京东广告大模型技术探索与新型模型体系建设实践

演讲人：张泽华
京东 / 算法总监

AiCon
全球人工智能开发与应用大会

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

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

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

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

4月 — 6月 — 8月 — 10月 — 12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

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

InfoQ 极客传媒

AiCon
全球人工智能开发与应用大会

目录

01

广告营销领域大模型的近期发展和挑战

广告营销领域大模型近期发展

技术范式：从模块化堆叠到生成式重构

传统CTR模型瓶颈

传统CTR预估模型因算术强度低、计算流碎片化，无法充分利用新一代GPU的算力，MFU仅个位数，难以满足广告营销对高效算力的需求。

生成式预估范式转变

2024年起，头部企业转向生成式预估范式，通过Transformer架构将用户行为和商品特征转化为统一Token序列，以GEMM主导计算流，使MFU提升至30%以上，实现算力效率质变。

三阶段范式

行业演进出PreTrain–PostTrain–Application三阶段范式，对齐内容空间与兴趣空间，先构建图文内容基础表征，再通过图搜召回率等中间指标实现内容空间与点击兴趣空间的对齐，最后完成CTR预估。

广告营销领域大模型近期发展

应用落地：AIGC 驱动营销全链路效率革命

1 AIGC渗透现状

AIGC已渗透53.1%广告主的创意流程，视频制作超半数环节由AI完成，成为广告创意生产的重要力量。

2 对营销的意义

生成式模型端到端接管创意全链路，使“小预算快迭代”成为主流运营节奏，为个性化推荐提供低成本创意。

对话式需求拆解

大模型以多轮对话将模糊需求拆成可计算属性，如“轻便跑步鞋”被解析为“运动场景+重量<300g+防滑”，突破了静态标签局限。

兴趣漂移追踪

通过时序建模捕捉兴趣漂移，从“科技政策”迁移至“AI伦理”，使推荐结果随用户意图演变而动态调整，提升了推荐精准度。

广告营销领域大模型近期发展

工程化趋势：基础设施与 Scaling Law 深度绑定

推理延迟问题

生成式模型自回归特性带来推理延迟，与广告投放毫秒级要求冲突，成为工程化落地的关键挑战。

技术优化手段

行业通过Flash Attention、定制CUDA Kernel等技术，将首Token延迟压缩至10ms内，支撑线上高并发。

双扩张阶段

多模态广告模型进入参数—数据双扩张阶段，通过持续增加算力、参数与数据，CTR提升仍呈稳定幂律关系。

Scaling Law验证

验证了Scaling Law在广告场景的有效性，但需配套工程优化以避免边际收益递减，为后续推理效率议题埋下伏笔。

广告营销领域大模型近期发展

合规与伦理：平衡创新与风险

典型错误案例

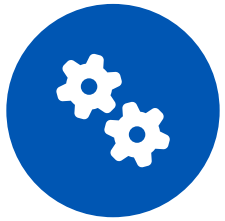
某广告曾出现“花生树上结果”一类常识错误，暴露生成模型缺乏世界知识的问题，损害品牌信任并可能触发合规处罚。

内容治理措施

行业正建立“AI生成+人工精修”协作流程，把创意策划、事实校验、法律审查设为必过节点，以提升内容质量。

过去一年我们又面临了哪些新挑战

从判别到生成：推荐范式拐点已至



传统判别式

多塔预估

Scaling-Law驱动范式转移

生成式推荐

序列生成

过去一年我们又面临了哪些新挑战

广告业务Token化，常识与幻觉

业务Token化

自改进Token与LLM内部表征对齐，Recall两位数提升。

幻觉抑制

通过知识图谱外接校验，虚构卖点率大幅降低。

过去一年我们又面临了哪些新挑战

极低延迟与算力供给

实时性优化

各种Attention优化、动态剪枝、算力优化
端到端从秒级降至50ms，满足广告毫秒级竞价。

模型MFU提升至
32%

推理延迟降至
50ms

02

生成式大模型
助力广告算法代际突破

广告生成式大模型业务模式变化

从传统推荐系统升级为："洞察–决策–执行"三角闭环，让算法决策业务可追溯。

快手 OneRec
全量上线，统一多场景

百度 GRAB
生成式排序，突破瓶颈

京东专用大模型
聚焦广告，深度定制

商业化深水区：注意力售卖终结

01

生成式AI重构广告逻辑

生成式AI把广告逻辑从卖注意力改写为卖结果，传统CPM模式遭遇挑战。Perplexity在2024Q4广告收入仅2万美元，揭示流量即收入公式失效。

02

巨头探索困境

传统平台和各流量入口，均面临算力成本飙升与站外流量下滑双重压力，稀缺广告位再难提价，行业共识转向按成交、按效果分成。

通用大模型为何难啃广告ROI

1

知识缺口

通用LLM缺乏行业趋势、投放策略等垂类知识，无法直接解决ROI优化、人群定向等具体业务问题。

2

落地障碍

通用模型仅能给出正确但无用的建议，如提高创意质量，难以被广告主采纳，无法满足实时融合商品、用户行为与促销节奏的需求。

3

核心问题

面对新品冷启动、老客复购、大促冲量等不同优化目标，通用模型缺乏可执行的动作空间，难以满足广告主的多样化需求。

GRAM：召排一体生成式大模型

1

模型架构

京东GRAM模型基于Reasoning架构，实现意图识别、检索、排序、机制端到端融合，参数化+非参数化内存动态更新商品/用户知识。

2

业务价值

GRAM模型解决传统推荐召回率低、冷启动难痛点，电商场景CTR提升XX%，推理延迟降低至50ms。

3

技术优势

模型采用混合架构设计，平衡效率与创新，通过芯片级优化+分布式训推，实现大模型毫秒级实时服务。

多层次训调打造广告专用大模型

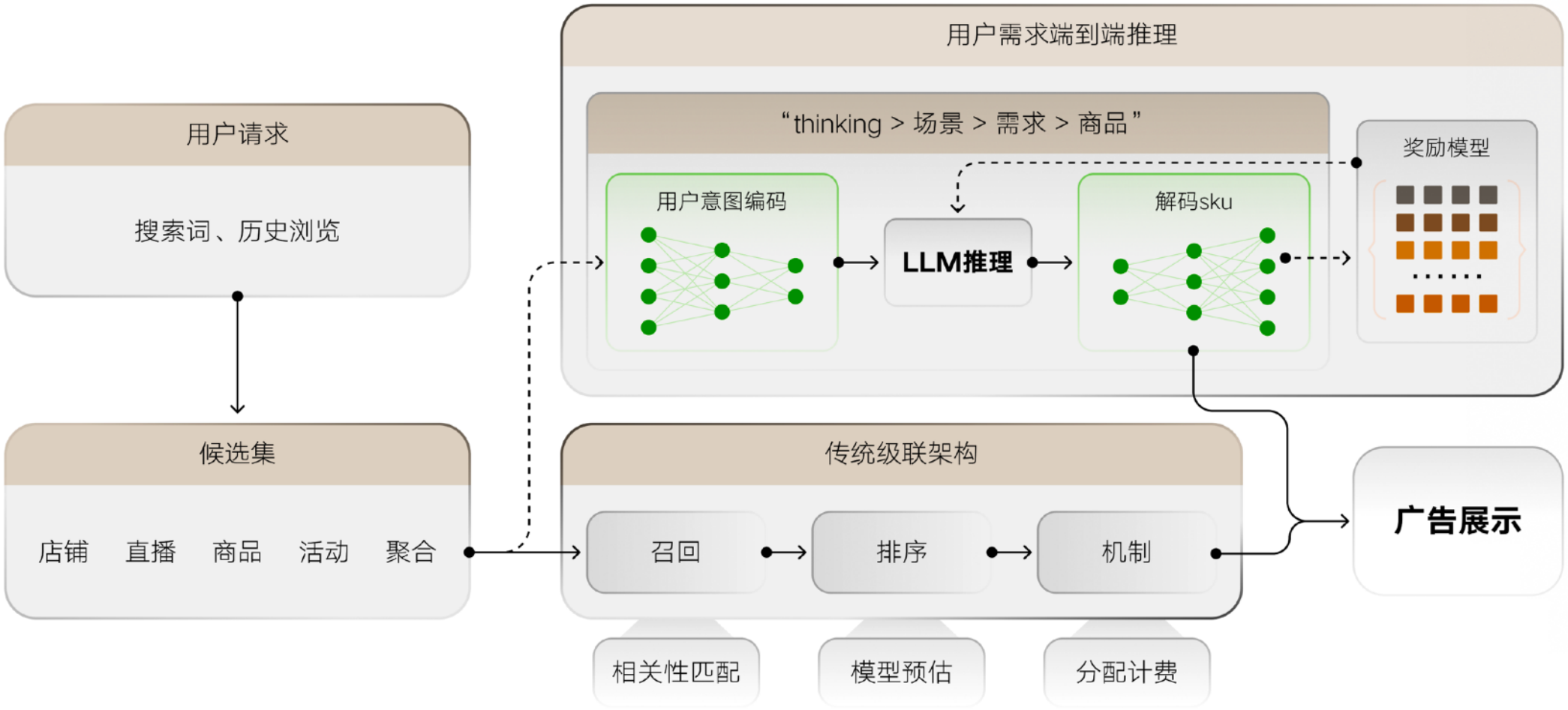
通用基座
保留基模泛化能力

行业知识
注入XX亿商品知识

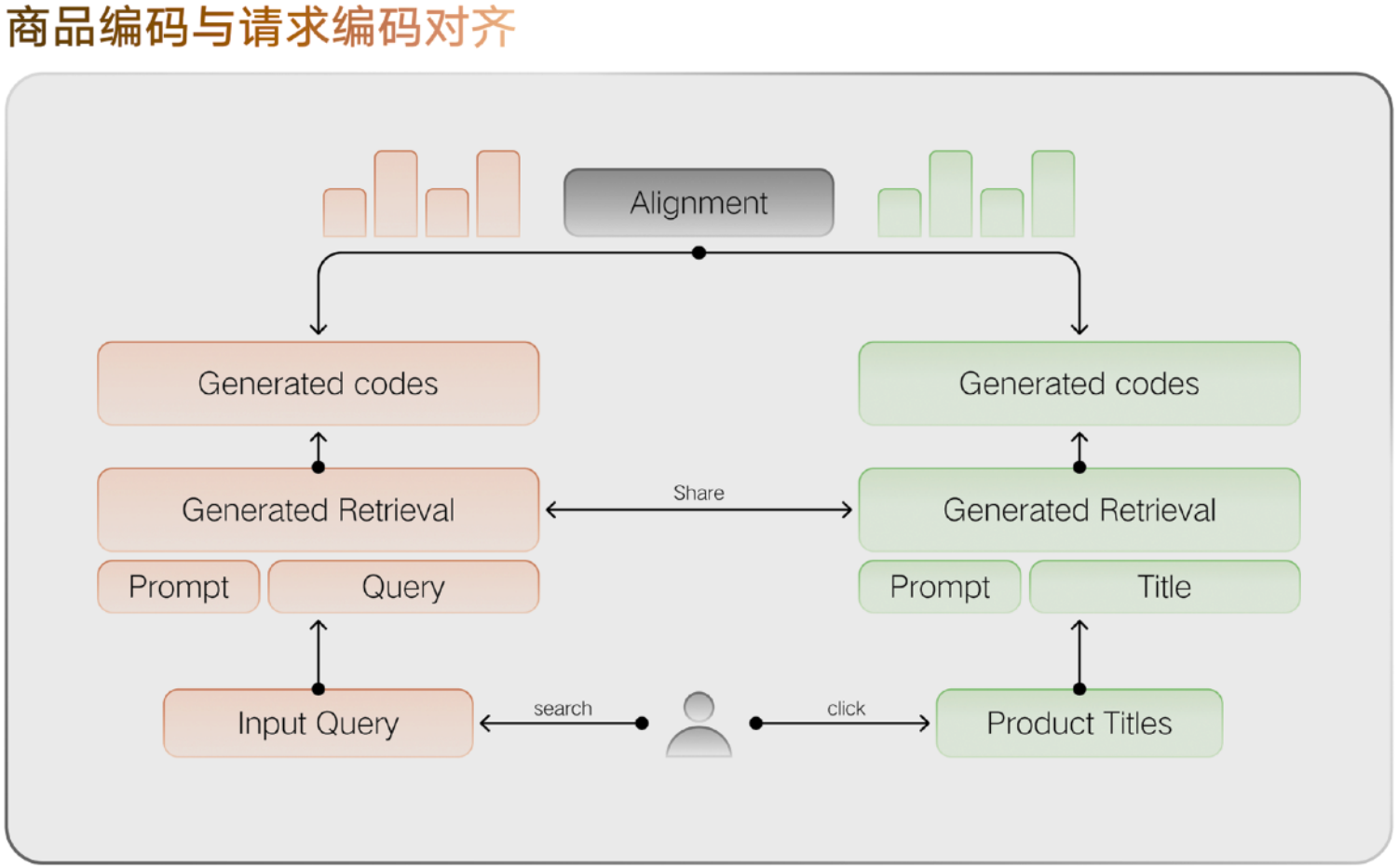
业务日志
10亿级广告日志对齐

推理成本	相关性	广告主ROI
降低 90%	**提升 XX%**	**+XX%**

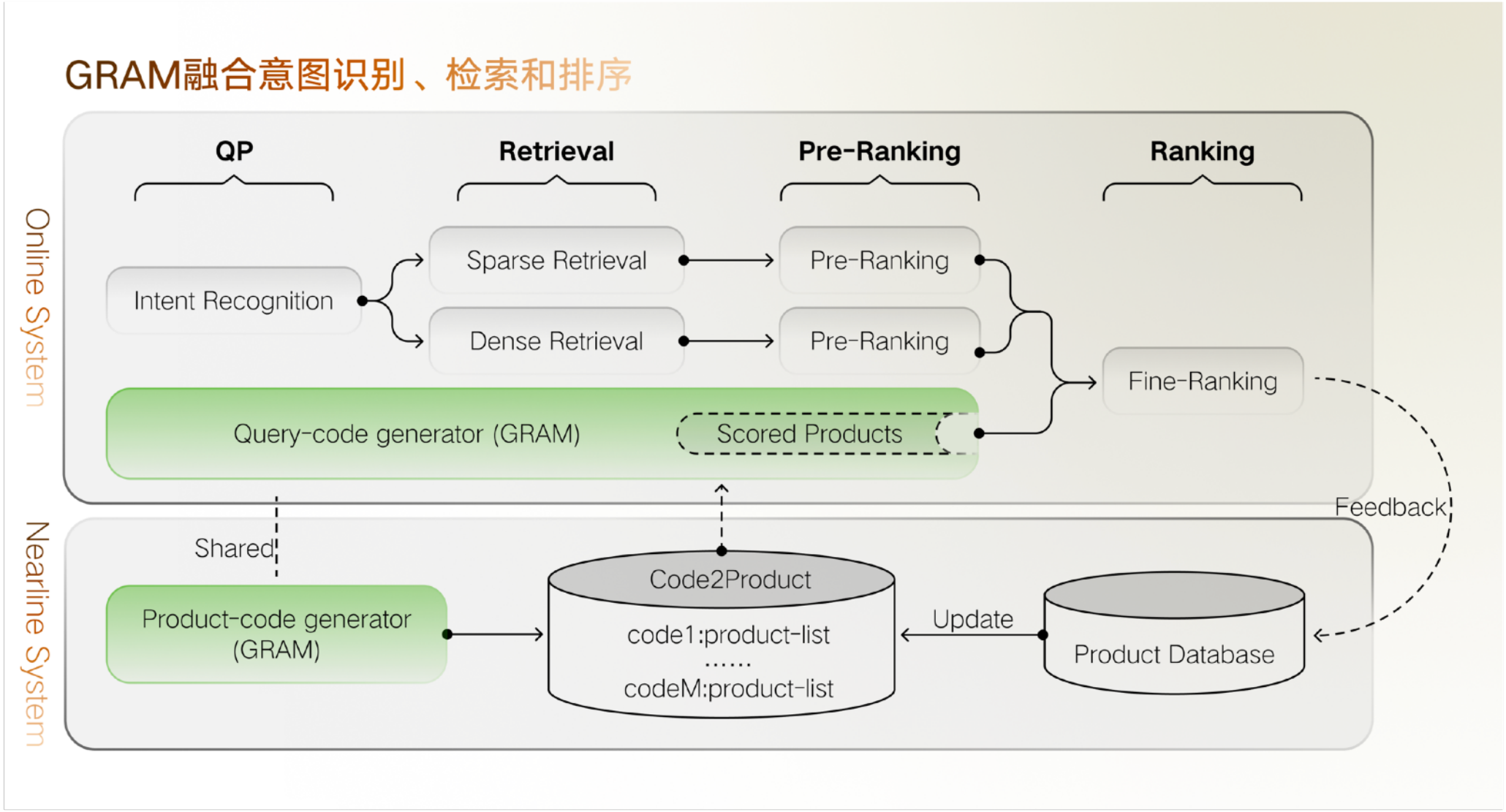
召排一体：LLM时代的广告代际突破



京东流量场中的实践：GRAM（Generative Retrieval and Alignment Model）



京东流量场中的实践：GRAM（Generative Retrieval and Alignment Model）



小样本与长尾：新用户新商品冷启动

1

问题描述

长尾商品与新用户行为稀疏，导致表征质量骤降，影响广告营销的效果和精准度。

2

解决方案

行业通过生成式内容先验、对比学习增强与元学习等技术，提高极少样本下的泛化能力。

3

平台责任

平台需构建通用商品知识库，以降低长尾获客成本，为中小企业应用提供技术抓手。

冷启动破解新品广告

1

新品属性编码

GRAM把新品属性、类目趋势、相似爆款经验实时编码到生成空间，无需等待累积点击数据。

冷启动效率提升

实测新品上架可快速获得首次曝光，冷启动期CTR较传统召回提升40%，为商家赢得新品流量窗口。

2

3

平台收益

平台缩短新品孵化周期，快速扩充可投货品池，提升整体广告投放效率和收益。

数据–推理–记忆 更深度理解

垂类数据处理

整合XX亿+商品数据、X亿+用户行为数据，形成可执行动作空间，为模型提供丰富的业务知识。

推理感知技术

在B端应用上，采用ReAct框架，让智能体思考+行动闭环，直接输出可执行方案，把AI从顾问升级为操盘手。

连续决策能力

通过长短期记忆管理，继承历史投放经验，实现大促爆款策略自动复用，提升投放效果。

03

生成式算法工程与知识工程实践

技术落地三道坎：实时、幻觉、门槛

1

实时性矛盾

电商广告场景要求毫秒级响应，生成式模型推理速度仅为判别式1/3 ~ 1/20，每100ms延迟可导致7%订单流失。

2

可靠性风险

模型幻觉带来的虚假宣传风险，使广告主合规顾虑陡增，平台需为每一句生成文案承担连带责任。

3

业务端新诉求

80%中小商家缺乏专业运营能力，传统投放操作复杂；多模态内容与投放决策的因果关联评估体系缺失。

技术深水区：常识推理与实时性平衡

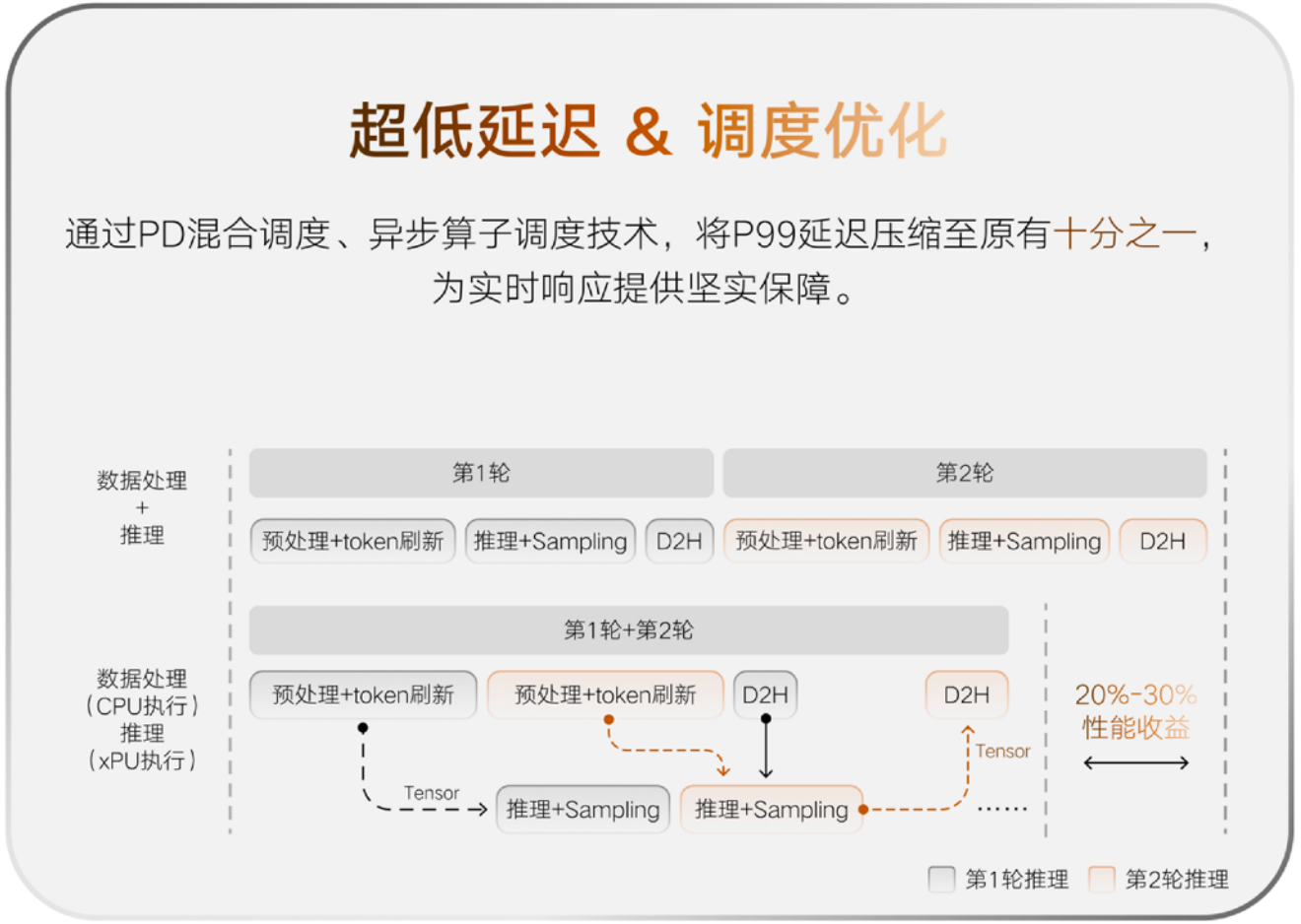
技术挑战

生成式模型仍会出现违反物理规律的商品描述，且自回归解码延迟难以压缩至毫秒以下，成为技术深水区。

前沿探索

前沿研究探索世界知识注入、投机解码与并行生成等技术，以同时提升生成内容的正确率与速度。

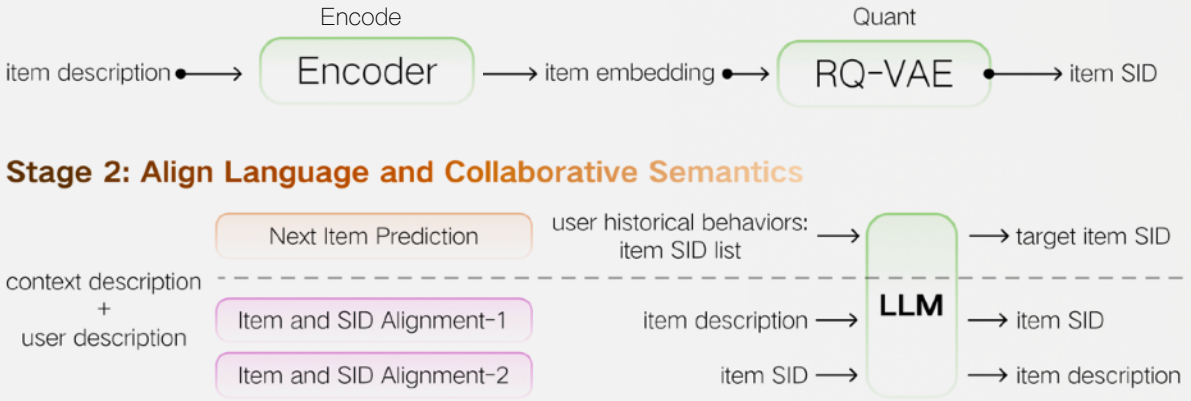
面向广告业务特性的超低延迟推理定制化解法



面向广告业务特性的超低延迟推理定制化解法

定制化模型 & 受限生成

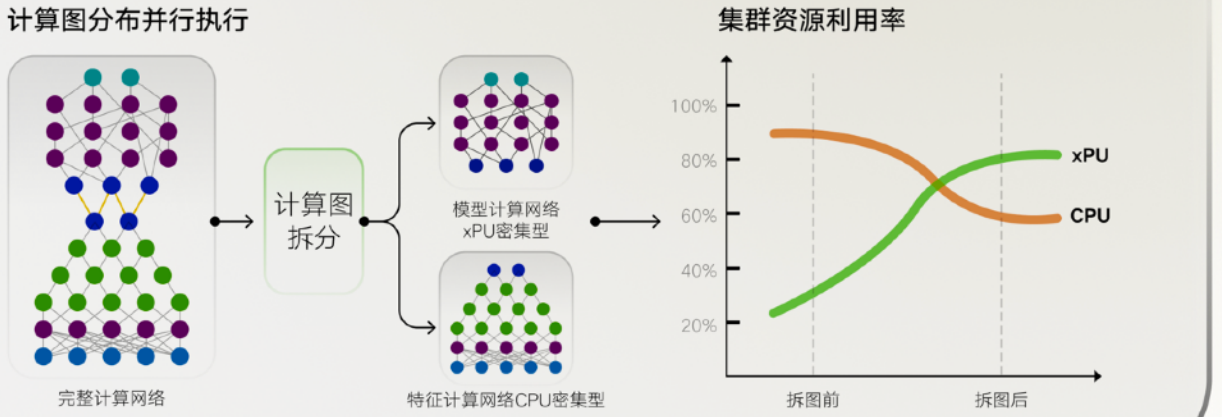
裁切LM Head权重Logits Mask，降低token概率空间和BeamSearch搜索空间，吞吐提升 70%+



面向广告业务特性的超低延迟推理定制化解法

算力安全 & 异构计算

引入CPU-xPU异构资源池与可信执行环境，兼顾高吞吐、低时延与数据安全。



多模态协同三阶段：图文对齐到兴趣迁移

1

三阶段任务

PreTrain阶段实现图文基础对齐，PostTrain阶段把内容表征迁移至兴趣空间，Application阶段完成CTR预估。

2

技术方案

采用对比学习、图搜召回率等中间指标以及Target Attention轻量交互，确保各阶段任务的顺利实现。

3

工程化优势

三阶段解耦使表征与任务模型可独立迭代，提升了上线效率，为工程复制提供了操作手册。

从特征库到知识工程：体系跃迁

传统特征库局限

传统特征库只能提供用户+商品浅层匹配，更新周期部分实时，部分T+1，无法满足实时性需求。

知识工程体系优势

知识工程体系用多模态图谱+非参数内存，支持实时注入新品、新潮流信息，实现因果级推理。

广告价值提升

广告价值从精准定向升级为策略生成与效果归因，平台获得按成交分成的新商业模型。

从特征库到知识工程：体系跃迁

内隐性知识

商品洞察

机制知识

场景知识

...

外显性知识

生成式画像

垂类知识

行业通识

...

从特征库到知识工程：知识图谱

01

图谱结构

营销知识图谱采用三层结构：行业层抽象营销目标与季节性规则，商品层聚合属性、卖点与竞品关系，用户层刻画跨类目兴趣演化，形成完整的营销知识体系。

02

统一本体

通过统一本体，将10亿商品、5亿用户行为映射至同一语义空间，使生成模型在解码阶段即可引用因果关联，显著减少幻觉，提高广告投放的精准度。

03

效果提升

这种知识图谱的构建方式使广告投放能够更好地理解用户需求和商品特性，从而实现更精准的广告推荐和更高的投放效果。

知识工程：非参数记忆实时更新

更新机制

系统采用增量向量索引+时序淘汰策略，在不停训的情况下分钟级写入新品、新促销信息，确保知识库的实时性和准确性。

性能表现

配合多版本并发控制，系统在大促高峰写入吞吐达1000万条每秒，同时线上查询P99延迟低于5ms，兼顾实效与稳定，为广告投放提供了强大的技术支持。

知识工程：多模态知识对齐

视频创意

以视频创意为例，系统先把素材帧级特征与商品图谱对齐，自动生成卖点标签，再与用户兴趣向量匹配，实现‘内容-商品-人’闭环。

实验结果

多模态知识融合对广告效果的直接增益，为广告创意提供了新的思路和方法。

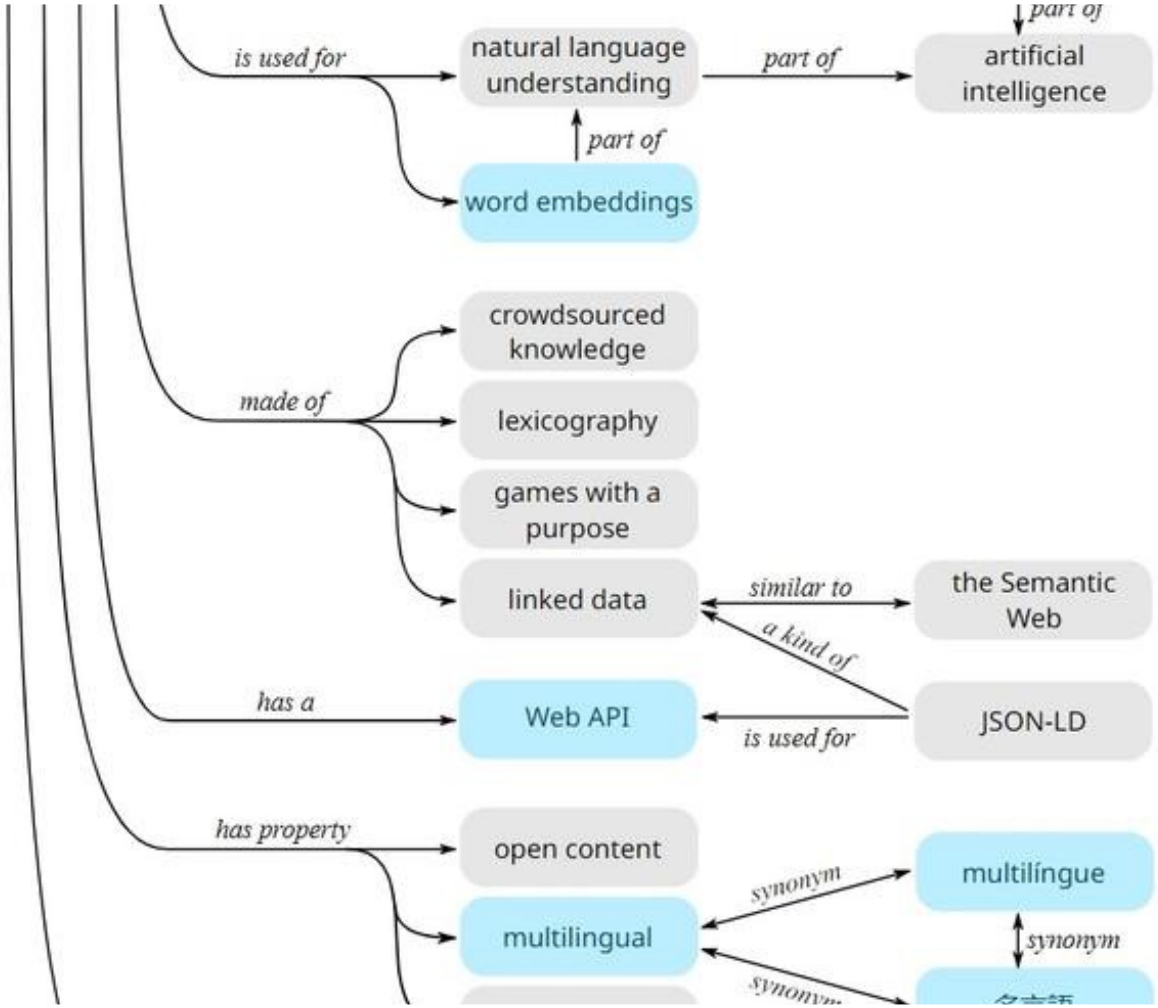
应用前景

这种多模态知识对齐技术为广告创意的生成和优化提供了新的方向，进一步提升广告的吸引力和效果。

知识工程：知识图谱如何根治创意幻觉

生成式最大风险是虚构。知识图谱成为“护栏”。

- **生成前核查：**基于商品属性、库存等事实，hallucination率降至5% 。
- **生成后校验：**自动过滤极限词、侵权词，保障品牌安全。
- **结果可追溯：**让每一条文案都可追溯、可审计、可赔付。



创意知识：卖点–人群–素材三域协同生成

卖点域
读取知识图谱
材质、功效、场景

+

人群域
匹配用户
长短期兴趣

+

素材域
调用引擎
生成图文视频

↓

20秒	成本 -XX%	CTR +XX%
输出XX套创意	创意制作成本	点击率提升

B端知识：秒级投放–中小商家门槛降七成

投放简化

一句话指令即可完成预算、人群、创意三步配置，系统秒级给出可执行方案，无需人工选词、出价。

门槛降低

测试显示，50%中小商家首次投放时间从3小时缩短至15分钟，入门门槛降低70%，投放技术被封装为黑箱式服务。

04 总结与展望

生成式技术三大启示：技术路径、引擎、价值

1

技术路径

大模型落地广告需走通用→垂类→知识增强递进路线，不可跳级，逐步提升模型的适应性和效果。

2

核心引擎

GRAM+多Agent构成新一代生成式引擎，打通创意、投放、归因全链路，提升广告投放的整体效率。

3

价值导向

衡量标准唯有降低门槛、提升效率、保障效果，以实现广告主、用户、平台三方的长期共赢。

端到端生成，就足够了吗?

幻觉不可0

复杂业务场景下，事实准确性风险仍高。

业务不可控

广告主难以干预生成逻辑，不符合营销诉求。

黑天鹅不可预测

突发舆情、合规变化等极端场景无法应对。

我们的探索：“生成式+”混合智能

知识增强 + 人机协同 + 规则兜底，极端风险事件损失降低 90% 。

混合增强：端到端并非终点

端到端技术局限

纯生成式方案存在可控性不足、算力成本高、伦理风险三大缺陷，难以满足广告投放的多样化需求。

混合增强范式

京东采用生成式+判别式协同，生成模型负责创新策略，判别模型实时校验品牌调性与合规。

长期目标

构建可信、可控、高效的三赢广告生态，实现广告主、用户、平台三方共赢。

下一代方向：动态环境感知

01 环境感知投放

结合天气、地理位置、人流密度实时调整创意内容与出价策略，实现线上线下联动，提升广告投放效果。

02 实验效果

初步实验显示，环境感知投放使线下到店率提升12%，为本地生活商家提供天时地利组合方案，拓展广告增量空间。

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

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

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

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

4月 — 6月 — 8月 — 10月 — 12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

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

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