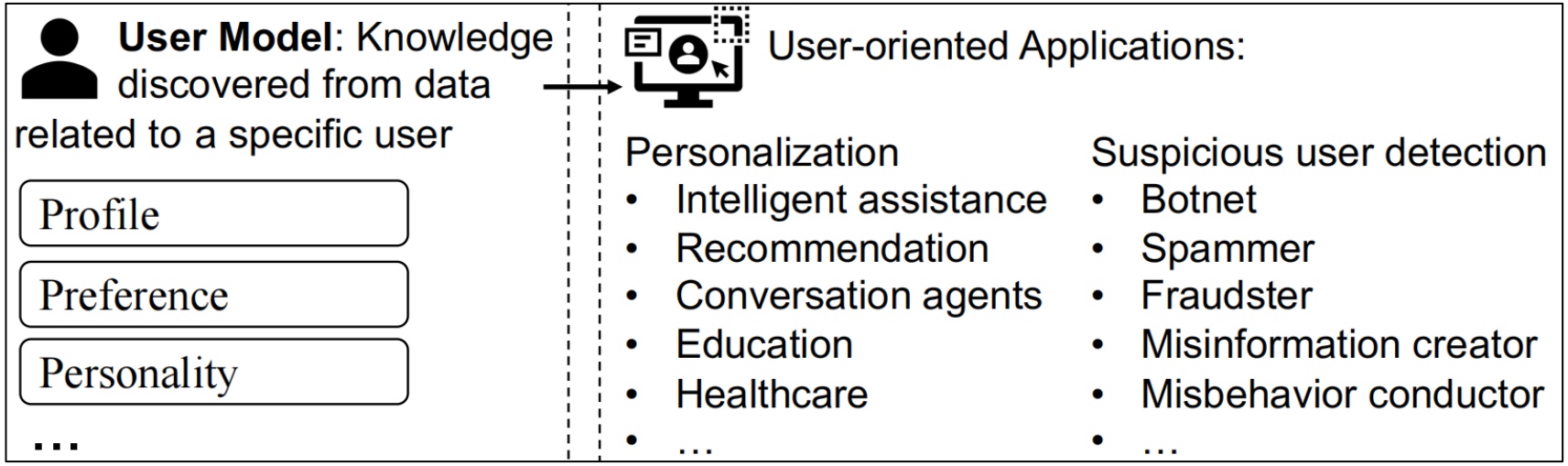
Instructing and Prompting Large Language Models for Explainable Cross-domain Recommendations(Recsys 2024)

总结与展望

- 目前基于大模型的可信用户偏好建模相关研究较少，多为评估类工作
- 研究点在于如何利用大模型的语义信息以增加可信性；或考虑LLMs是否会给用户偏好建模带来额外的不可信因素，是否可以用LLMs解决可信的相关手段来辅助解决基于大模型的用户偏好建模中存在的不可信问题
- 考虑交叉视角的可信问题，例如可解释的公平性

未来工作

- 写一篇有关可信用户建模的综述(预投IJCAI 2025 Survey Track)



- 基于解耦方法的可迁移公平性问题(预投SIGIR 2025)

国内研究团队

沈华伟 研究员

研究方向： 网络数据挖掘；社交网络分析；图神经网络

所属部门： 智能算法安全重点实验室

导师类别： 博导计算机软件与理论

联系方式： shenhuawei AT ict.ac.cn

个人网页： http://www.bigdatalab.ac.cn/~shenhuawei/

计算所 沈华伟

Xiangnan HE

何向南

Professor

School of Data Science
School of Information Science and Technology
University of Science and Technology of China

443 Huangshan Road, Hefei, China 230027

Email: xiangnanhe at gmail.com

中科大 何向南

Min Zhang

z-m AT tsinghua.edu.cn

Min Zhang, 张敏 in Chinese

Ph.D & Professor,
Information Retrieval Group
National Lab of Intelligent Tech. & Sys.
Department of Computer Sci.& Tech.
Tsinghua University, Beijing, China

z-m AT tsinghua.edu.cn
Telephone: +86-10-62798279

Research area: Web Information Retrieval and Recommendation, User Behavior Analysis and Profiling, Machine Learning, Data Mining.

清华大学 张敏

Chao Huang

Assistant Professor

Data Intelligence Lab@HKU

Department of Computer Science & Institute of Data Science

University of Hong Kong (HKU)

chaohuang75@gmail.com Google Scholar Lab Github

团队公众号 Twitter Linkedin 小红书

香港大学 黄超

国内研究团队

Mi Zhang

Professor
Head of Whitzard AI Team
System Software and Security Lab
School of Computer Science and Technology
Fudan University
Shanghai, China
Email: mi_zhang at fudan.edu.cn
Google Scholar| DBLP| Research Gate

复旦大学 张谧

郑小林 博士

教授 | 博士生导师

单位 计算机科
职务 人工智能

电话 0571-87951453
邮箱 xlzheng@zju.edu.cn

研究方向
- 金融科技
- 人工智能
- 推荐系统
- 隐私计算

浙江大学 郑小林

Q&A