InfoQ 极客传媒

大模型在荣耀推荐和广告场景的应用实践

演讲人：冯晓东

荣耀终端股份有限公司 / AI算法专家

AiCon

全球人工智能开发与应用大会

AiCon
全球人工智能开发与应用大会

目录

极客邦科技 2026 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

InfoQ 极客传媒

荣耀终端股份有限公司介绍

HONOR

全球领先的AI终端生态公司

14000+	70%+	11.5%	300+	52000+	2.5亿+
员工	研发人员占比	研发投入营收占比	专利月申请量	体验店与专区专柜	在网设备数

注：数据截至2023年12月

荣耀公司新定位

新愿景

以人为本，科技与人文结合，最大程度释放人类潜能

新定位

从智能手机制造商转型
成为全球领先的AI终端生态公司

新战略

第一步 打造智能手机
第二步 构筑智慧生态
第三步 拥抱智慧世界

01

荣耀的推荐场景介绍

搜索、推荐、广告、游戏

荣耀的推荐场景

搜索

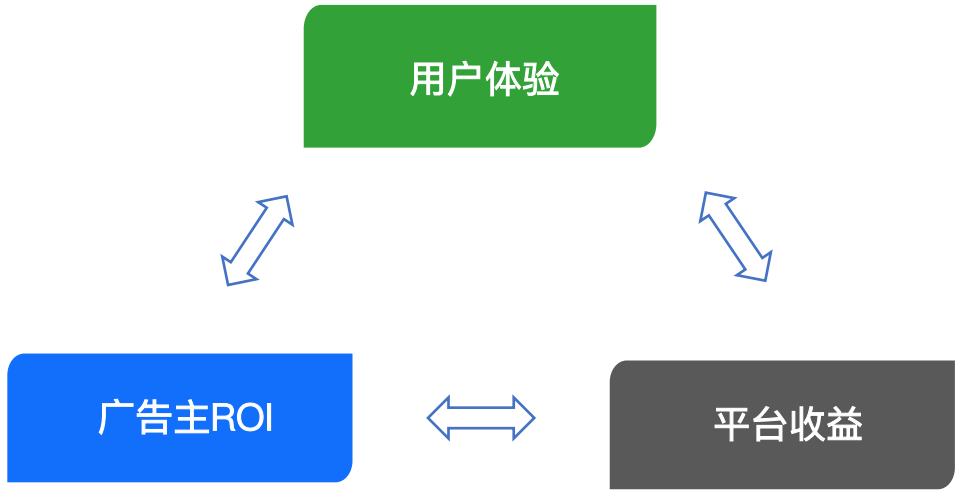
推荐

广告

游戏

荣耀的推荐场景

广告业务追求三角的平衡，对算法提出更高要求



多模态内容理解

- 内容侧：图文和视频内容的特征挖掘
- 用户侧：刻画用户对不同类型营销素材的偏好

更精准的排序模型

- 推荐：精准的顺序保证用户体验和分发效率
- 广告：精准的预估分保证流量的收益

20+应用30+场景：**多场景**　　单场景10+目标：**多目标**　　亿级活跃用户：**海量样本**　　用户日均10+行为：**超长行为序列**

02

推荐算法模型和大模型的特点分析

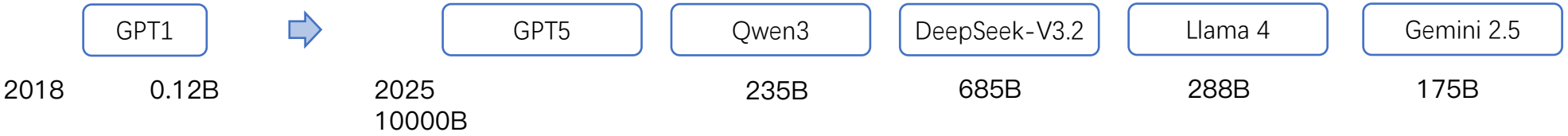
推荐算法结构演进及特点
大语言模型的演进及优势
大模型与推荐的结合方式

推荐算法结构演进及特点

线下天花板：大量特征挖掘和样本归因

线上天花板：50ms时延限制计算量和模型大小

大语言模型的演进及优势



预训练能力

多模态数据（文本、图像、音频、视频）

最高 1,000,000 Token 上下文长度

覆盖更多领域和长尾知识

序列处理能力

Self-Attention

Flash Attention

DeepSeek Sparse Attention

训练推理能力

MOE

KV Cache

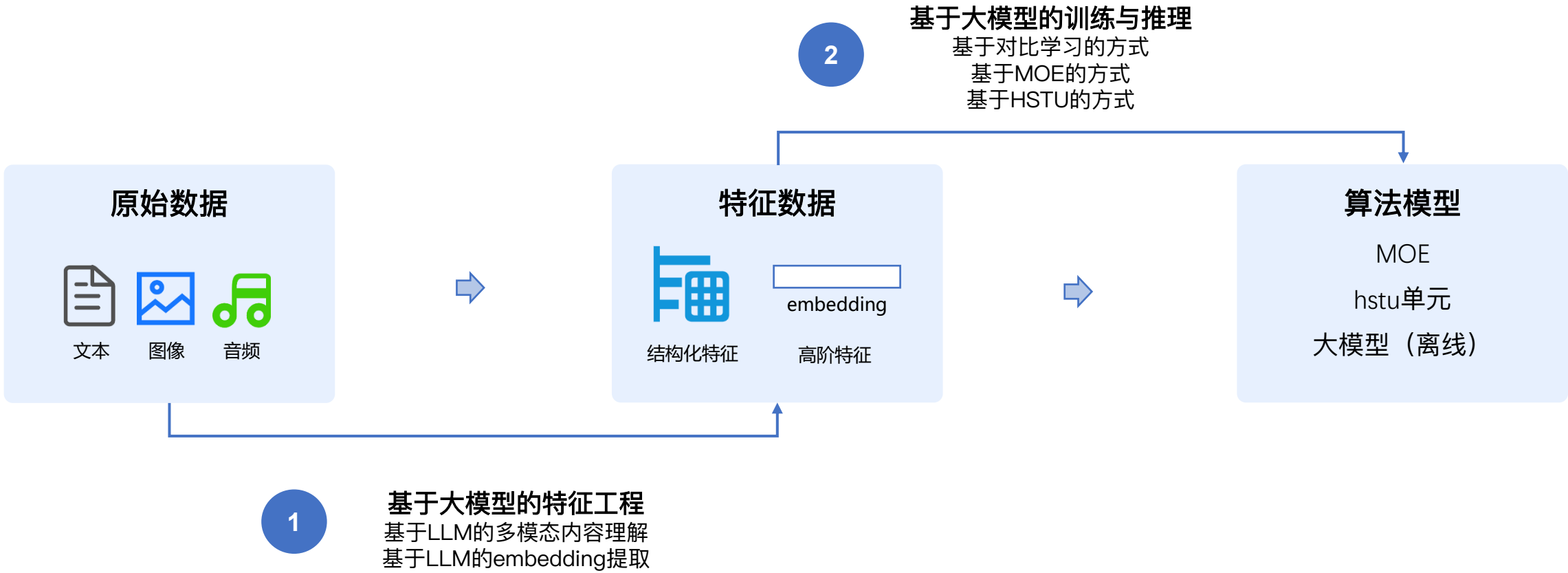
强化学习

特征工程和冷启动

用户超长/全生命周期序列

多场景多任务

大模型与推荐的结合方式



03

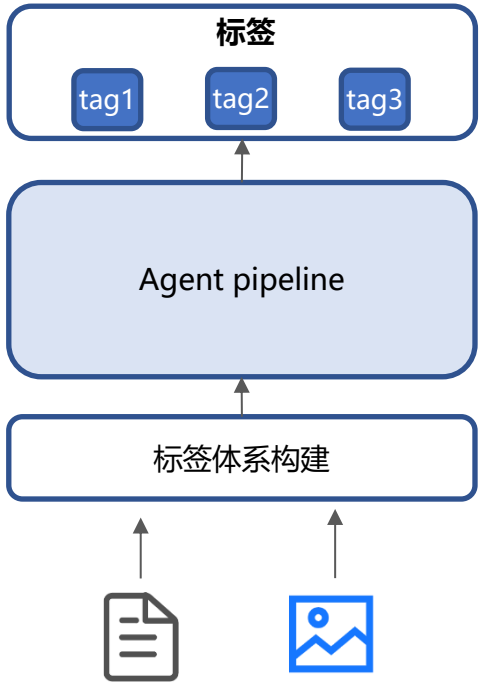
基于大模型的特征工程

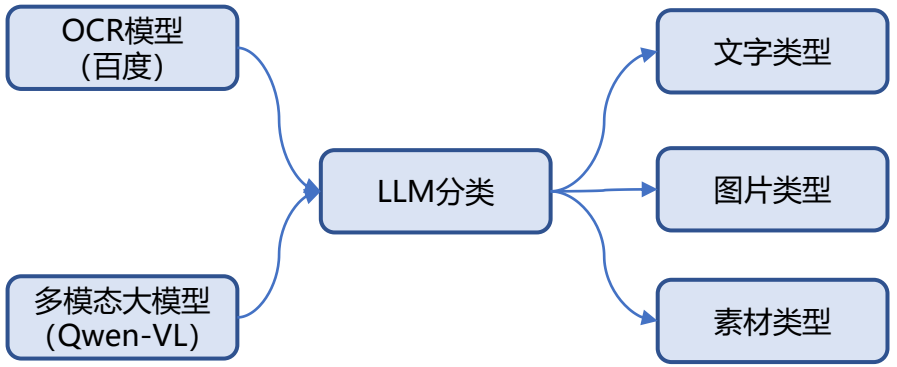
基于LLM的多模态内容理解

基于LLM的embedding提取

基于大模型的特征工程

基于LLM的多模态内容理解





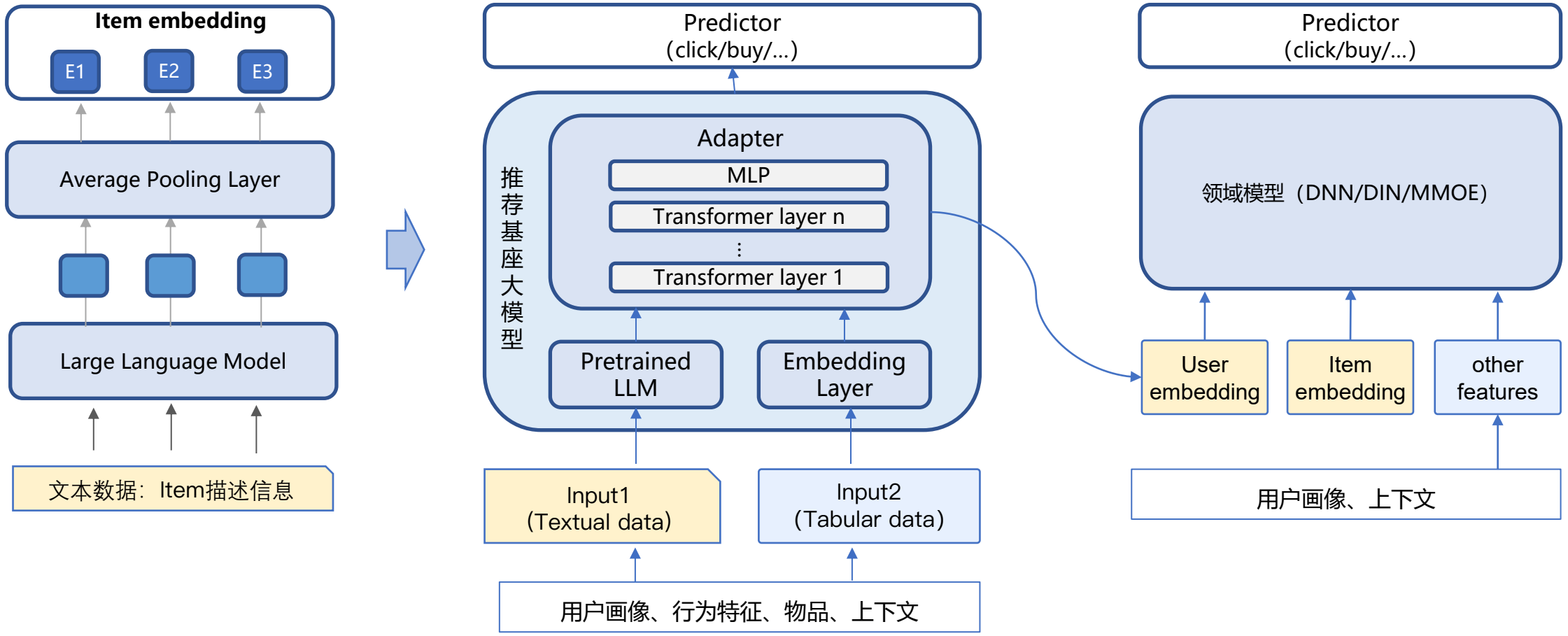
- Prompt设计：限定标签体系和步骤
- 流程编排：利用不同预训练模型的优势

一层分类准确率	二层分类准确率	三层分类准确率
95%	87.5%	90.7%

字段名称	注释
creative_id	创意ID
screen_orientation	素材展示屏幕方向
duration	视频类型资产的时长信息
ratio	素材长宽比例
color	色系
clarity	清晰度
brigthness	明暗度
tone	色温
person	人物
text	原始文本
text_length	文本长度
text_key_words	文本关键词
text_color	文字颜色
text_type	广告素材类型
object_type	物体类型

基于大模型的特征工程

基于LLM的embedding特征提取



基于大模型的特征工程

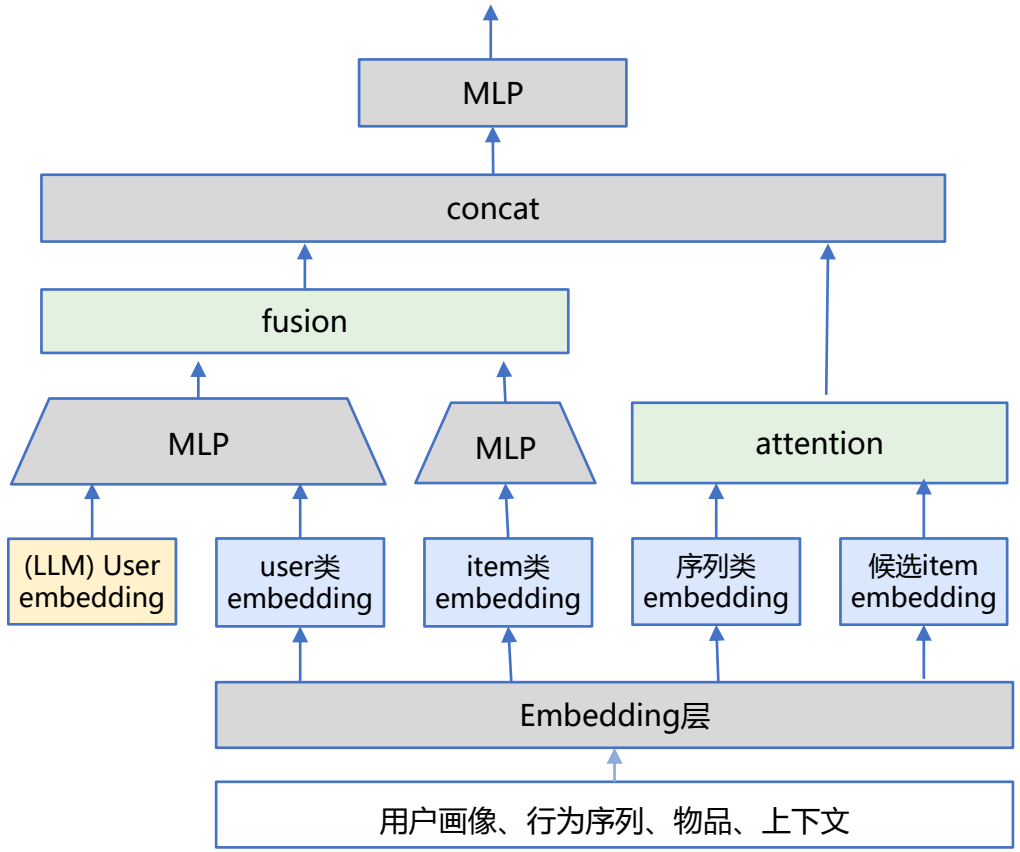
基于LLM的embedding特征提取—广告场景应用实践

文本特征构建

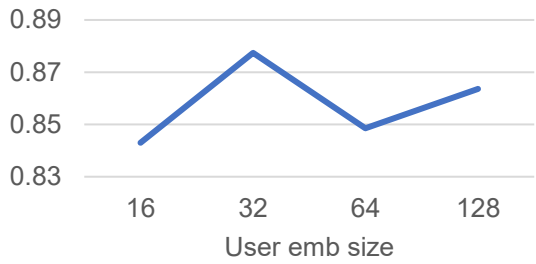
candidate_i_app_package_name	u_candidate_item_time	age	gender	u_item_download_name_seq	u_item_download_time_seq
途游斗地主（比赛版	28:20.2	44	0	开心消消乐,开心消消乐,开心消消乐,开心消消乐,开心消消乐	2024-08-15 21:36
蛋仔派对	04:13.0	44	0	蛋仔派对,火影忍者,巅峰坦克,创世战车,极品飞车：集结	2024-08-19 10:01
三国志·战略版	47:38.1	23	1	第五人格,第五人格,天姬变,第五人格-新赛季预约	2024-07-26 12:06
蛋仔派对	20:53.2	44	1	和平精英,和平精英,第五人格,和平精英,使命召唤手游,使命召唤手游,欢乐钓鱼大师,王者荣耀,地	2024-08-20 20:04

该用户最近的APP点击序列为地铁跑酷、三国志·战略版，APP下载序列为和平精英、和平精英、王者荣耀、贪吃蛇大作战、地铁跑酷、蛋仔派对，其中王者荣耀的分类为动作射击-MOBA;三国志·战略版的分类为经营策略-战争策略;蛋仔派对的分类为休闲益智-休闲竞技;地铁跑酷的分类为动作射击-跑酷;贪吃蛇大作战的分类为休闲益智-IO;和平精英的分类为动作射击-吃鸡游戏。

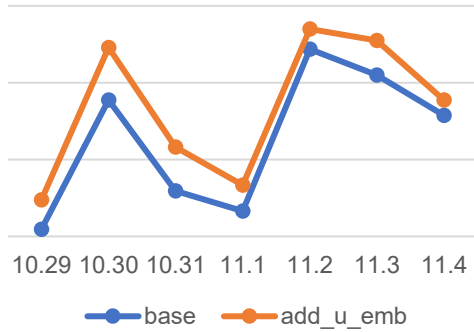
上线模型



实验效果



Emb size设为32，离线AUC最优



加user_emb特征后，DTR平均提升**0.85%**

04

基于大模型的训练与推理

基于对比学习的方式
基于MOE的方式
基于HSTU的方式

基于大模型的训练与推理

基于对比学习的方式

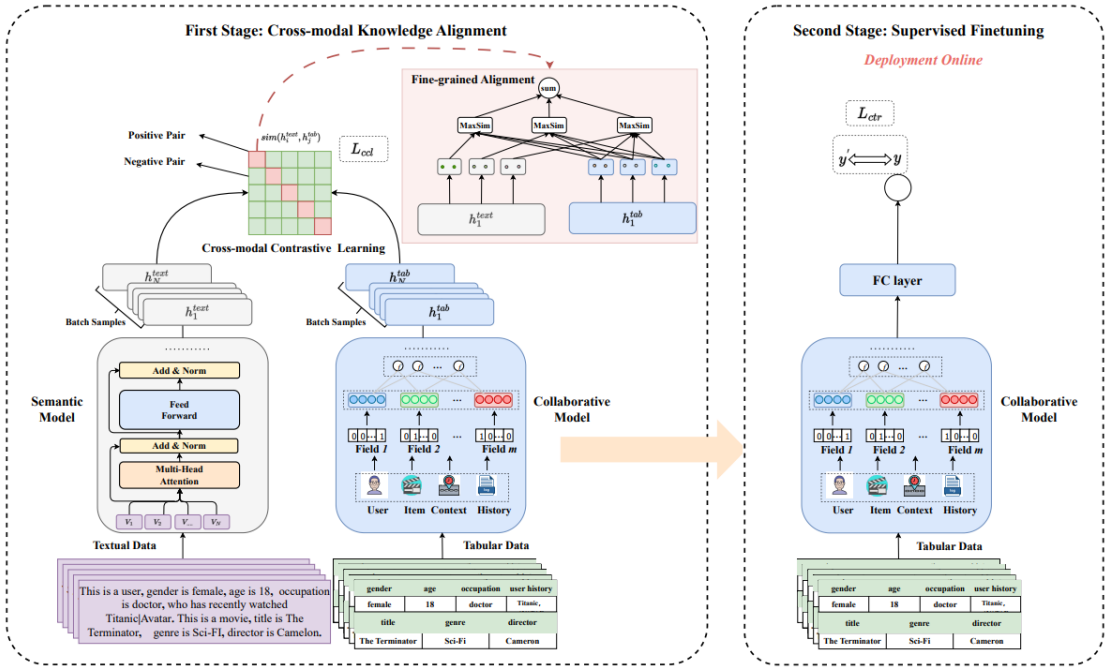


Figure 3: An intuitive illustration of the CTRL, which is a two-stage framework, in the first stage, cross-modal contrastive learning is used to fine-grained align knowledge of the two modalities. In the second stage, the lightweight collaborative model is fine-tuned on downstream tasks. The red square represents a positive pair in the batch, while the green square represents a negative pair.

Xiangyang Li, et al. (2023) CTRL: Connect Collaborative and Language Model for CTR Prediction.

实现的关键点

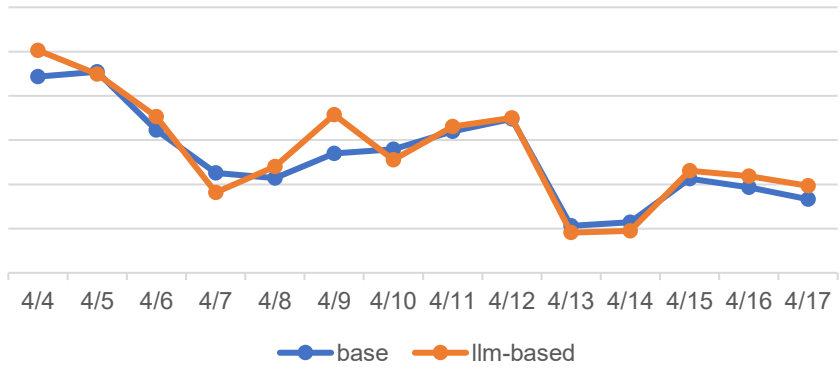
a. 语义信息如何编写？ --定制prompt模板

b. 选取什么LLM模型？ --冻结预训练LLM模型

c. 预训练阶段，损失函数如何设计？

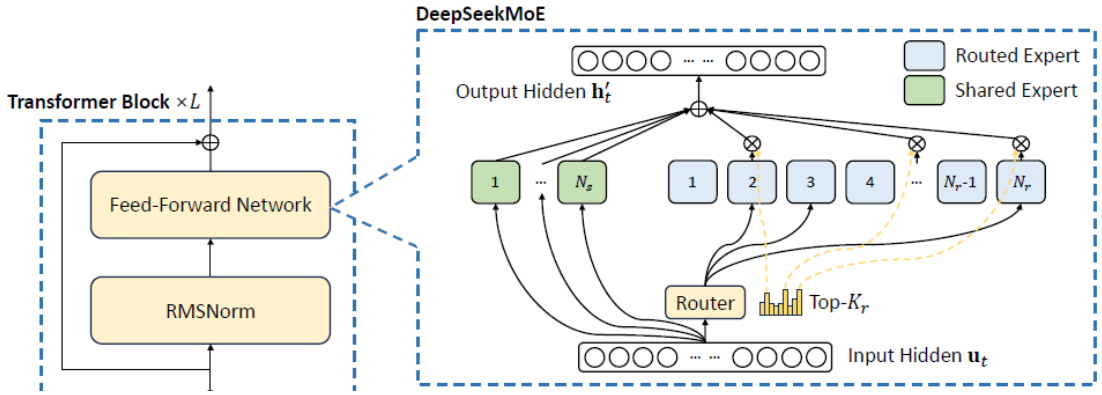
--$L^{textual2tabular}$ 与 $L^{tabular2textual}$ 具有对称性

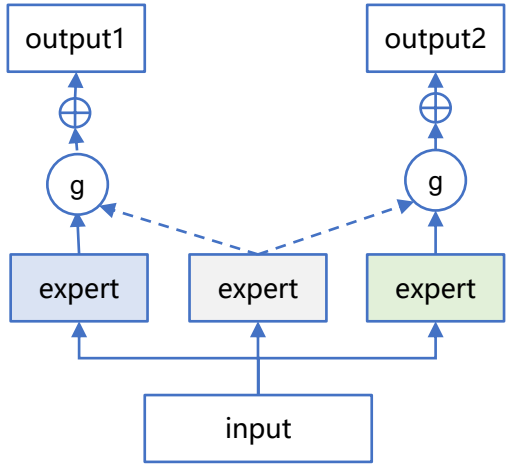
实验效果：14天平均提升1.07%

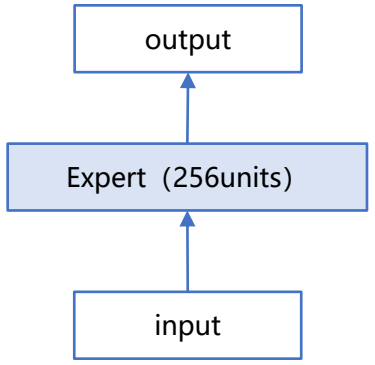


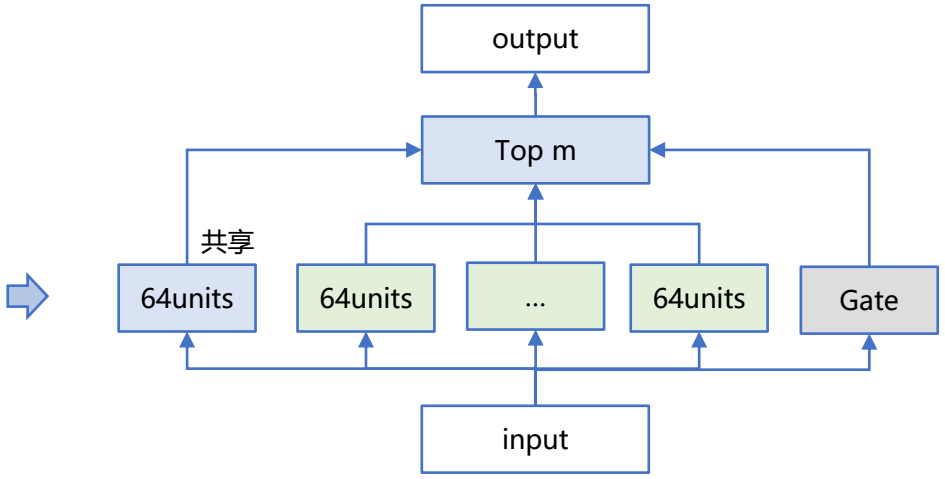
基于大模型的训练与推理

基于MOE的方式







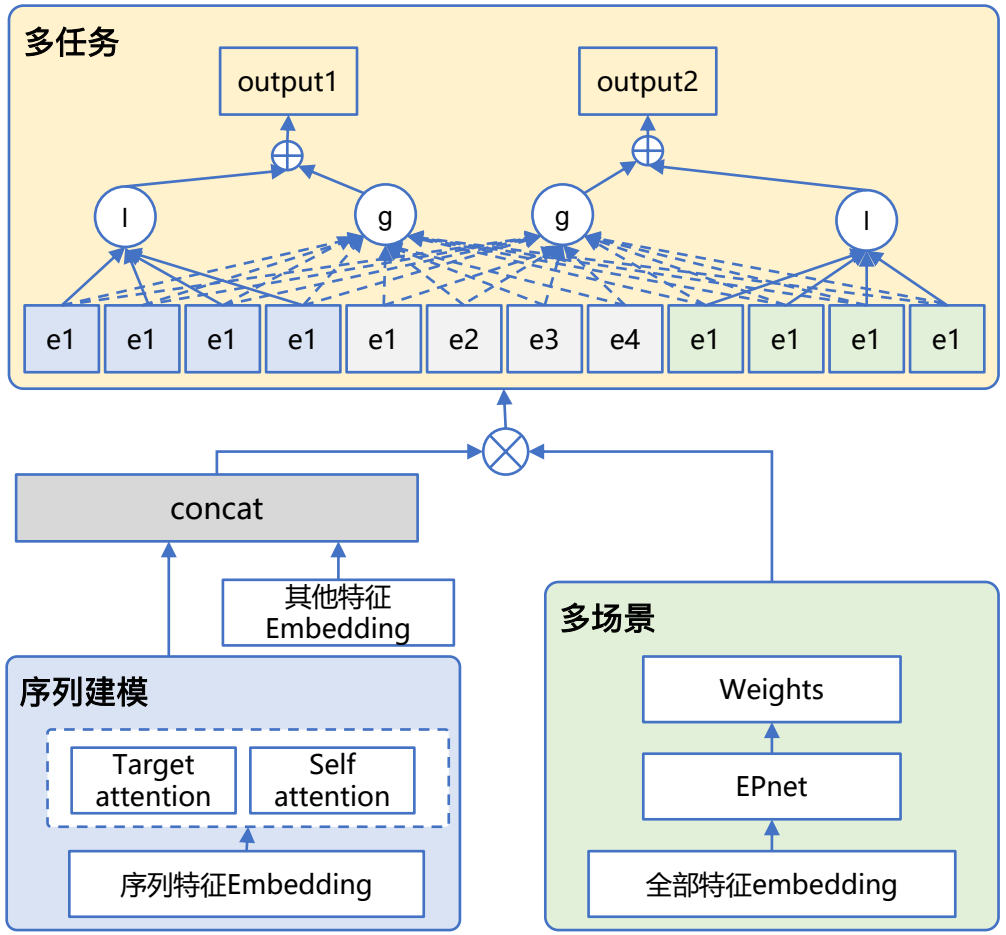


专家群结构关键点：

- 专家数量：数量太多会有负载均衡问题
- 路由方式：无路由与稀疏路由，共享路由与全局路由
- 稀疏路由：样本量小的目标拟合能力弱

基于大模型的训练与推理

基于MOE的方式—广告场景应用实践



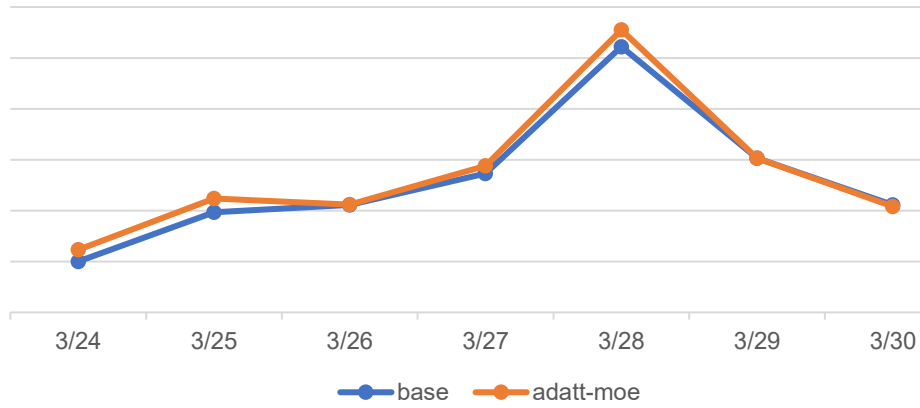
设计的关键点

a. 专家网络1拆4

b. 专家路由经过多轮实验效果不佳，采用无路由方式

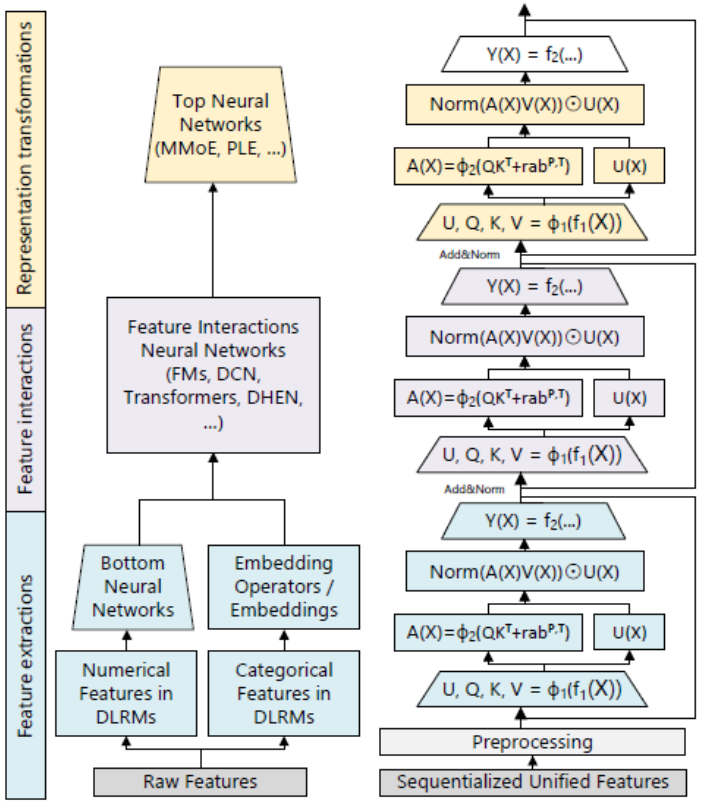
c. 为了专家计算复杂度，对特征重要度做筛选

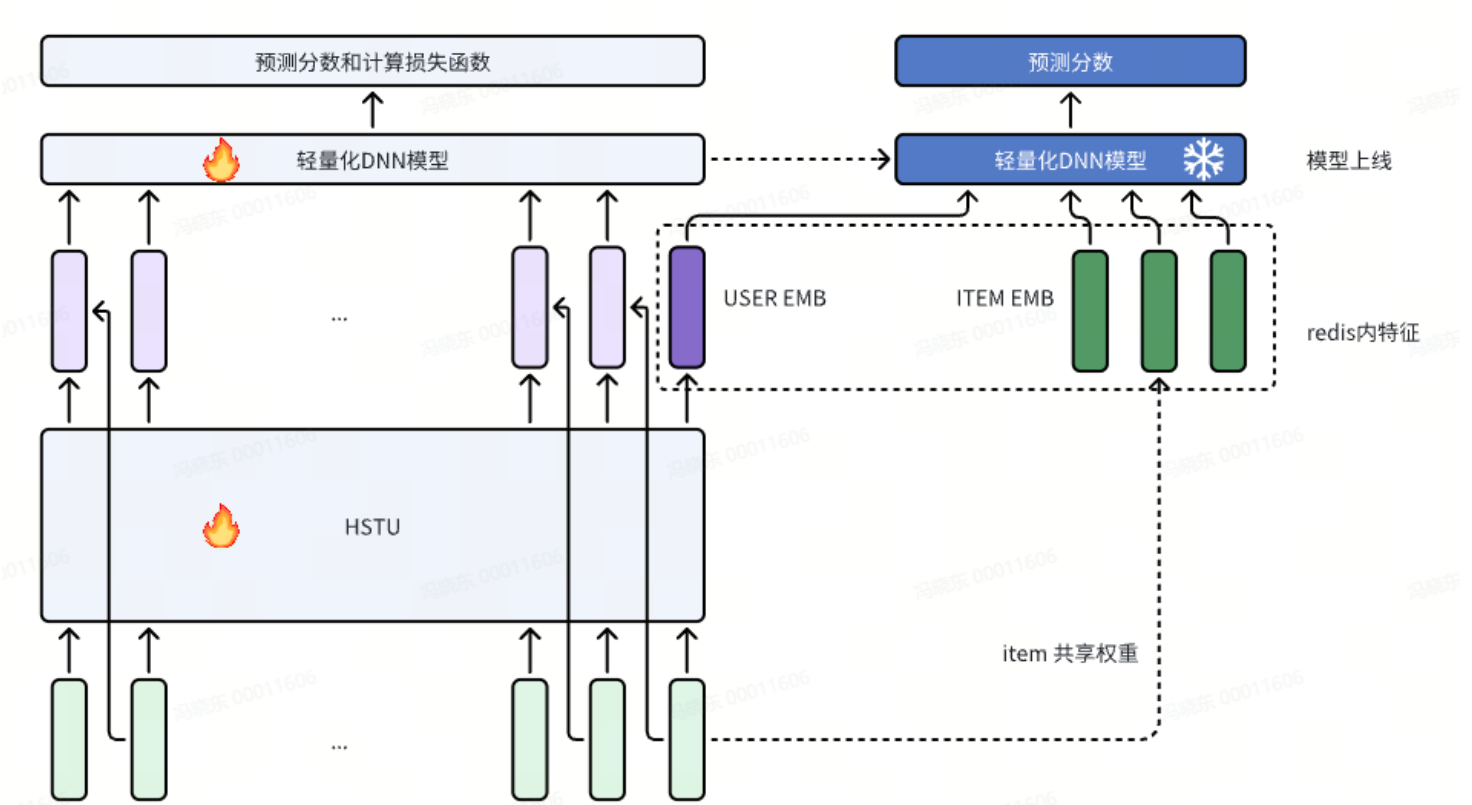
实验效果：单位收入平均提升0.63%



基于大模型的训练与推理

基于HSTU的方式



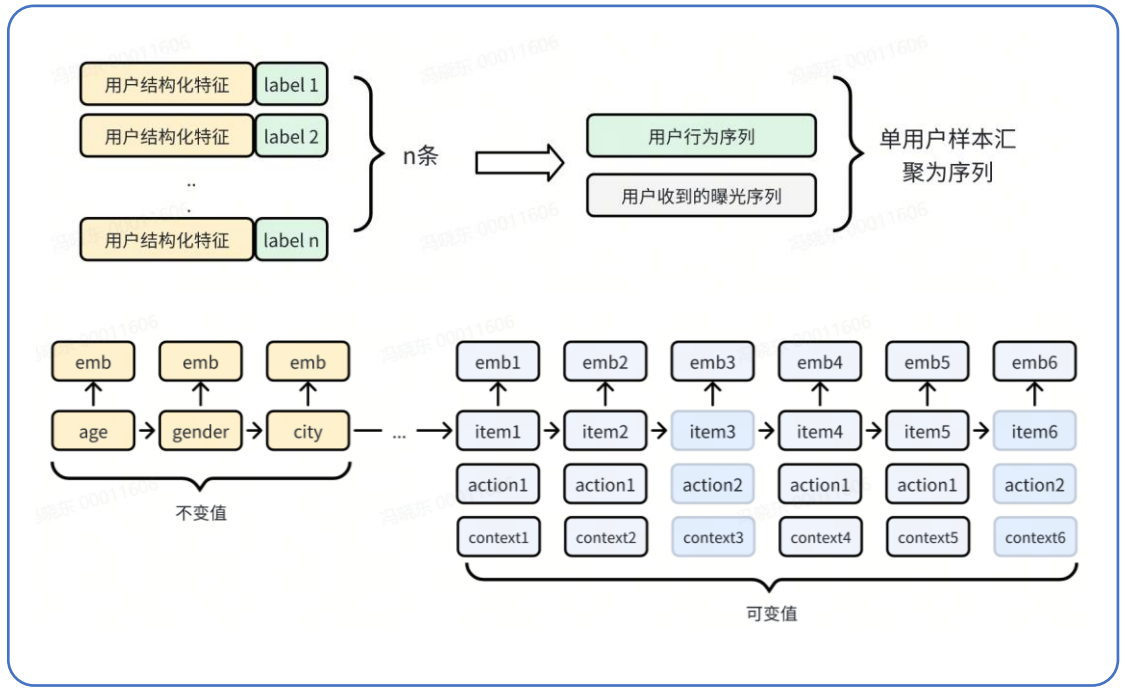


Jiaqi Zhai, et al. (2024) Actions Speak Loudera than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations.

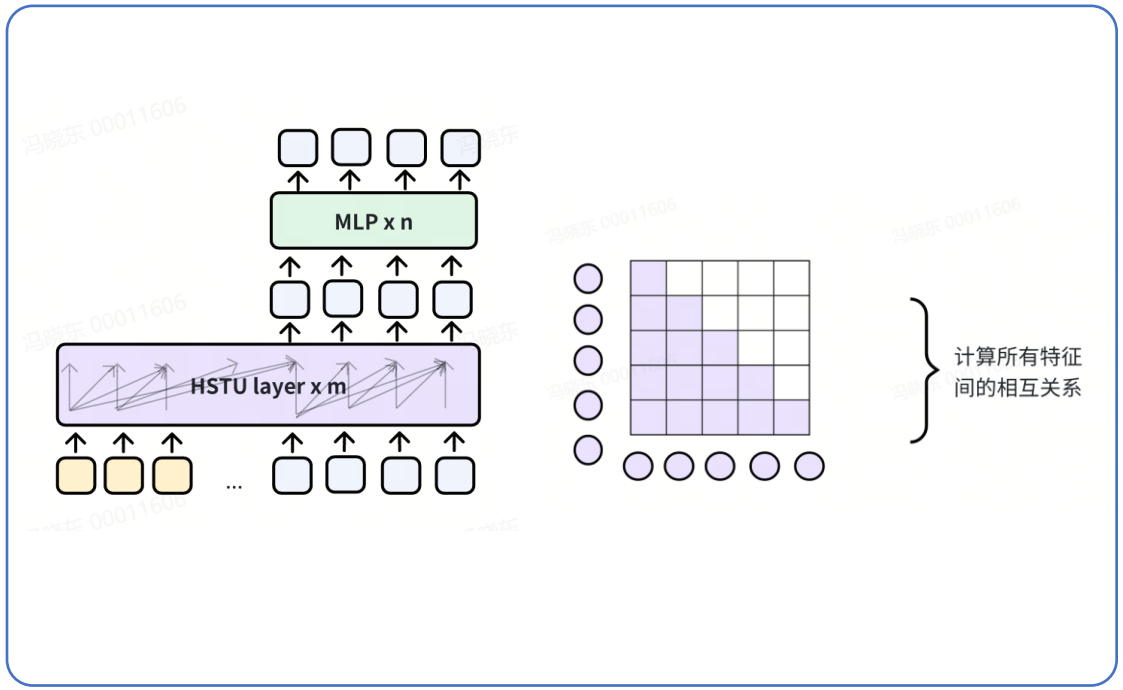
基于大模型的训练与推理

基于HSTU的方式

（1）**特征构建**：按照用户粒度构建特征和样本，减少数据重复，保持更多用户信息，保留重要画像和上下文特征



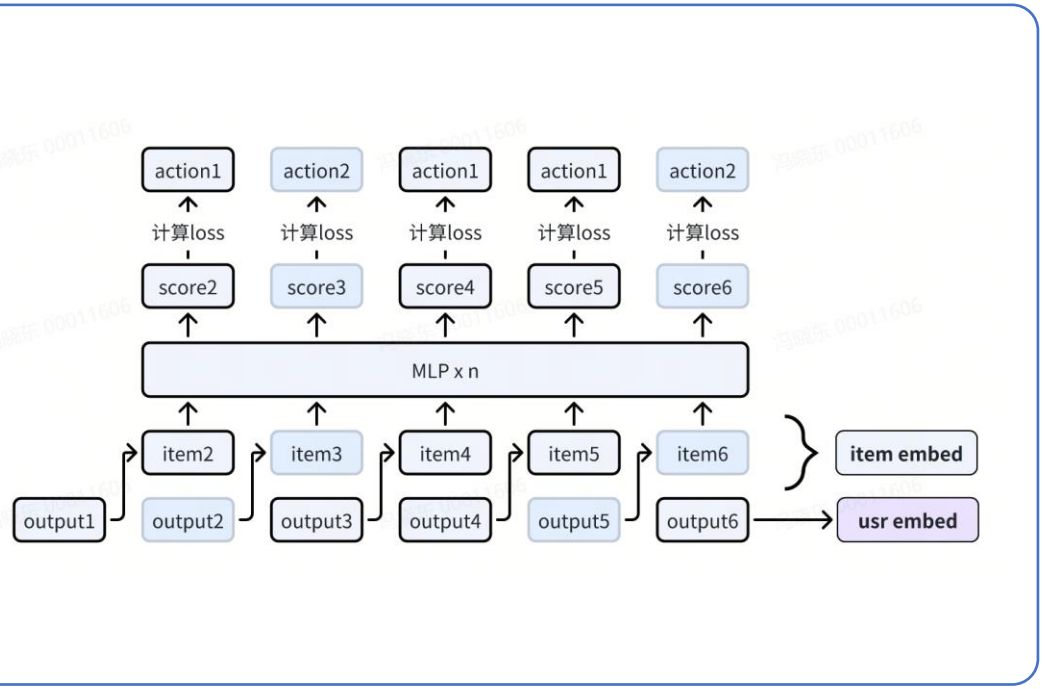
（2）**模型结构**：HSTU+ DNN，在传统transformer基础上增加行为间的相对时间关系和对物料embedding的门控机制



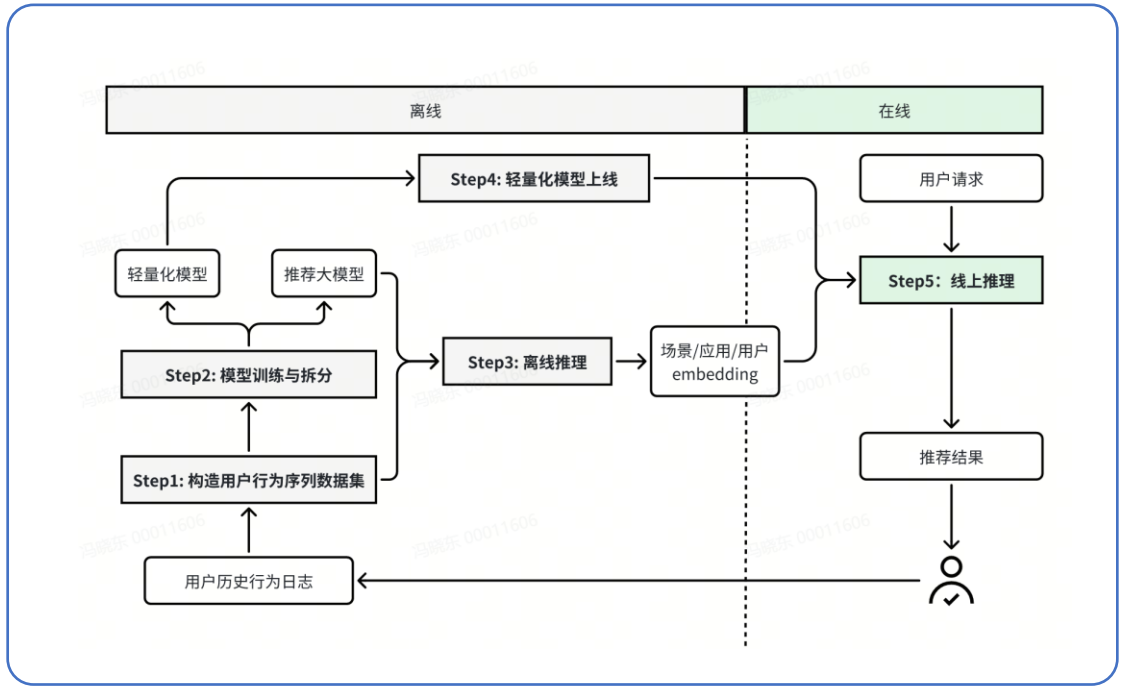
基于大模型的训练与推理

基于HSTU的方式

（3）**训练策略**：在用户的每个曝光行为后，都基于该时刻之前的用户特征推理出用户向量embedding，用于预测下一个行为



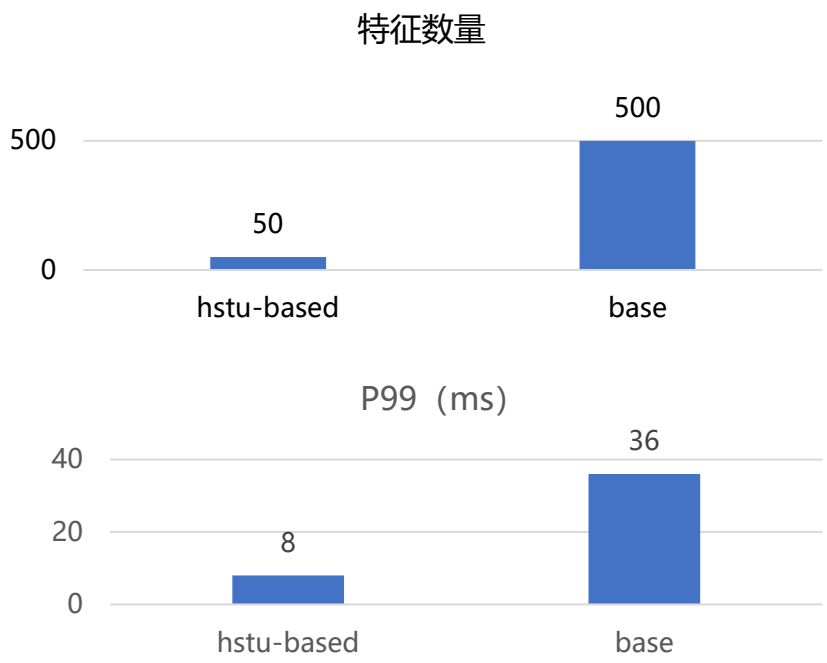
（4）**线上推理**：大模型参数量多具备scaling law，拆分轻量化模型结构上线，推理性能更佳



基于大模型的训练与推理

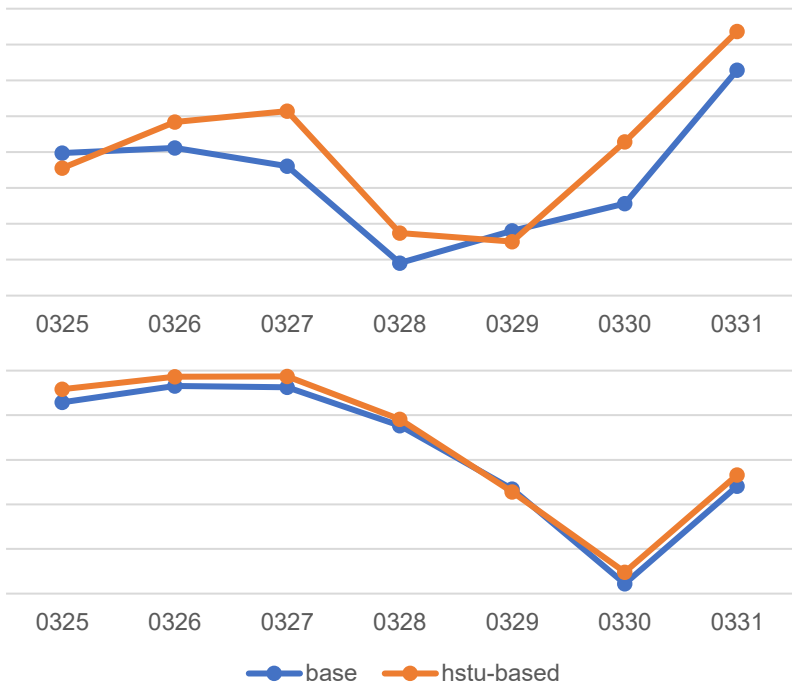
基于HSTU的方式—广告场景应用实践

模型性能



Base特征数量500+，hstu-based特征50+，性能极大提升

业务效果



CTR平均提升**0.98%**，单位收入平均提升**0.72%**

05

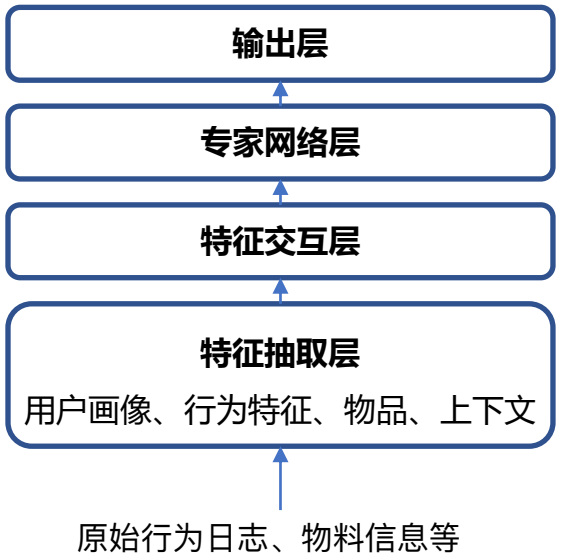
总结与展望

快速落地推荐大模型

生成式推荐覆盖推荐全链路

总结：快速落地推荐大模型

传统DLRM → 预训练LLM + 微调 + 传统DLRM → 生成式推荐 + 传统DLRM → 生成式推荐



基于LLM的DLRM学习

- 基于对比学习的方式
- 基于模型蒸馏的方式

多模态内容理解

- 基于LLM的标签特征提取
- 基于LLM的embedding特征提取

基于长序列的生成式推荐

- 用户全生命周期序列抽取
- 基于hstu的序列模型构建
- 基于hstu+moe的多任务模型构建

生成式推荐

- 多模态
- Hstu
- Moe
- 强化学习

展望：生成式推荐覆盖推荐全链路

原始数据

文本　图像　音频

特征工程不断精简，甚至消失

模型具备scalling law，同时满足训练推理性能

算法模型

MOE

hstu单元

大模型（离线）

极客邦科技 2026 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

InfoQ 极客传媒

InfoQ 极客传媒

THANKS

探索 AI 应用边界

Explore the limits of AI applications

AiCon

全球人工智能开发与应用大会