

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

推荐模型如何从大语言模型种取长补短，从而提升推荐性能，优化用户体验？

02 何处运用大语言模型 (Where)

大语言模型可以用于特征工程、特征编码、打分排序、流程控制

03 如何运用大语言模型 (How)

总结大语言模型用于推荐系统的两个关键趋势，并分别介绍两个技术方案

04 挑战和展望

从应用视角出发，总结大语言模型用于推荐系统的挑战，并展望未来趋势

01

背景和问题

DataFunSummit # 2023



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×

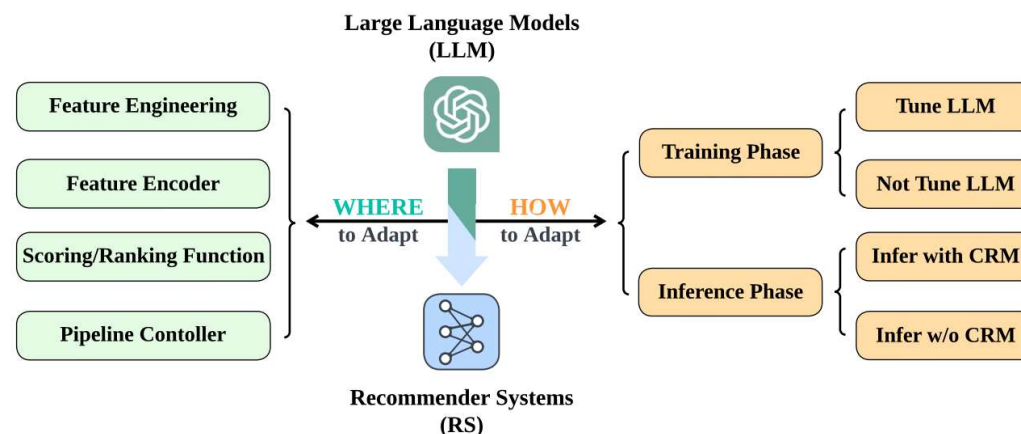
■ 核心研究问题

- 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

02

何处运用大语言模型

DataFunSummit # 2023

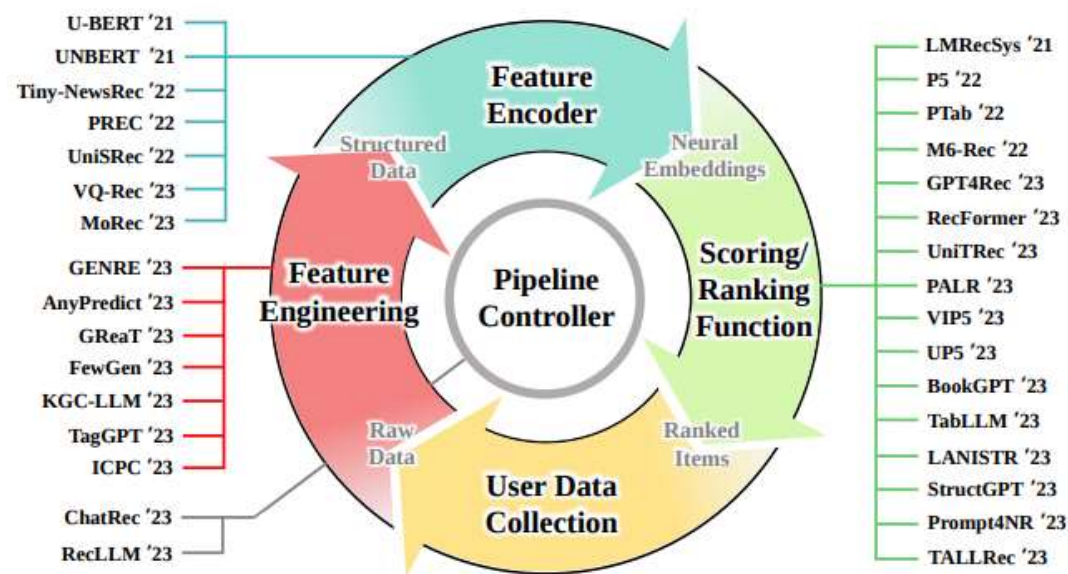


何处运用大语言模型 (WHERE to adapt LLM)



■ 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：

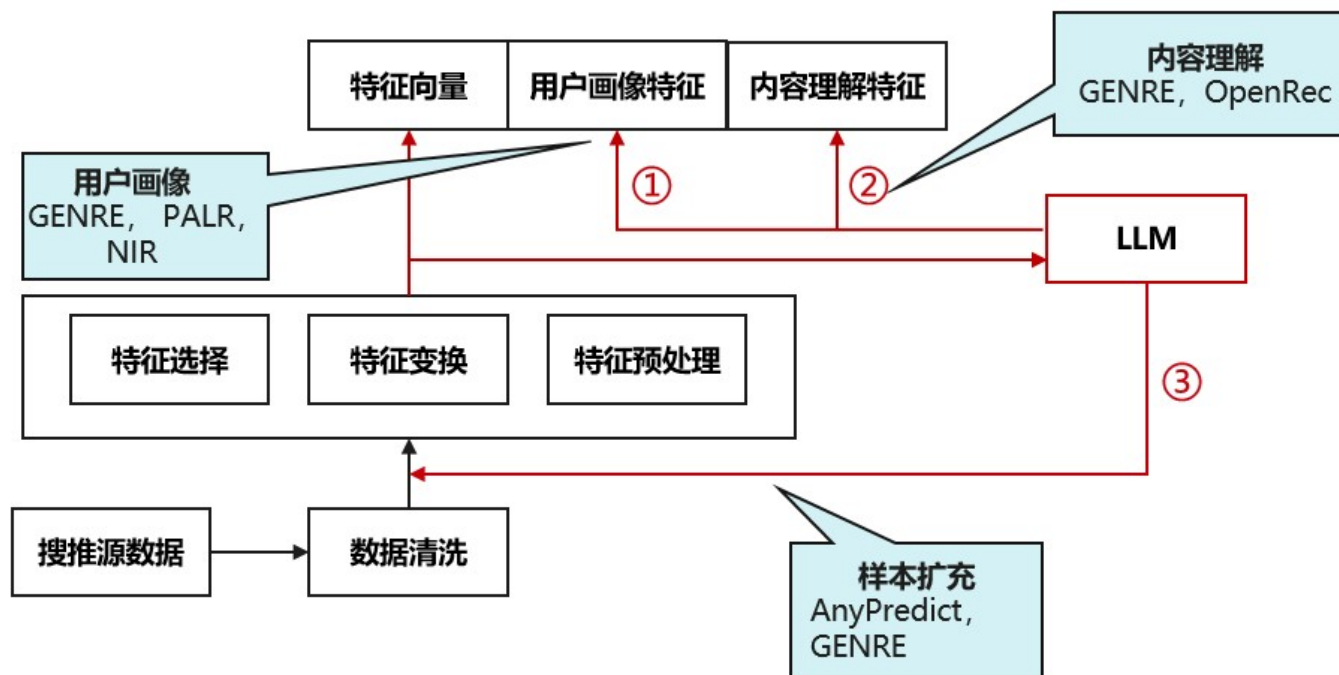
- **数据采集阶段**：线上收集用户行为和记录，得到**原始数据** (raw data)
- **特征工程阶段**：对原始数据进行筛选、加工、增强，得到可供下游深度模型使用的**结构化数据** (structured data)
- **特征编码阶段**：对结构化数据进行编码，得到对应的**稠密向量表示** (neural embeddings)
- **打分排序阶段**：对候选物品进行打分排序，得到要呈现给用户的排序列表 (recommended items)
- **推荐流程控制**：作为中央控制器，把控推荐系统的整体流程。也可以细化到对排序阶段的召回、粗排、精排的控制



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

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

■ 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

ChatGPT

Enhance news titles based on given information in the following format:

```
[title] {title}
[abstract] {abstract}
[category] {category}
```

You should rephrase the title to be clear, complete, objective, and neutral. Only provide the new title in the following format:

```
[newtitle] {newtitle}
```

[title] Here's Exactly When To Cook Every Dish For Thanksgiving Dinner
[abstract] Time out turkey day down to the minute.
[category] foodanddrink

[newtitle] Perfectly Timed Thanksgiving Dinner: A Comprehensive Guide to Cooking Every Dish with Minute-by-Minute precision

(a) News summarizer

ChatGPT

Describe user profile based on browsed news list in the following format:

```
(1) {title}
(2) {title}...
```

You should describe the related topics and regions in the following format:

```
[topics]
- {topic1} ...
[region]
- {region1} ...
```

1) 'Go back to work': Outcry over deaths on Amazon's warehouse floor
2) Four flight attendants were arrested in Miami's airport
3) America's cheapest cities where everyone wants to to live right now

Topics:
- travel - economy
- business - labor rights

Regions:
- Florida

(b) User profiler

ChatGPT

Generate a news article based on user history list in the following format:

```
(1) {{category}} {title}
(2) {{category}} {title}...
```

Provide one news article, which should be diverse to the original news list, in the following format:

```
[title] {title}
[abstract] {abstract}
[category] {category}
```

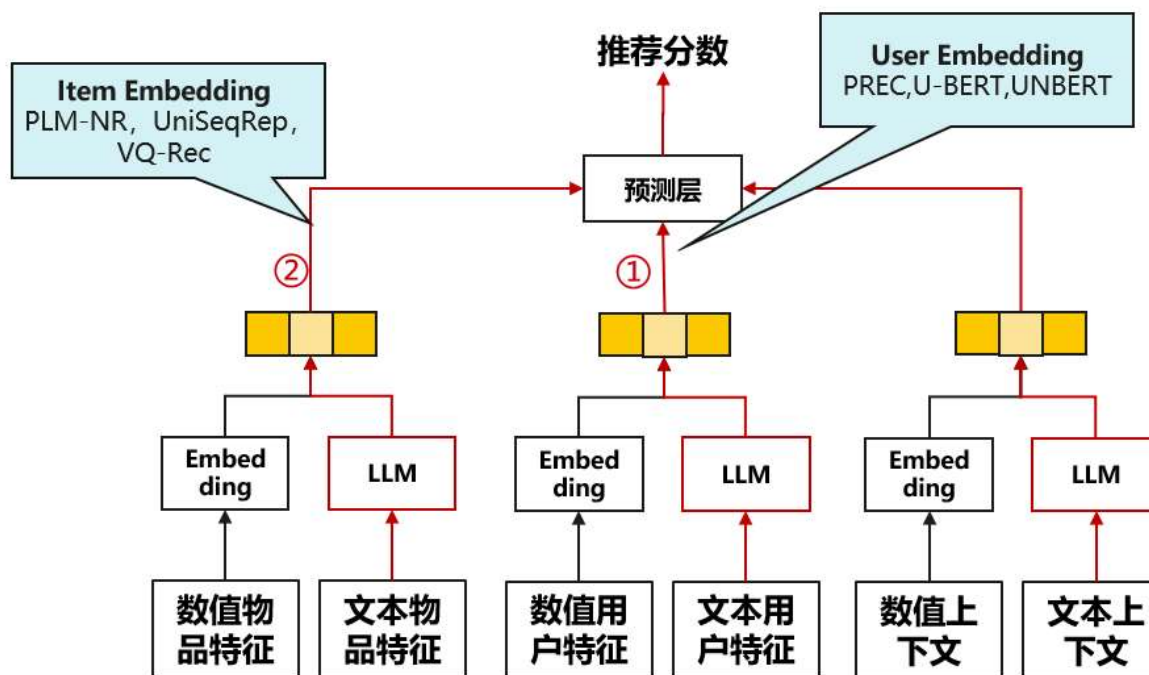
1) (tv) 'The Little Mermaid Live!': TV Review

[title] Live-action 'Mulan' hits theaters in March
[abstract] Disney's much anticipated live-action remake of 'Mulan' hits theaters in March, promising a new take on the beloved classic with stunning visuals and a talented cast.
[category] movie

(c) Personalized news generator

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

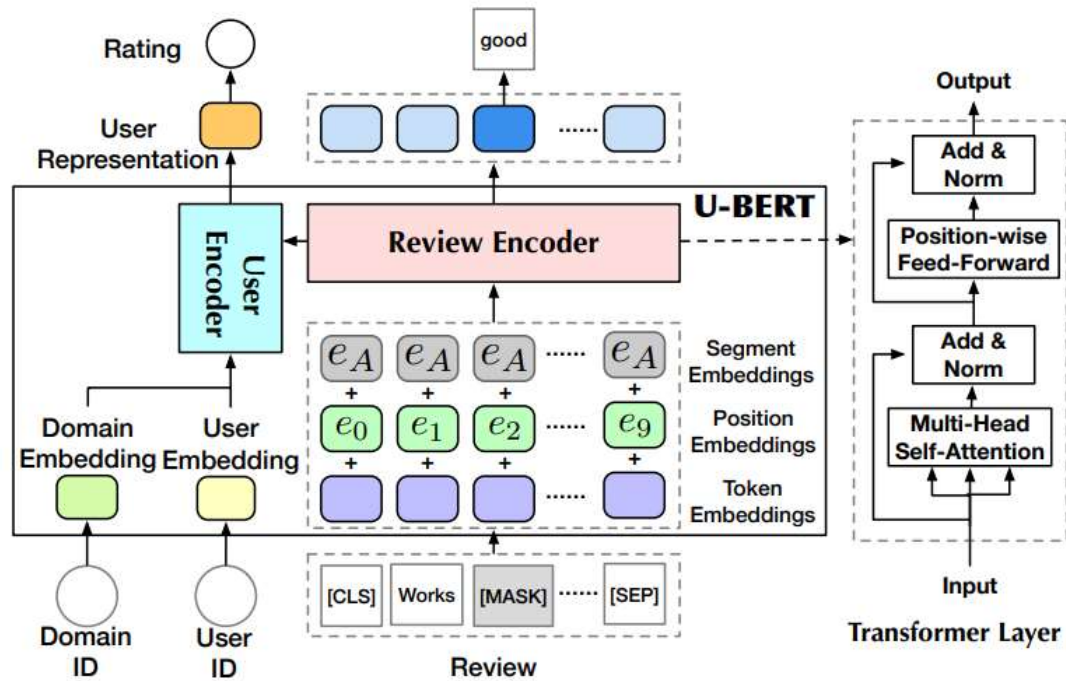


大语言模型在特征编码中的应用

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



以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/898005122045006033>