推荐系统如何从大语言模型中取长补短：从应用视角出发

唐睿明 ---- 华为诺亚方舟实验室

DataFunSummit # 2023

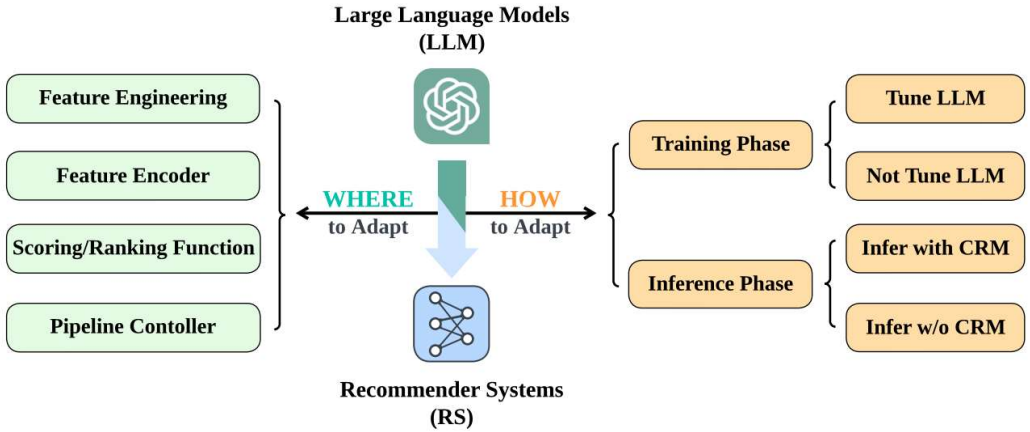
目录
CONTENT

01

背景和问题

背景和问题

- **传统的推荐系统**
  - 模型相对较小，时间空间开销低√
  - 可以充分利用协同信号√
  - 只能利用数据集内的知识×
  - 缺乏语义信息和深度意图推理×
- **大语言模型**
  - 引入外部开放世界知识，语义信号丰富√
  - 具备跨域推荐能力，适合冷启动场景√
  - 协同信号缺失×
  - 计算复杂度高，难以处理海量样本×
- **核心研究问题**
  - 推荐模型如何从大模型中取长补短，从而提升推荐性能，优化用户体验?
- 从应用角度出发，我们进一步将该问题拆解为
  - **何处运用大语言模型（WHERE to adapt）**
  - **如何运用大语言模型（HOW to adapt）**



LLM+RS：核心研究问题拆解

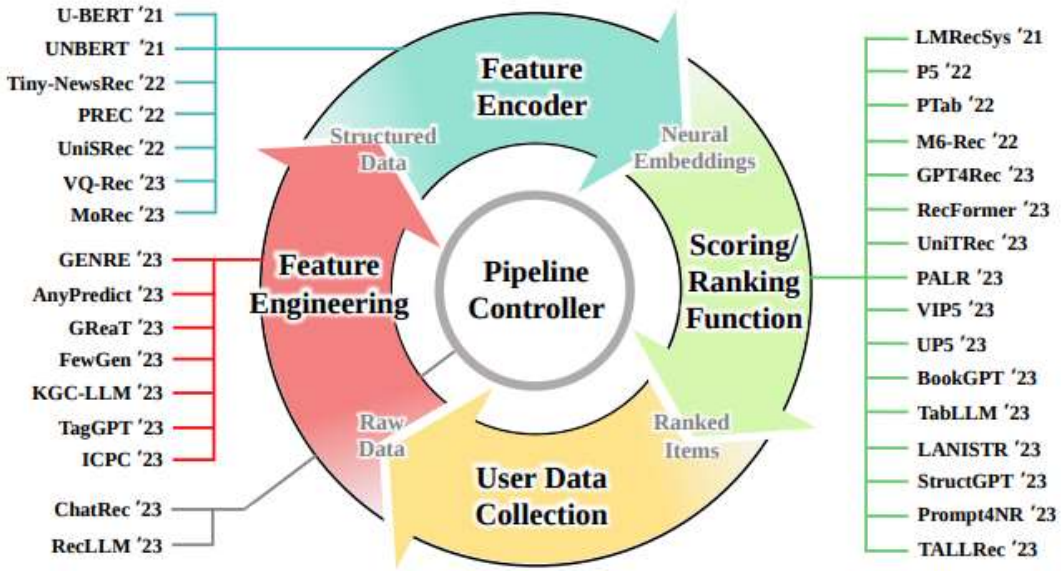
Lin J, Dai X, Xi Y, et al. How Can Recommender Systems Benefit from Large Language Models: A Survey. arXiv preprint, 2023.

02
何处运用大语言模型

DataFunSummit # 2023

何处运用大语言模型 (WHERE to adapt LLM)

- 根据现代基于深度学习的推荐系统的流程，我们抽象出以下五个环节：
  - **数据采集**阶段：线上收集用户行为和记录，得到原始数据（raw data）
  - **特征工程**阶段：对原始数据进行筛选、加工、增强，得到可供下游深度模型使用的结构化数据 (structured data)
  - **特征编码**阶段：对结构化数据进行编码，得到对应的稠密向量表示 (neural embeddings)
  - **打分排序**阶段：对候选物品进行打分排序，得到要呈现给用户的排序列表(recommended items)
  - 推荐**流程控制**：作为中央控制器，把控推荐系统的整体流程。也可以细化到对排序阶段的召回、粗排、精排的控制

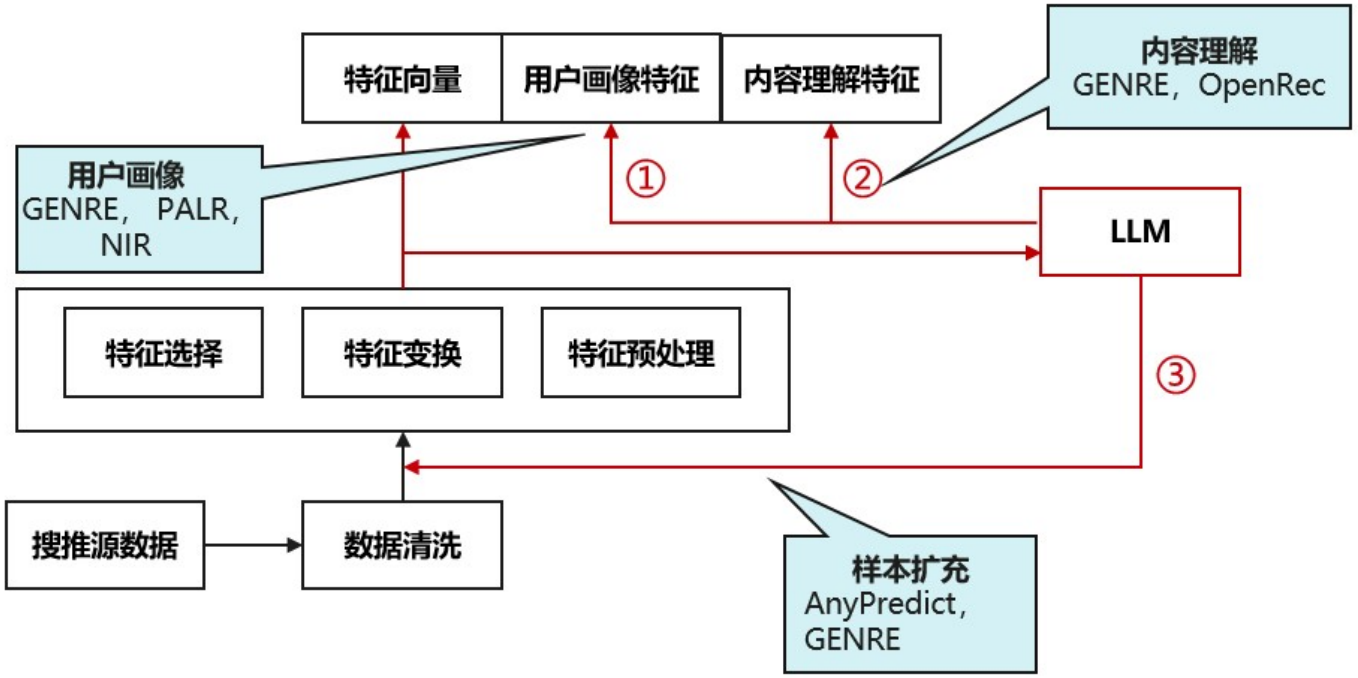


基于深度学习的推荐系统流程和不同阶段LLM应用的代表性工作

Lin J, Dai X, Xi Y, et al. How Can Recommender Systems Benefit from Large Language Models: A Survey. arXiv preprint, 2023.

特征工程

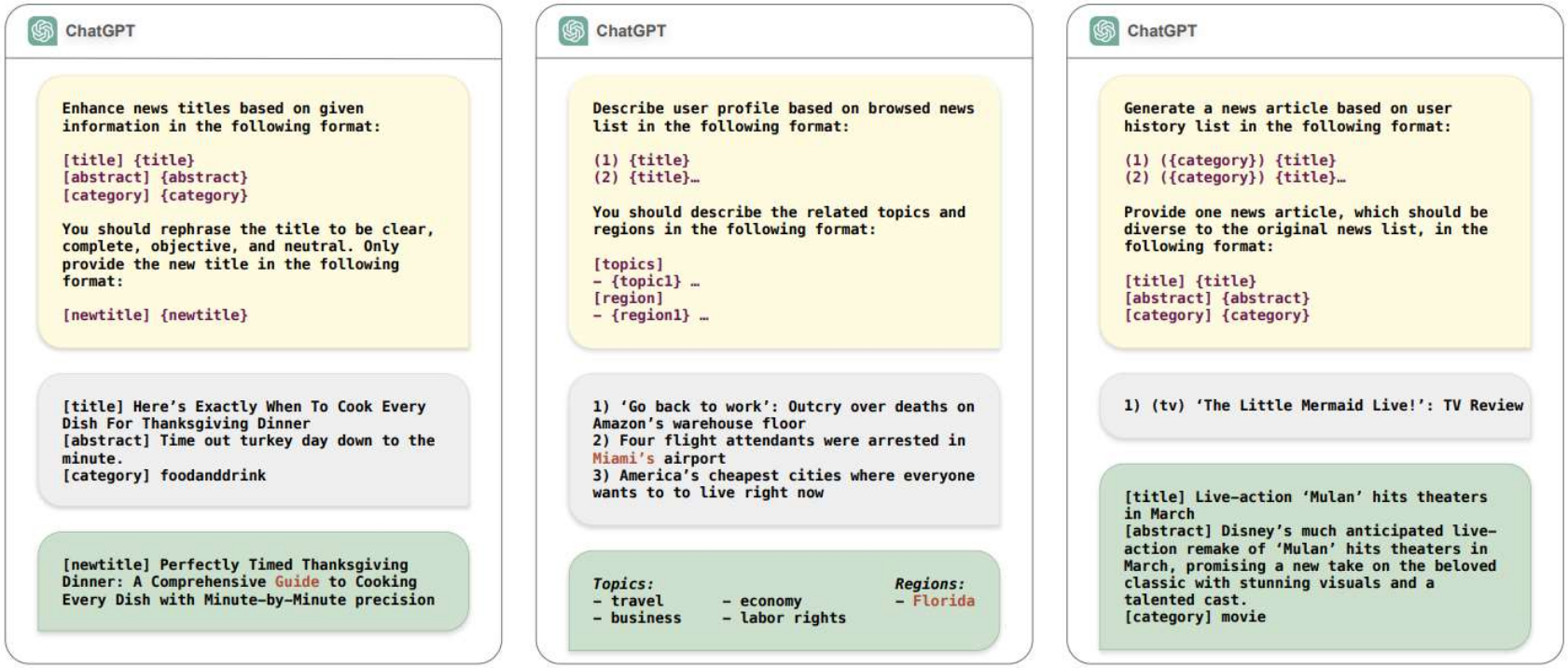
- 利用大语言模型的**外部通用知识**和**逻辑推理**能力，进行特征增强
  - 1. 丰富用户画像 2. 理解推荐内容 3.样本扩充



大语言模型在特征工程中的应用

特征工程

- GENRE
  - 在**新闻推荐**的场景下，利用LLM进行新闻摘要，用户画像和个性化新闻内容生成

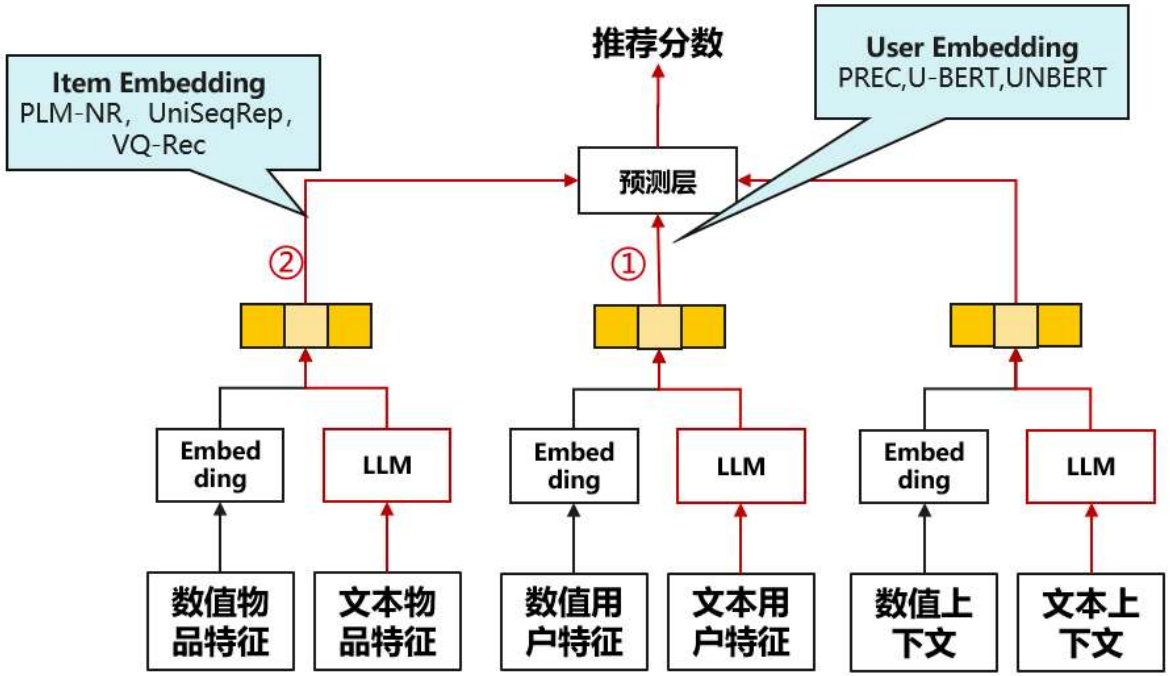


(a) News summarizer　(b) User profiler　(c) Personalized news generator

Liu Q, Chen N, Sakai T, et al. A First Look at LLM-Powered Generative News Recommendation. arXiv preprint, 2023.

特征编码

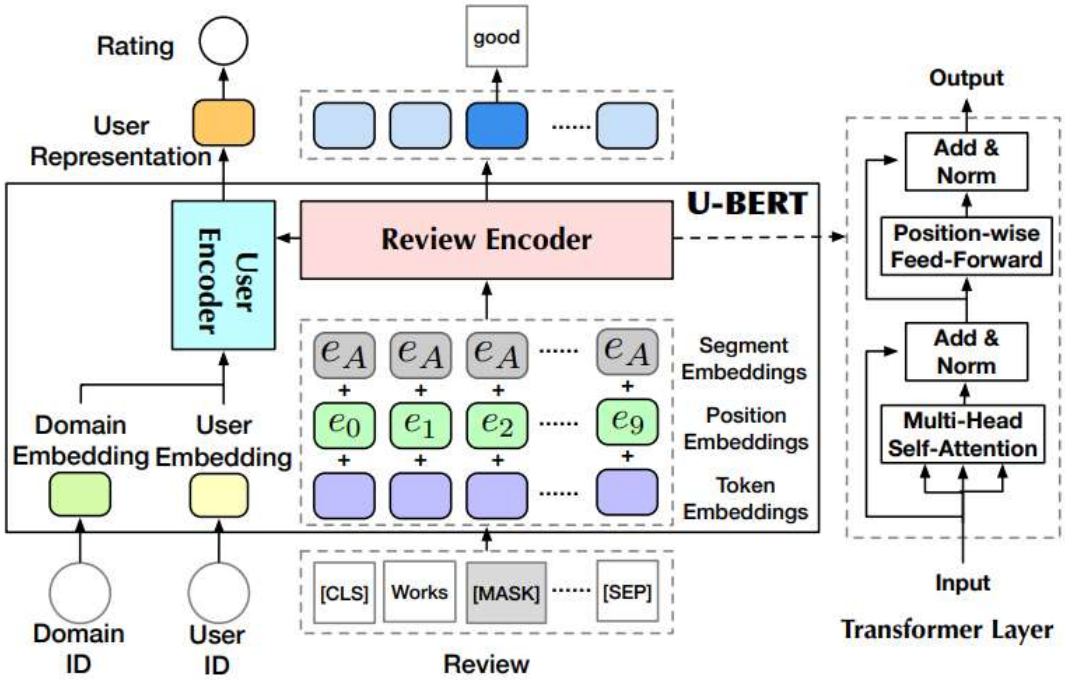
- 利用LLM的通用语义信息丰富推荐特征表示
  - 1.增强文本特征 (用户表征、物品表征) 表示 2.改善基于ID的特征表示的跨场景迁移能力



大语言模型在特征编码中的应用

特征编码

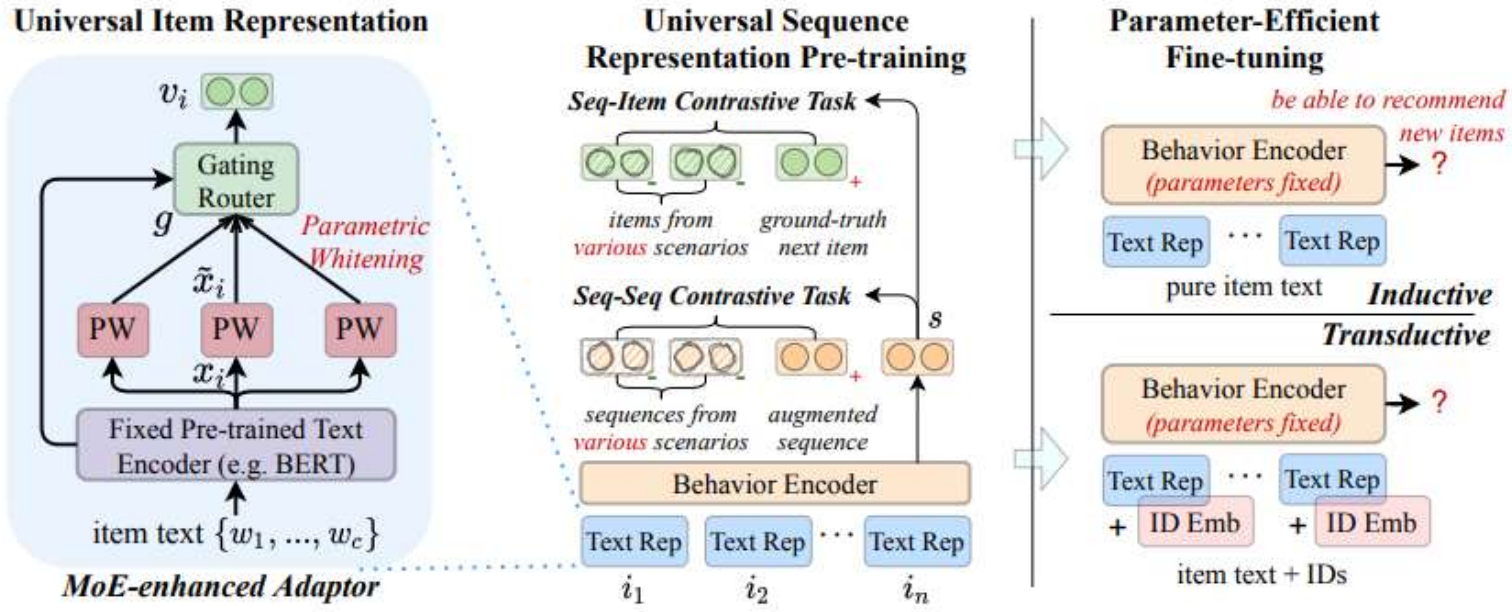
- U-BERT
  - **用户表征**：用语言模型对用户评论内容编码，增强用户的个性化表征



Qiu Z, Wu X, Gao J, et al. U-BERT: Pre-training user representations for improved recommendation. AAAI, 2021.

特征编码

- UniSRec
  - **物品表征**：通过对物品标题/描述进行编码，来实现跨域推荐的目标



Hou Y, Mu S, Zhao W X, et al. Towards universal sequence representation learning for recommender systems. KDD, 2022.

打分/排序

- **打分/排序**是推荐系统的核心任务，目标是得到和用户偏好相符的物品(列表)
- 根据如何得到最终排序列表的形式，我们将大语言模型应用于打分/排序的工作分成以下三种
  - **物品评分任务 (Item Scoring Task)**
    - 大语言模型对候选物品逐一评分，最后根据分数排序得到最终的排序列表
  - **物品生成任务 (Item Generation Task)**
    - 通过生成式的方式生成下一个物品的ID，或者直接生成排序列表
  - **混合任务 (Hybrid Task)**
    - 大语言模型天然地适合多任务场景，因此很多工作会利用大语言模型来实现多个推荐任务，其中包括评分任务和生成任务

物品评分任务 (Item Scoring Task)

- 探究语言模型分别在零样本 (Zero-Shot)，少样本 (Few-Shot)和微调场景下的评分预测的能力
  - 零样本和少样本

Zero-Shot User Rating Predictor

Given a user's past movie ratings in the format: Title, Genres, Rating. Ratings range from 1.0 to 5.0.

Babe (1995), Children's|Comedy|Drama, 4
There's Something About Mary (1998), Comedy, 4
Awakenings (1990), Drama, 4
Simple Plan, A (1998), Crime|Thriller, 4
Bug's Life, A (1998), Animation|Children's|Comedy, 5
Twelve Monkeys (1995), Drama|Sci-Fi, 5
Pleasantville (1998), Comedy, 4
Apollo 13 (1995), Drama, 4
Misery (1990), Horror, 4
South Park: Bigger, Longer and Uncut (1999), Animation|Comedy, 4

The candidate movie is Player, The (1992), Comedy|Drama. What's the rating that the user will give?

Few-Shot User Rating Predictor

You're required to predict user ratings for movie recommendations. The rating ranges from 1.0 to 5.0. You're given the user's past rating history, in the format of in the format: Title, Genres, Rating.

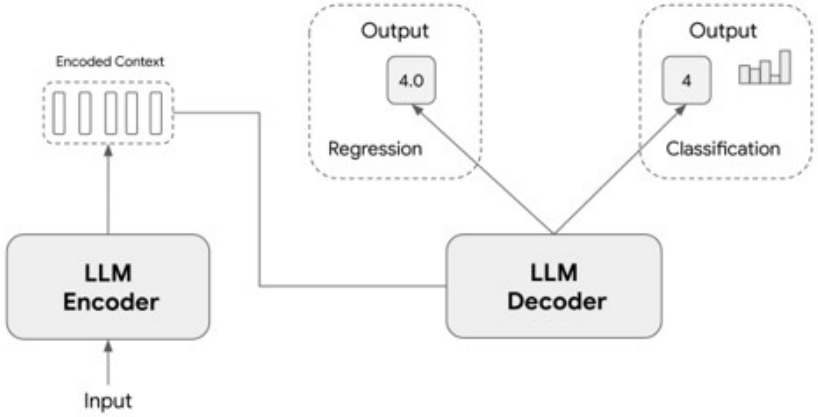
Example:
Q: The user history is:
Jerry Maguire (1996), Drama|Romance, 2.0
Fight Club (1999), Action|Crime|Drama|Thriller, 4.0
…
Ghostbusters (a.k.a. Ghost Busters) (1984), Action|Comedy|Sci-Fim 4.0
The user now watched (Fifth Element, Sci-fi|Action), give a single number as rating without saying anything else. Do not give reasoning.
A: 4.0

Q: The user history is:
Babe (1995), Children's|Comedy|Drama, 4
There's Something About Mary (1998), Comedy, 4
…
Misery (1990), Horror, 4
South Park: Bigger, Longer and Uncut (1999), Animation|Comedy, 4
The user now watched (Player, The (1992), Comedy|Drama), give a single number as rating without saying anything else. Do not give reasoning.
A:

Kang W C, Ni J, Mehta N, et al. Do LLMs Understand User Preferences? Evaluating LLMs On User Rating Prediction. arXiv preprint, 2023.

物品评分任务 (Item Scoring Task)

- 探究语言模型分别在零样本 (Zero-Shot)，少样本 (Few-Shot)和微调场景下的评分预测的能力
  - 微调语言模型

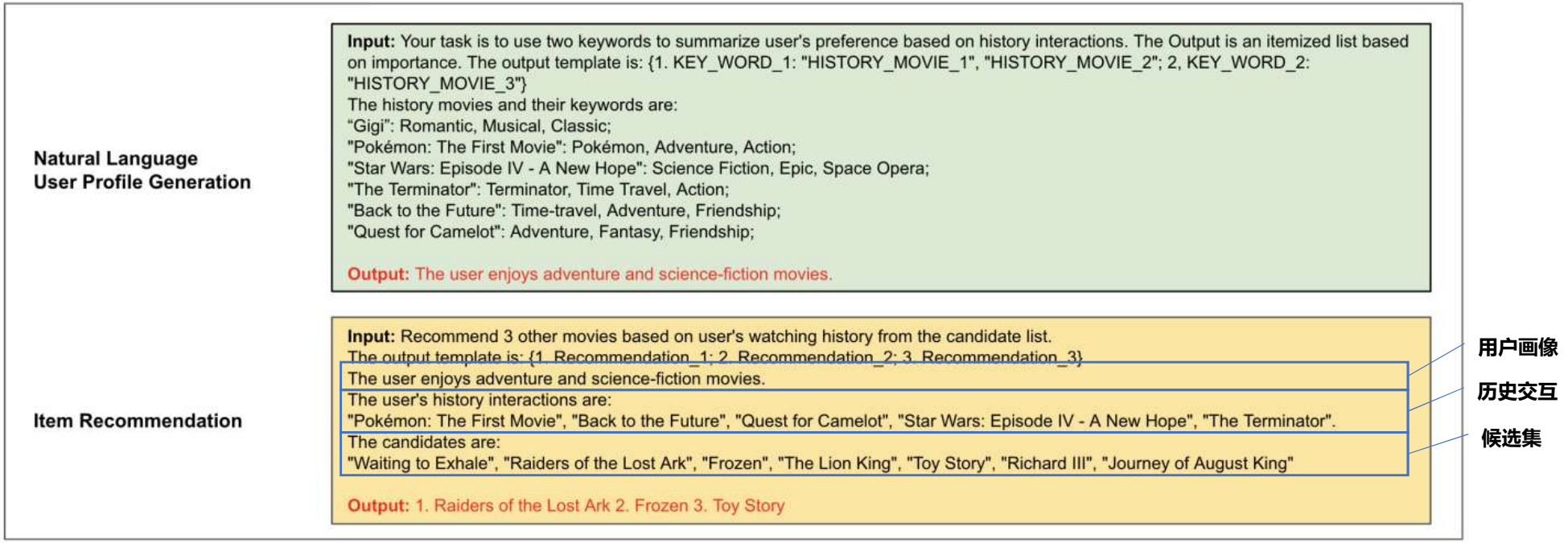


Model	MovieLens			Amazon-Books		
	RMSE↓	MAE↓	AUC↑	RMSE↓	MAE↓	AUC↑
Zero-Shot LLMs						
Flan-U-PALM	1.0677	0.7740	0.7084	0.9565	0.5569	0.7676
ChatGPT	1.0081	0.8193	0.6794	1.0081	0.8093	0.6778
text-davinci-003	1.0460	0.7850	0.6951	0.8890	0.5442	0.7416
Few-Shot LLMs						
Flan-U-PALM	1.0721	0.7605	0.7094	1.0712	0.5855	0.7439
ChatGPT	1.0862	0.8203	0.6930	1.0618	0.7760	0.7470
text-davinci-003	1.0867	0.8119	0.6963	1.0716	0.7753	0.7739
Simple Dataset Statistics						
Global Avg. Rating	1.1564	0.9758	0.5	0.9482	0.7609	0.5
Candidate Item Avg. Ratings	0.9749	0.7778	0.7395	0.9342	0.7078	0.6041
User Past Avg. Ratings	1.0196	0.7959	0.7266	0.8527	0.5502	0.8047
Supervised Recommendation Methods						
MF	0.9552	0.7436	0.7734	1.7960	1.1070	0.7638
MLP	0.9689	0.7452	0.7393	0.8607	0.6384	0.6932
Transformer+MLP	**0.8848**	0.7036	0.7979	**0.8143**	0.5541	0.8042
Fine-tuned LLMs						
Flan-T5-Base (classification)	1.0110	0.6805	0.7590	0.9856	0.4685	0.6292
Flan-T5-Base (regression)	0.9187	0.7092	0.7949	0.8413	0.5317	0.8182
Flan-T5-XXL (regression)	0.8979	**0.6986**	**0.8042**	0.8301	**0.5122**	**0.8312**

Kang W C, Ni J, Mehta N, et al. Do LLMs Understand User Preferences? Evaluating LLMs On User Rating Prediction. arXiv preprint, 2023.

物品生成任务 (Item Generation Task)

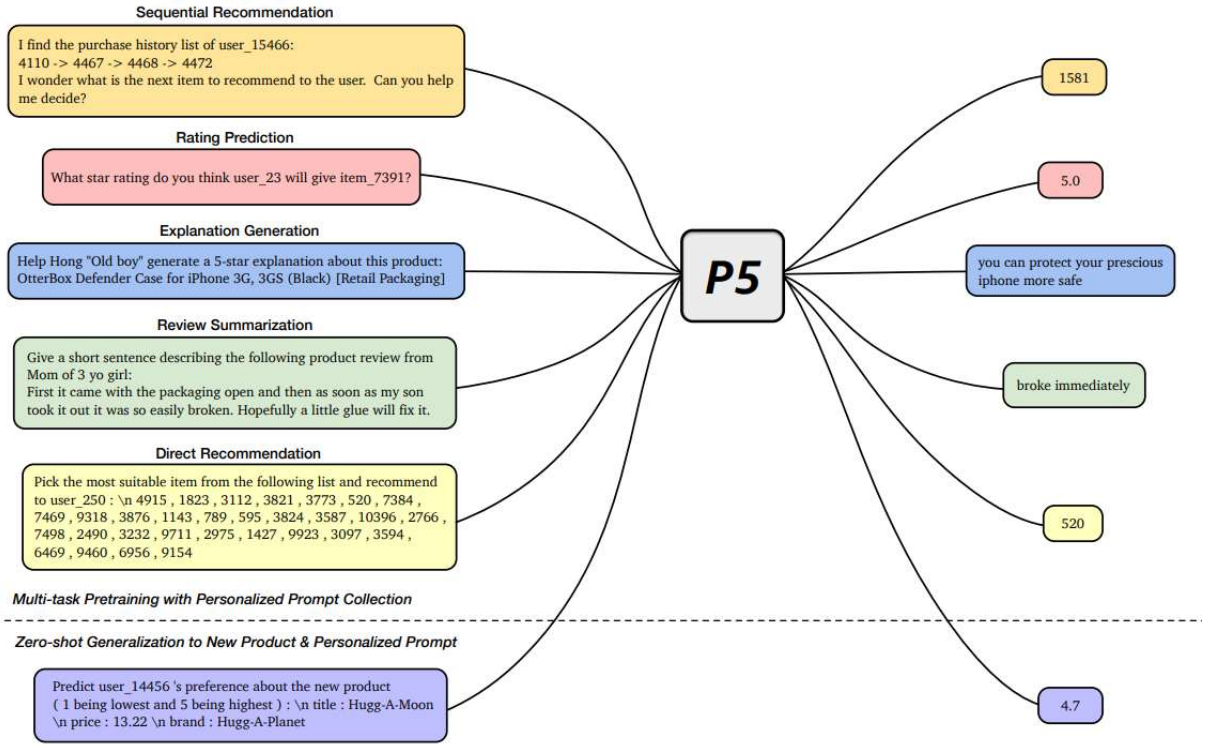
- PALR
  - 利用用户历史交互得到用户画像，然后基于用户画像、历史交互和提前过滤得到的候选集信息生成推荐列表



Yang F, Chen Z, Jiang Z, et al. PALR: Personalization Aware LLMs for Recommendation. arXiv preprint, 2023.

混合任务 (Hybrid Task)

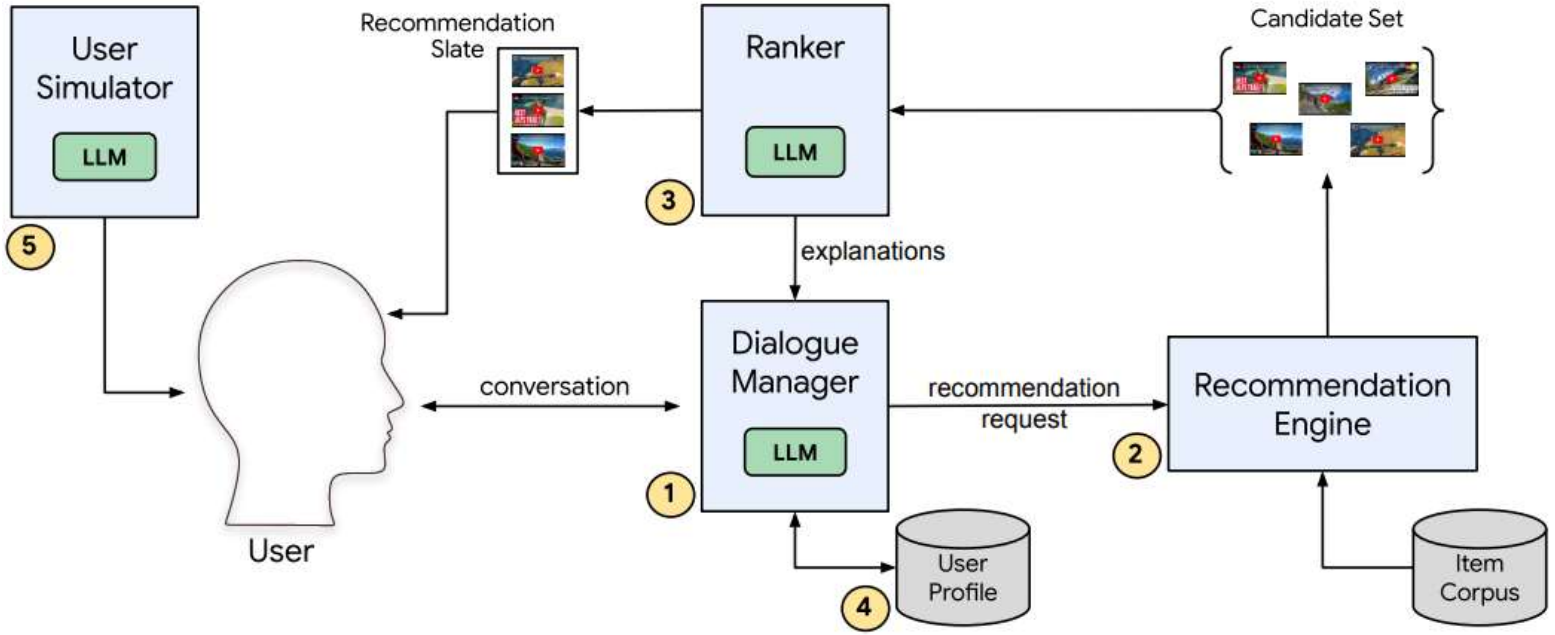
- P5
  - 用一个统一的大语言模型在不同的推荐任务上进行预训练，针对不同任务使用不同推荐模版



Geng S, Liu S, Fu Z, et al. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5) .RecSys. 2022.

流程控制

- **RecLLM**
  - 提出了一种使用LLM来集成推荐系统流程各模块(检索、排序、用户画像、用户模拟)的一个**对话式推荐系统**路线图



Friedman L, Ahuja S, Allen D, et al. Leveraging Large Language Models in Conversational Recommender Systems. arXiv preprint, 2023.

03

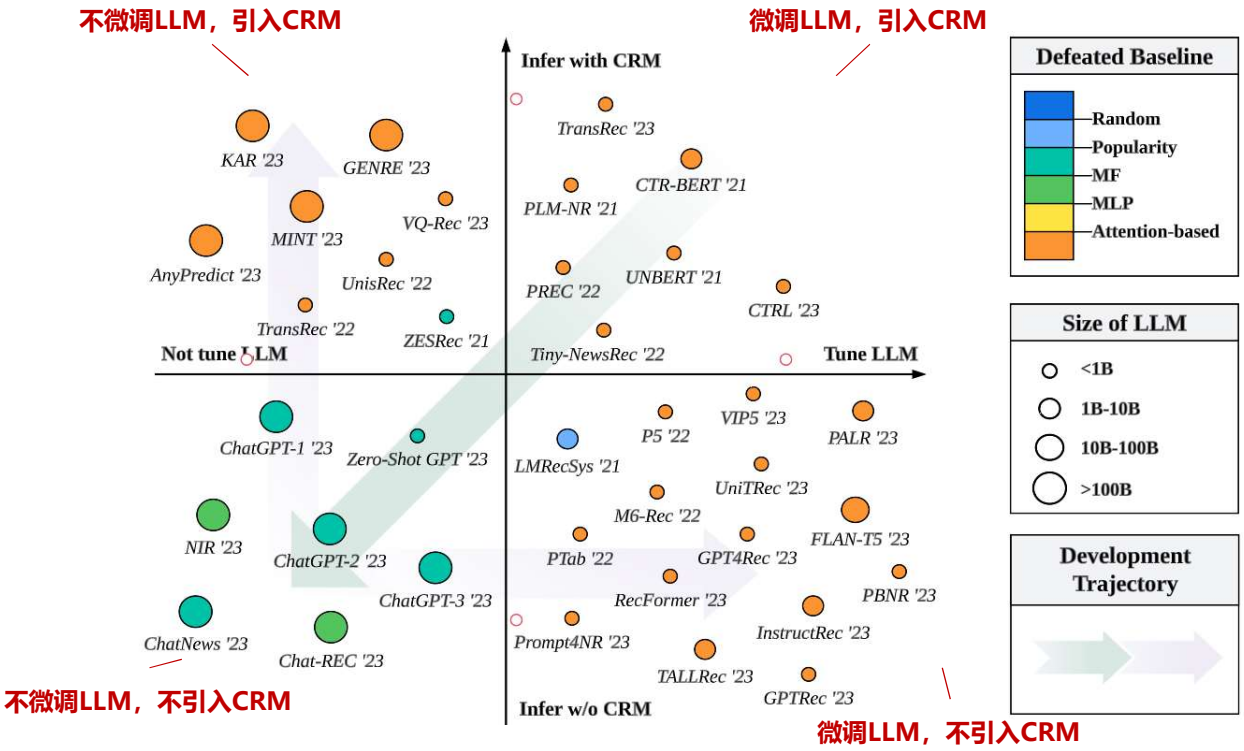
如何运用大语言模型

DataFunSummit # 2023

如何运用大语言模型 (HOW to adapt LLM)

- 从训练和推理两个阶段出发，我们根据以下的两个维度将现有工作分为四个象限：

- 在**训练阶段**，大语言模型是否需要微调。这里微调的定义包含了全量微调和参数高效微调。
- 在**推理阶段**，是否需要引入传统推荐模型(Conventional Recommendation Model，CRM)。其中，如果CRM知识作为一个预先过滤candidate的作用，则不被考虑在内。



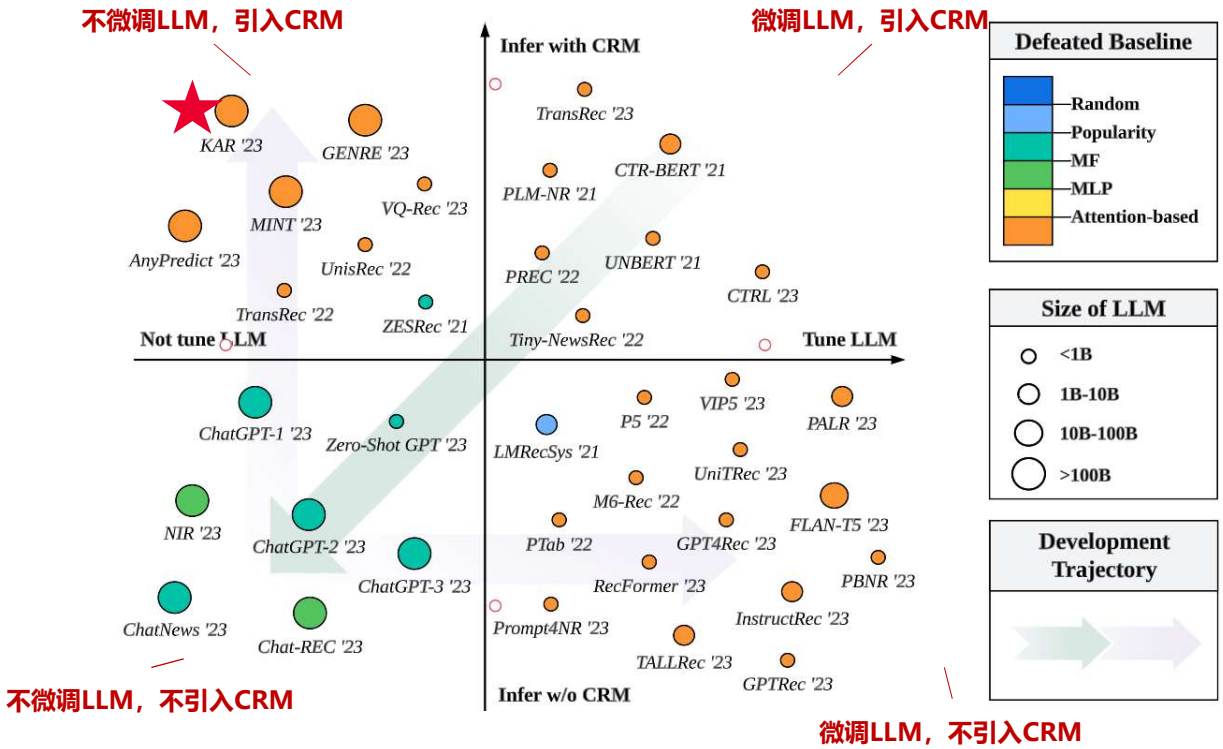
大语言模型在推荐系统应用的四象限图及代表性工作

两个趋势

- **模型：**通过引入传统推荐模型（CRM）为语言模型注入协同信号。
- **数据：**通过引入推荐场景的数据，结合微调技术，为语言模型注入协同信号。

Lin J, Dai X, Xi Y, et al. How Can Recommender Systems Benefit from Large Language Models: A Survey. arXiv preprint, 2023.

利用大语言模型开放知识辅助推荐的通用推荐框架KAR



大语言模型在推荐系统应用的四象限图及代表性工作

两个趋势

- **模型：**通过引入传统推荐模型（CRM）为语言模型注入协同信号。
- **数据：**通过引入推荐场景的数据，结合微调技术，为语言模型注入协同信号。

仅使用语言模型进行推荐目前效果不理想。

Xi Y, Liu W, Lin J, et al. Towards Open-World Recommendation with Knowledge Augmentation from Large Language Models. arXiv preprint,2023.

方法创新性

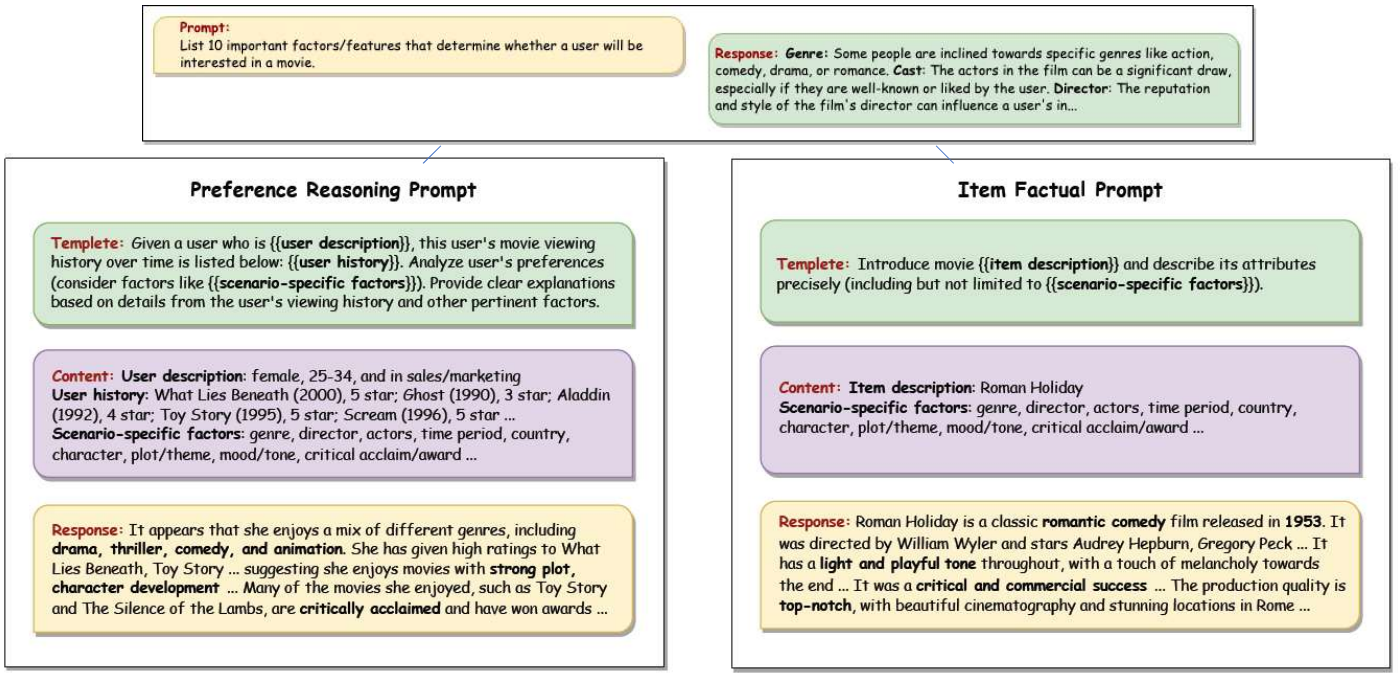
有效结合大语言模型的通用世界知识与传统推荐系统的领域知识

- 通过**Factorization Prompting**，有效激发大语言模型针对用户兴趣的推理和知识获取能力，缓解构成差距的问题
- 提出**混合多专家网络知识适配器**，将语言模型生成的世界知识由语义空间适配至推荐空间，进行降维、噪声处理的同时保存有效信息。
- 推理复杂度与传统推荐模型一致 $O(f(n,m))$。

通用性强，适配各类型工业推荐场景和任务

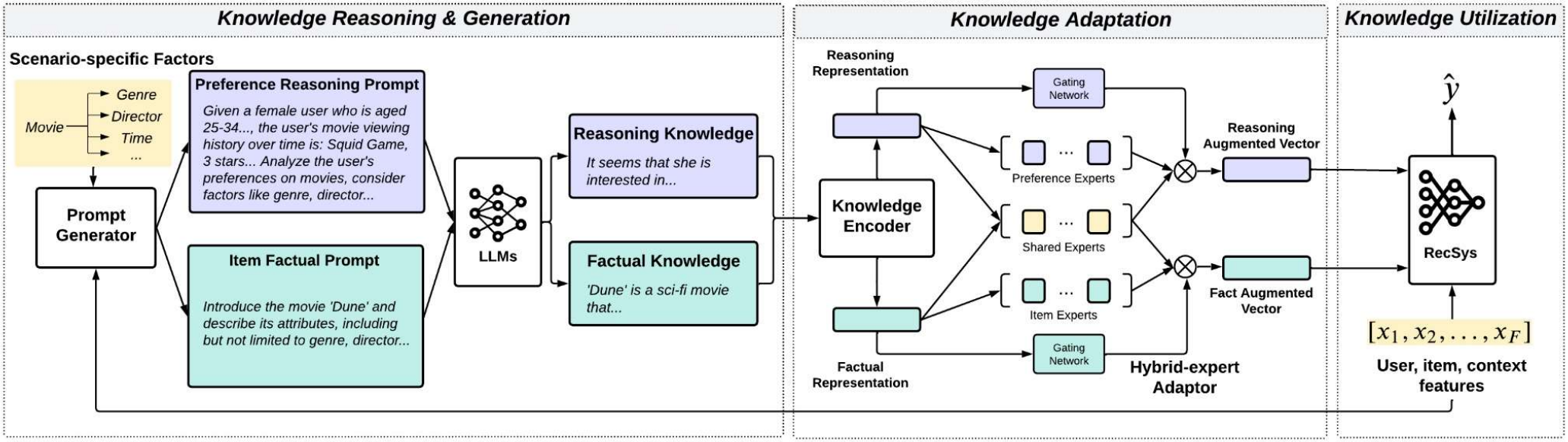
■ **大语言模型的构成差距问题（Compositional Gap）**

- 无法准确回答复合推理问题，但可以准确回答复合问题的各个子问题
- 用户偏好复杂多面，世界知识混合海量，难以直接生成有效知识
- 受因式分解的启发，将开放知识生成问题动态分解为多个关键子因素，按因素分别进行用户兴趣推理和知识提取。



Xi Y, Liu W, Lin J, et al. Towards Open-World Recommendation with Knowledge Augmentation from Large Language Models. arXiv preprint,2023.

技术方案



知识推理和生成：

- 基于推荐场景对于决定用户偏好，动态分解出相应的关键因素，对于用户偏好和物品外部知识分别对大语言模型提问
- 生成相应的兴趣推理知识和物品事实知识文本

知识适配：

- 所生成的文本信息内容复杂多面（500~1000 tokens），且存在幻觉问题，推荐系统无法直接理解和利用
- 设计多专家网络进行知识提取、压缩、映射，适配至推荐空间，输出结果鲁棒。

知识利用：

- 将所生成的知识增强向量作为额外的特征域，结合原本数据特征，进行特征交互，输出最终结果。

Xi Y, Liu W, Lin J, et al. Towards Open-World Recommendation with Knowledge Augmentation from Large Language Models. arXiv preprint,2023.

有益效果

- 【通用性】在9个SOTA的推荐算法上，平均AUC显著提升1.5%（AUC 3‰以上的提升即为显著）

Backbone Model	AUC			Logloss		
	base	KAR	improv.	base	KAR	improv.
DCNv2	0.7924	**0.8049***	1.58 %	0.5451	**0.5315***	2.50 %
DCN	0.7929	**0.8043***	1.46 %	0.5457	**0.5319***	2.53 %
DeepFM	0.7928	**0.8041***	1.44 %	0.5462	**0.5321***	2.57 %
FiBiNet	0.7925	**0.8051***	1.59 %	0.5450	**0.5310***	2.56 %
AutoInt	0.7934	**0.8060***	1.59 %	0.5440	**0.5297***	2.65 %
FiGNN	0.7944	**0.8054***	1.39 %	0.5424	**0.5307***	2.16 %
xDeepFM	0.7942	**0.8041***	1.25 %	0.5457	**0.5317***	2.57 %
DIEN	0.7960	**0.8059***	1.25 %	0.5469	**0.5298***	3.13 %
DIN	0.7975	**0.8066***	1.15 %	0.5387	**0.5304***	1.55%

∗ denotes statistically significant improvement (t-test with p-value < 0.05) over the backbone model.

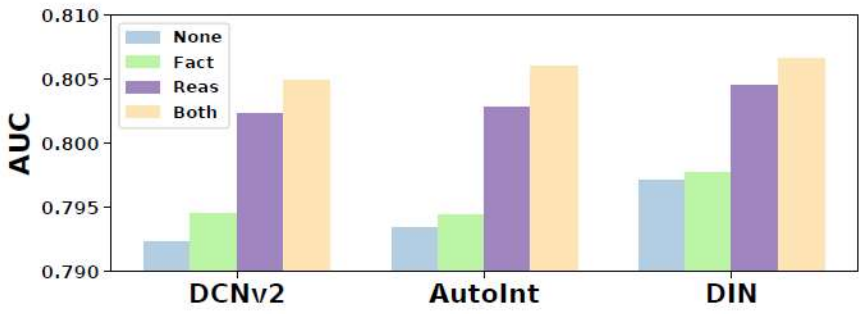
- 【可落地性】推理复杂度与传统推荐模型相当。

Model	Inference time (s)
LLM API	5.54
$KAR_{w/\ apt}$	8.08×10^{-2}
$KAR_{w/o\ apt}$	6.64×10^{-3}
base	6.42×10^{-3}

- 【有效性】相比SOTA预训练推荐模型，AUC显著提升1%以上。且用户推理知识和物品事实知识都提供显著的增强效果，二者联合使用效果更优。

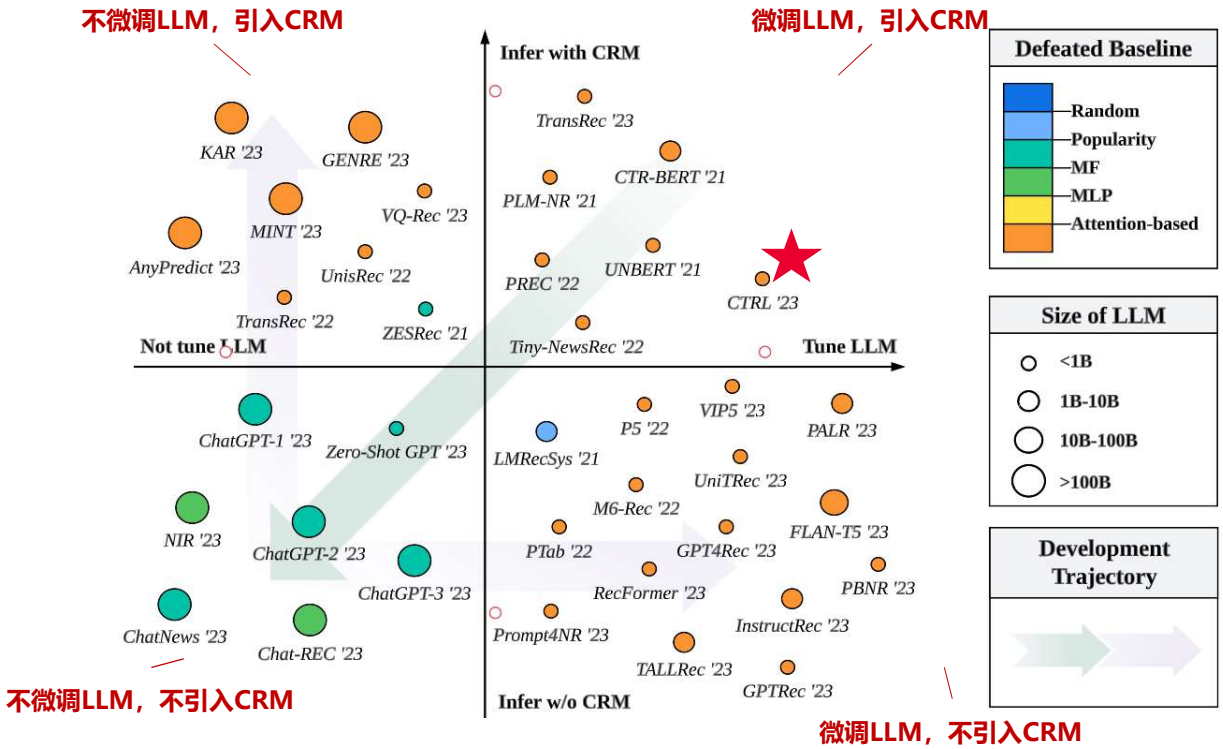
Model	AUC	Logloss
UnisRec	0.7891	0.5496
VQ-Rec	0.7914	0.5456
base(DIN)	0.7975	0.5387
KAR(DIN)	**0.8066***	**0.5304***

∗ denotes statistically significant improvement (t-test with p-value < 0.05) over the baseline/backbone models.



Xi Y, Liu W, Lin J, et al. Towards Open-World Recommendation with Knowledge Augmentation from Large Language Models. arXiv preprint,2023.

一种对齐语言模型和协同模型的框架CTRL



大语言模型在推荐系统应用的四象限图及代表性工作

两个趋势

- **模型：**通过引入传统推荐模型（CRM）为语言模型注入协同信号。
- **数据：**通过引入推荐场景的数据，结合微调技术，为语言模型注入协同信号。

仅使用语言模型进行推荐目前效果不理想。

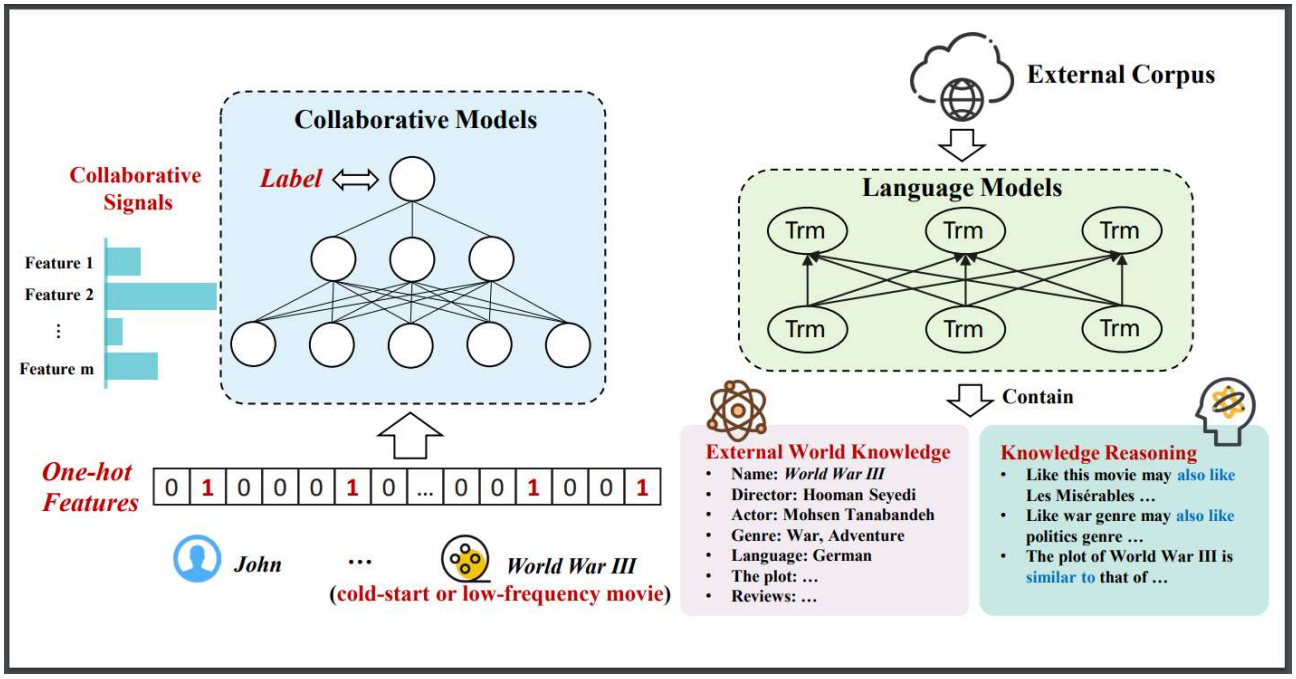
Li X, Chen B, Hou L, et al. CTRL: Connect Tabular and Language Model for CTR Prediction. arXiv preprint, 2023.

方法创新性

- 以**混合粒度知识对齐**的方式，同时建模协同信号和语义信号
- 从数据角度进行**双向知识注入**，语言模型与推荐模型互相解耦
- 可以**单侧推理**，推理复杂度低

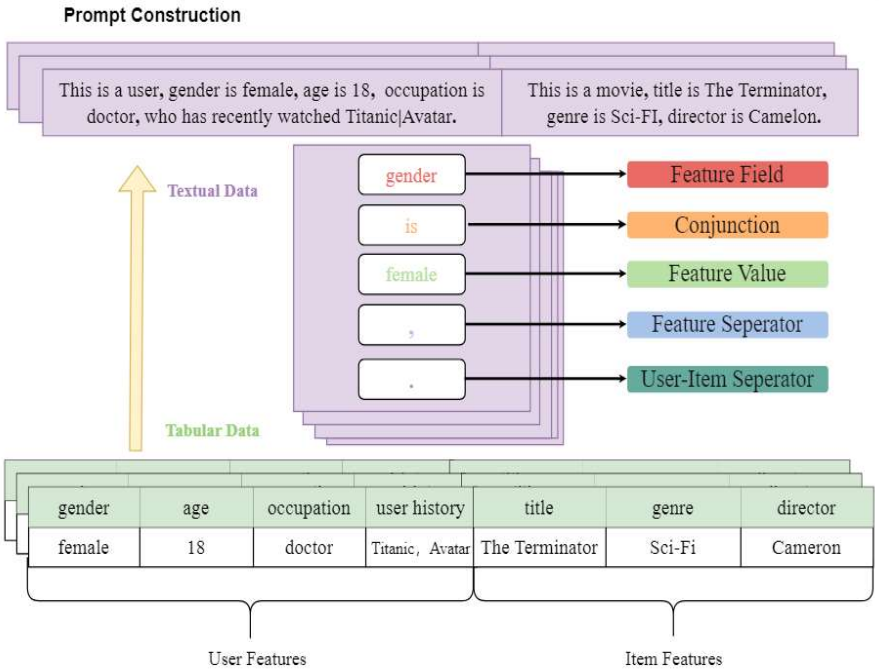
现有LLM4Rec的缺陷

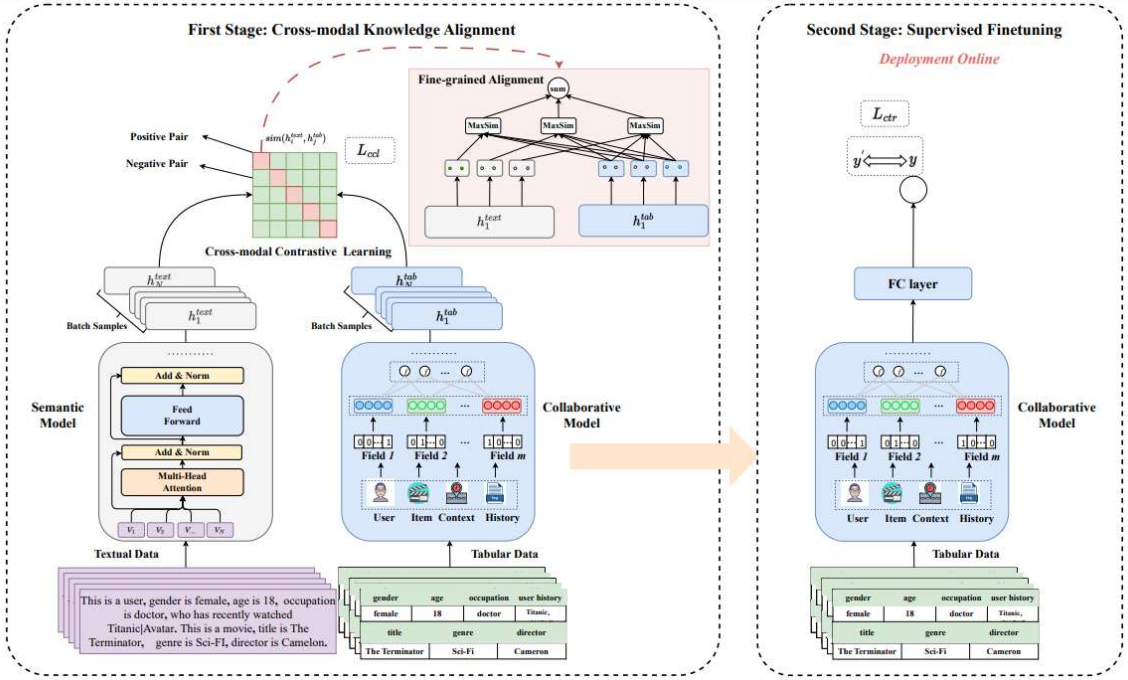
- 缺乏协同信号，在推荐下游任务准确率较低
- 在线推理时延过高，难以满足工业需求



Li X, Chen B, Hou L, et al. CTRL: Connect Tabular and Language Model for CTR Prediction. arXiv preprint, 2023.

技术方案





Prompt construction:

通过**7个**模板把表格数据转换为文本数据：

- 用户和物品特征：**特征名+连接词+特征值**
- 用户历史行为序列：**用户的历史类型+动作连接词+历史1|历史2|历史3**
- 采用，作为特征之间的分隔符；采用。作为用户信息和物品信息的分隔符

Cross-model Knowledge Alignment:

- 将协同模型和语言模型的知识进行**对齐**
- 利用**对比学习预训练**融合两种模态的信息
- 进一步利用**细粒度**的对比学习使信息融合更加充分

Supervised Finetuning:

- 在经过细粒度对比学习预训练之后，两种模态的信息已经进行充分融合
- 使用监督信号使协同模型**适配下游任务**
- 通过在不同的任务上微调，可以**适配不同的推荐任务**

Li X, Chen B, Hou L, et al. CTRL: Connect Tabular and Language Model for CTR Prediction. arXiv preprint, 2023.

有益效果

- 【推荐效果】和推荐模型和语言模型基线相比，AUC取得显著提升

Category	Model	MovieLens		Amazon		Alibaba	
		AUC	Logloss	AUC	Logloss	AUC	Logloss
Collaborative Models	DSSM	0.7901	0.4826	0.6481	0.4815	0.5696	0.3559
	Wide&Deep	0.8261	0.4248	0.6968	0.4645	0.6272	0.1943
	DeepFM	0.8268	0.4219	0.6969	0.4645	0.6280	0.1951
	DCN	0.8313	0.4165	0.6999	0.4642	0.6281	0.1949
	AutoInt	0.8290	0.4178	0.7012	0.4632	0.6279	0.1948
Semantic Models	P5	0.5541	0.5841	0.5333	0.5475	0.5556	0.3584
	CTR-BERT	0.7650	0.4944	0.6934	0.4629	0.6005	0.2020
	P-Tab	0.8031	0.4612	0.6942	0.4625	0.6112	0.3584
CTRL		**0.8376**⋆	**0.4025**⋆	**0.7074**⋆	**0.4577**⋆	**0.6338**⋆	**0.1890**⋆
Rel Impr.		0.76%	3.36%	0.88%	1.18%	0.91%	2.97%

- 【推理效率】和传统的推荐模型相比，推理效率**基本一致；**和语言模型相比，**显著减少推理时延**

	Alibaba		Amazon	
Model	Params	Inf Time	Params	Inf Time
DSSM	6.71×10^7	15s	3.35×10^7	0.51s
DeepFM	8.82×10^7	18s	3.45×10^7	0.58s
DCN	8.84×10^7	19s	3.46×10^7	0.58s
AutoInt	8.82×10^7	19s	3.45×10^7	0.59s
P5	2.23×10^8	10832s	1.10×10^8	440s
CTR-Bert	1.10×10^8	4083s	1.10×10^8	144s
CTRL	8.82×10^7	19s	3.45×10^7	0.59s

Li X, Chen B, Hou L, et al. CTRL: Connect Tabular and Language Model for CTR Prediction. arXiv preprint, 2023.

04

挑战和展望

工业应用场景下的挑战

- **训练效率**
  - 问题：显存用量过大、训练时间过长
  - 可能解决思路：1. 参数高效微调(PEFT)方案 2.调整模型更新频率（e.g.长短更新周期结合）
- **推理时延**
  - 问题：推理时延过高
  - 可能解决思路：
    - 预存部分输出或中间结果，以空间换时间；
    - 通过蒸馏、剪枝、量化等方法，降低推理模型的真实规模；
    - 仅用于特征工程和特征编码，避免直接在线上做模型推理
- **推荐领域的长文本建模**
  - 问题：长用户序列、大候选集、多元特征都会导致推荐文本过长，不仅难以被大模型有效捕捉，甚至可能会超过语言模型的**上下文窗口限制** (Context Window Limitation)
  - 可能解决思路：通过过滤、选择、重构，提供真正简短有效的文本输入
- **ID特征的索引和建模**
  - 问题：纯ID类特征(e.g. 用户ID)天然不具备语义信息，无法被语言模型理解
  - 可能解决思路：探索更适合语言模型的ID索引和建模策略

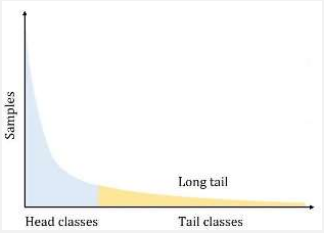
总结和展望

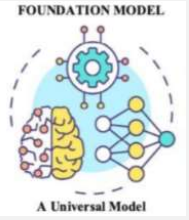
总结

- 从应用视角出发，以推荐系统为核心，我们调研了以下两个核心问题：
  - 何处运用大语言模型（WHERE to adapt）
  - 如何运用大语言模型（HOW to adapt）
- 现有的语言模型在推荐系统中的应用存在以下两个发展趋势：
  - **突破传统定位，重塑推荐流程**
  - LLM在推荐系统中扮演的角色逐渐突破传统定位，从简单的编码器、打分器逐渐向外延伸，在特征工程，乃至推荐流程控制都发挥重要作用
  - **语义协同兼顾，跨域知识融合**
  - 需要通过微调大语言模型（数据层面）或引入传统推荐模型（模型层面）的方式来为语言模型注入推荐的域内知识

展望

- **缓解稀疏场景**
  - LLM的zero-shot和few-shot能力可以用于解决冷启动和长尾问题
- **引入外部知识**
  - LLM拥有大量关于Item的世界知识，对于资讯类场景这种通用知识的引入可以大大丰富Item侧的信息
- **改善交互体验**
  - 用户可以主动通过交互式界面自由地描述他们的需求，从而实现精准推荐






感谢观看
DataFun.