# 大模型的模型结构

在推荐系统中，**大模型**（通常指大语言模型或多模态大模型）往往采用 Transformer 编码器–解码器架构。例如，快手提出的 OneRec 模型使用了一个编码器对用户的行为序列进行压缩以提取兴趣特征，同时在解码器中引入了大量专家混合（MoE）参数来精确解码推荐内容[[1]](https://ar5iv.labs.arxiv.org/html/2506.13695v1#:~:text=architecture.%20Built%20as%20an%20encoder,training%2C%20we%20develop%20a)。这种设计显著提升了模型容量和表达能力。对于物品表示，有的工作采用了**语义 ID（Semantic ID）**的方法：首先利用预训练语言/视觉编码器将物品内容（如标题、描述、图片）编码成连续向量，再通过量化方法将其离散化为一组语义代码字序列（Semantic ID），然后用 Transformer 来预测下一个要推荐的语义 ID[[2]](https://ar5iv.labs.arxiv.org/html/2305.05065#:~:text=refer%20to%20our%20method%20as,Transformer%20model%20on%20the%20sequential)。这种方式相当于用“语义”特征取代传统的原子型 Item ID，使模型能更好地理解新物品和冷启动场景。值得注意的是，有研究将召回和排序任务统一为**生成式序列建模**问题；例如 HSTU 提出将排名和检索任务视作序列到序列的生成问题进行建模[[3]](https://ar5iv.labs.arxiv.org/html/2402.17152#:~:text=Our%20key%20insights%20are%2C%20a,Due%20to)。HSTU 基于此设计了新型 Transformer 单元，在 1.5 万亿参数规模下，在线上 A/B 测试中带来了 12.4% 的指标提升[[4]](https://ar5iv.labs.arxiv.org/html/2402.17152#:~:text=high%20cardinality%2C%20non,More%20importantly)。这些尝试表明，大模型可以通过结构创新，将推荐的各个阶段（如召回、粗排、精排）融合为统一的生成式框架。

## 大模型的世界知识

**物品侧知识**：大模型通过理解多模态内容来获取物品语义。例如，类似快手 M3CSR 框架会提取视频的视觉、文本和音频特征，并将这些多模态内容聚类成类别 ID，得到可学习的内容表征，从而支持冷启动推荐[[5]](https://recsys.acm.org/recsys24/accepted-contributions/#:~:text=specific%20interest%20and%20behavioral%20interest,start%20scenarios)。此外，Tiger 等工作用语义 ID 代替原子 Item ID，将物品内容映射为语义嵌入，然后在生成式模型中预测下一个物品的语义 ID，从而增强推荐系统对新物品的泛化能力[[2]](https://ar5iv.labs.arxiv.org/html/2305.05065#:~:text=refer%20to%20our%20method%20as,Transformer%20model%20on%20the%20sequential)。**用户侧知识**：大模型可对用户历史行为进行摘要提炼，生成表示用户兴趣的向量。例如 OneRec 的编码器能够压缩用户的终身行为序列以提取兴趣特征[[1]](https://ar5iv.labs.arxiv.org/html/2506.13695v1#:~:text=architecture.%20Built%20as%20an%20encoder,training%2C%20we%20develop%20a)，并通过后续的强化学习阶段对模型输出进行优化，使推荐结果更符合用户偏好[[6]](https://ar5iv.labs.arxiv.org/html/2506.13695v1#:~:text=Mixture,framework%20to%20refine%20recommendations%20by)。这种方式有效利用了大模型的序列建模能力，将用户过去的点击、观看等行为转换为能够指导推荐的表征。此外，多模态特征也能延伸到用户侧：除了常规的浏览历史外，可以将用户画像、环境上下文等转换为文本或其他模态输入，通过 LLM 推理用户当前意图。总的来说，大模型通过深度理解用户和物品的语义信息，实现了更丰富的世界知识建模。

## 训练范式的创新

近期**预训练 + 微调 + 强化学习**的流程在推荐中开始流行。例如，Meta 提出的 ExFM 框架先训练一个通用的大语言模型作为教师，然后将教师在多种任务上的预测结果作为标签对专业学生模型进行外部蒸馏[[7]](https://ar5iv.labs.arxiv.org/html/2502.17494v7#:~:text=In%20light%20of%20the%20above,Foundation%20Model%20that%20can%20serve)。这种“外部蒸馏”思路使得万亿级参数的模型可以通过离线标注的方式指导线上服务模型，兼顾了效率和效果。另外，像 OneRec 这样的生成式推荐模型会在离线训练后引入强化学习，用定制化的奖励函数进一步校准模型输出，使推荐结果在多目标指标上达到平衡[[6]](https://ar5iv.labs.arxiv.org/html/2506.13695v1#:~:text=Mixture,framework%20to%20refine%20recommendations%20by)。为提高训练效率，还可以采用 LoRA 等参数高效微调方法，甚至将多个小模型合并（如近期提出的 SOLAR、SPHINX 等方法）以快速扩大模型规模。在模型联合训练方面，也出现了将搜索广告与推荐模型共同训练的尝试。例如，京东利用 TensorRT-LLM 框架在推荐广告和搜索广告业务中部署了统一的生成式召回模型，实现了跨场景的一体化训练与推理[[8]](https://developer.nvidia.com/zh-cn/blog/nvidia-tensorrt-llm-ad-generation/#:~:text=%E7%9B%AE%E5%89%8D%EF%BC%8C%E7%94%9F%E6%88%90%E5%BC%8F%E5%8F%AC%E5%9B%9E%E4%B8%80%E6%9C%9F%E5%B7%B2%E5%9C%A8%E4%BA%AC%E4%B8%9C%E6%8E%A8%E8%8D%90%E5%B9%BF%E5%91%8A%E5%8F%8A%E6%90%9C%E7%B4%A2%E5%B9%BF%E5%91%8A%E7%AD%89%E4%B8%BB%E8%A6%81%E4%B8%9A%E5%8A%A1%E7%BA%BF%E6%88%90%E5%8A%9F%E5%AE%9E%E6%96%BD%E3%80%82%E5%9C%A8%E6%8E%A8%E8%8D%90%E5%B9%BF%E5%91%8A%E6%96%B9%E9%9D%A2%EF%BC%8C%E5%80%9F%E5%8A%A9%E7%94%9F%E6%88%90%E5%BC%8F%E6%A8%A1%E5%9E%8B%E7%9A%84%E5%8F%82%E6%95%B0%E8%A7%84%E6%A8%A1%E5%8F%8A%E8%AF%AD%E4%B9%89%E7%90%86%E8%A7%A3%E4%BC%98%E5%8A%BF%EF%BC%8CAB%20%E5%AE%9E%E9%AA%8C%E7%BB%93%E6%9E%9C%E6%98%BE%E7%A4%BA%E5%95%86%E5%93%81%E7%82%B9%E5%87%BB%E7%8E%87%E4%B8%8E%E6%B6%88%E8%B4%B9%E5%B8%A6%E6%9D%A5%E4%BA%86%E6%98%BE%E8%91%97%E7%9A%84%E6%8F%90%E5%8D%87%E3%80%82%E5%9C%A8%E6%90%9C%E7%B4%A2%E5%B9%BF%E5%91%8A%E6%96%B9%E9%9D%A2%EF%BC%8C%E9%80%9A%E8%BF%87%20LLM%20%E6%89%80%E5%85%B7%E5%A4%87%E7%9A%84%E8%AF%AD%E4%B9%89%E7%90%86%E8%A7%A3%E8%83%BD%E5%8A%9B%EF%BC%8C%E6%98%BE%E8%91%97%E6%8F%90%E5%8D%87%E4%BA%86%E5%AF%B9%E6%9F%A5%E8%AF%A2%E4%B8%8E%E5%95%86%E5%93%81%E7%9A%84%E8%AE%A4%E7%9F%A5%E8%83%BD%E5%8A%9B%EF%BC%8C%E7%89%B9%E5%88%AB%E6%98%AF%E5%9C%A8%E5%A4%84%E7%90%86%E6%90%9C%E7%B4%A2%E4%B8%AD%E7%9A%84%E9%95%BF%E5%B0%BE%E6%9F%A5%E8%AF%A2%E6%97%B6%EF%BC%8C%E5%A1%AB%E5%85%85%E7%8E%87%E6%9C%89%E6%98%8E%E6%98%BE%E6%8F%90%E5%8D%87%EF%BC%8CAB%20%E5%AE%9E%E9%AA%8C%E5%90%8C%E6%A0%B7%E5%8F%96%E5%BE%97%E4%BA%86%E7%82%B9%E5%87%BB%E7%8E%87%E4%B8%8E%E6%B6%88%E8%B4%B9%E5%87%A0%E4%B8%AA%E7%99%BE%E5%88%86%E7%82%B9%E7%9A%84%E6%94%B6%E7%9B%8A%E5%A2%9E%E9%95%BF%E3%80%82)。总之，通过预训练赋能、微调精调、以及强化学习调优的组合，大模型正逐步融入工业推荐系统的端到端训练流程中。

## 推理范式的创新

在大模型推理时，**缓存机制**是一项重要优化。针对大模型的自回归生成，Key-Value Cache 可以避免重复计算前面层的特征；近来有工作进一步提出在前几层生成“注册令牌”（register tokens）来压缩历史信息，在后续层只关注这些令牌，大幅减少 KV Cache 的大小，从而获得数倍的加速效果[[9]](https://arxiv.org/abs/2507.00715#:~:text=Extensive%20experiments%20on%20three%20datasets%2C,effectiveness%20gap%20in%20LLMRec%2C%20offering)。同时，推理过程中的**加速方法**也在迅速演进。例如，NVIDIA 的 TensorRT-LLM 平台通过流水线优化、Flash Attention 和并行采样等技术，将原本在基准方案上需要 100ms 完成的推理，在相同约束下实现了 5 倍以上的吞吐提升[[10]](https://ar5iv.labs.arxiv.org/html/2506.13695v1#:~:text=Infrastructure%20and%20Efficiency%20We%20utilize,we%20achieve%20a%205%20throughput)。下图展示了利用优化框架后延迟和吞吐的对比：

*图：NVIDIA TensorRT-LLM 优化前后的延迟对比。改进后延迟显著降低（数据来源见*[*[11]*](https://developer.nvidia.com/zh-cn/blog/nvidia-tensorrt-llm-ad-generation/#:~:text=%E5%9C%A8%20NVIDIA%20GPU%C2%A0%C2%A0%E4%B8%8A%E8%BF%9B%E8%A1%8C%E7%9A%84%E6%B5%8B%E8%AF%95%E4%B8%AD%EF%BC%8C%E9%80%9A%E8%BF%87%E5%AF%B9%E6%AF%94%20TensorRT,LLM%20%E8%BF%9B%E8%A1%8C%E6%8E%A8%E7%90%86%EF%BC%8C%E5%85%B6%E5%90%9E%E5%90%90%E9%87%8F%E7%9B%B8%E8%BE%83%E4%BA%8Ebaseline%E6%8F%90%E5%8D%87%E4%BA%86%E4%BA%94%E5%80%8D%E4%BB%A5%E4%B8%8A%E3%80%82%E8%BF%99%E7%9B%B8%E5%BD%93%E4%BA%8E%E5%B0%86%E9%83%A8%E7%BD%B2%E6%88%90%E6%9C%AC%E9%99%8D%E8%87%B3%E5%8E%9F%E6%9D%A5%E7%9A%84%E4%BA%94%E5%88%86%E4%B9%8B%E4%B8%80%E3%80%82)*）。*

除了框架优化，**解码策略**也不断改进。对比搜索（beam search）、多 token 预测等技术可以在一定程度上降低实时推理耗时。实际中，合理设置 beam 宽度能有效提高召回准确率，不过也可能增加延迟，为此需要在效果和性能间权衡。下图为吞吐量对比：在固定延迟预算下，采用 TensorRT-LLM 推理使每秒查询数（QPS）提高了 5 倍以上[[11]](https://developer.nvidia.com/zh-cn/blog/nvidia-tensorrt-llm-ad-generation/#:~:text=%E5%9C%A8%20NVIDIA%20GPU%C2%A0%C2%A0%E4%B8%8A%E8%BF%9B%E8%A1%8C%E7%9A%84%E6%B5%8B%E8%AF%95%E4%B8%AD%EF%BC%8C%E9%80%9A%E8%BF%87%E5%AF%B9%E6%AF%94%20TensorRT,LLM%20%E8%BF%9B%E8%A1%8C%E6%8E%A8%E7%90%86%EF%BC%8C%E5%85%B6%E5%90%9E%E5%90%90%E9%87%8F%E7%9B%B8%E8%BE%83%E4%BA%8Ebaseline%E6%8F%90%E5%8D%87%E4%BA%86%E4%BA%94%E5%80%8D%E4%BB%A5%E4%B8%8A%E3%80%82%E8%BF%99%E7%9B%B8%E5%BD%93%E4%BA%8E%E5%B0%86%E9%83%A8%E7%BD%B2%E6%88%90%E6%9C%AC%E9%99%8D%E8%87%B3%E5%8E%9F%E6%9D%A5%E7%9A%84%E4%BA%94%E5%88%86%E4%B9%8B%E4%B8%80%E3%80%82)。

*图：TensorRT-LLM 优化后吞吐量与基线方案的对比，吞吐提升显著*[*[11]*](https://developer.nvidia.com/zh-cn/blog/nvidia-tensorrt-llm-ad-generation/#:~:text=%E5%9C%A8%20NVIDIA%20GPU%C2%A0%C2%A0%E4%B8%8A%E8%BF%9B%E8%A1%8C%E7%9A%84%E6%B5%8B%E8%AF%95%E4%B8%AD%EF%BC%8C%E9%80%9A%E8%BF%87%E5%AF%B9%E6%AF%94%20TensorRT,LLM%20%E8%BF%9B%E8%A1%8C%E6%8E%A8%E7%90%86%EF%BC%8C%E5%85%B6%E5%90%9E%E5%90%90%E9%87%8F%E7%9B%B8%E8%BE%83%E4%BA%8Ebaseline%E6%8F%90%E5%8D%87%E4%BA%86%E4%BA%94%E5%80%8D%E4%BB%A5%E4%B8%8A%E3%80%82%E8%BF%99%E7%9B%B8%E5%BD%93%E4%BA%8E%E5%B0%86%E9%83%A8%E7%BD%B2%E6%88%90%E6%9C%AC%E9%99%8D%E8%87%B3%E5%8E%9F%E6%9D%A5%E7%9A%84%E4%BA%94%E5%88%86%E4%B9%8B%E4%B8%80%E3%80%82)*。*

在**线上链路覆盖**方面，生成式推荐正逐渐打破传统多级过滤的结构。JD 的实践表明，生成式召回已经能与传统检索模块并行运行，并在推荐广告和搜索广告场景下同时部署。在推荐广告线上实验中，基于大模型的生成式召回显著提升了商品点击率和消费额；在搜索广告中，利用 LLM 提升了对长尾查询的理解能力，同样带来了点击率和消费增长（提升了几个百分点）[[8]](https://developer.nvidia.com/zh-cn/blog/nvidia-tensorrt-llm-ad-generation/#:~:text=%E7%9B%AE%E5%89%8D%EF%BC%8C%E7%94%9F%E6%88%90%E5%BC%8F%E5%8F%AC%E5%9B%9E%E4%B8%80%E6%9C%9F%E5%B7%B2%E5%9C%A8%E4%BA%AC%E4%B8%9C%E6%8E%A8%E8%8D%90%E5%B9%BF%E5%91%8A%E5%8F%8A%E6%90%9C%E7%B4%A2%E5%B9%BF%E5%91%8A%E7%AD%89%E4%B8%BB%E8%A6%81%E4%B8%9A%E5%8A%A1%E7%BA%BF%E6%88%90%E5%8A%9F%E5%AE%9E%E6%96%BD%E3%80%82%E5%9C%A8%E6%8E%A8%E8%8D%90%E5%B9%BF%E5%91%8A%E6%96%B9%E9%9D%A2%EF%BC%8C%E5%80%9F%E5%8A%A9%E7%94%9F%E6%88%90%E5%BC%8F%E6%A8%A1%E5%9E%8B%E7%9A%84%E5%8F%82%E6%95%B0%E8%A7%84%E6%A8%A1%E5%8F%8A%E8%AF%AD%E4%B9%89%E7%90%86%E8%A7%A3%E4%BC%98%E5%8A%BF%EF%BC%8CAB%20%E5%AE%9E%E9%AA%8C%E7%BB%93%E6%9E%9C%E6%98%BE%E7%A4%BA%E5%95%86%E5%93%81%E7%82%B9%E5%87%BB%E7%8E%87%E4%B8%8E%E6%B6%88%E8%B4%B9%E5%B8%A6%E6%9D%A5%E4%BA%86%E6%98%BE%E8%91%97%E7%9A%84%E6%8F%90%E5%8D%87%E3%80%82%E5%9C%A8%E6%90%9C%E7%B4%A2%E5%B9%BF%E5%91%8A%E6%96%B9%E9%9D%A2%EF%BC%8C%E9%80%9A%E8%BF%87%20LLM%20%E6%89%80%E5%85%B7%E5%A4%87%E7%9A%84%E8%AF%AD%E4%B9%89%E7%90%86%E8%A7%A3%E8%83%BD%E5%8A%9B%EF%BC%8C%E6%98%BE%E8%91%97%E6%8F%90%E5%8D%87%E4%BA%86%E5%AF%B9%E6%9F%A5%E8%AF%A2%E4%B8%8E%E5%95%86%E5%93%81%E7%9A%84%E8%AE%A4%E7%9F%A5%E8%83%BD%E5%8A%9B%EF%BC%8C%E7%89%B9%E5%88%AB%E6%98%AF%E5%9C%A8%E5%A4%84%E7%90%86%E6%90%9C%E7%B4%A2%E4%B8%AD%E7%9A%84%E9%95%BF%E5%B0%BE%E6%9F%A5%E8%AF%A2%E6%97%B6%EF%BC%8C%E5%A1%AB%E5%85%85%E7%8E%87%E6%9C%89%E6%98%8E%E6%98%BE%E6%8F%90%E5%8D%87%EF%BC%8CAB%20%E5%AE%9E%E9%AA%8C%E5%90%8C%E6%A0%B7%E5%8F%96%E5%BE%97%E4%BA%86%E7%82%B9%E5%87%BB%E7%8E%87%E4%B8%8E%E6%B6%88%E8%B4%B9%E5%87%A0%E4%B8%AA%E7%99%BE%E5%88%86%E7%82%B9%E7%9A%84%E6%94%B6%E7%9B%8A%E5%A2%9E%E9%95%BF%E3%80%82)。未来，可进一步考虑将稀疏模型与大模型联合推理，例如先用经典的索引召回，再用大模型精细排布，以兼顾效率与精度。

## 探索性方向与未来趋势

面向未来，大模型在推荐领域有多种前沿方向值得探索。一是**多智能体协同推荐**，即让多个专门的模型/代理协作完成推荐任务。最近的研究如 MACRec 提出通过引入“管理者”、“用户分析者”、“搜索者”等多种角色，让不同 agent 协作来解决推荐问题[[12]](https://arxiv.org/abs/2402.15235#:~:text=%3E%20Abstract%3ALLM,Analyst%2C%20Reflector%2C%20Searcher%2C%20and%20Task)。这种方法有望将推荐分工更加细化，实现更灵活的系统设计。二是引入**推理能力和链式思维**。目前 LLM 已表现出强大的归纳推理能力，在推荐中可通过特殊的提示或模型结构让大模型生成可解释的推荐理由，或在生成过程中采用分步推理，以改善推荐可信度。三是**工具调用与检索增强生成（RAG）**。可让大模型在推荐过程中主动查询知识库、外部数据库或调用插件来获得最新信息，例如调用搜索接口获取热点新闻，再结合个性化历史给出推荐。四是**全模态生成和沉浸式推荐**。随着 AR/VR 技术的发展，未来推荐可能不再局限于列表或卡片式，而是生成沉浸式的交互式内容（如虚拟购物场景、3D 商品展示等），提供身临其境的体验。大模型具备生成不同模态内容的潜力，这一方向或能为用户带来更丰富的推荐体验。

**参考文献：** 本文论述引用了多篇来自工业界的最新研究与实践报告，如 OneRec、ExFM、Tiger 等[[1]](https://ar5iv.labs.arxiv.org/html/2506.13695v1#:~:text=architecture.%20Built%20as%20an%20encoder,training%2C%20we%20develop%20a)[[7]](https://ar5iv.labs.arxiv.org/html/2502.17494v7#:~:text=In%20light%20of%20the%20above,Foundation%20Model%20that%20can%20serve)[[8]](https://developer.nvidia.com/zh-cn/blog/nvidia-tensorrt-llm-ad-generation/#:~:text=%E7%9B%AE%E5%89%8D%EF%BC%8C%E7%94%9F%E6%88%90%E5%BC%8F%E5%8F%AC%E5%9B%9E%E4%B8%80%E6%9C%9F%E5%B7%B2%E5%9C%A8%E4%BA%AC%E4%B8%9C%E6%8E%A8%E8%8D%90%E5%B9%BF%E5%91%8A%E5%8F%8A%E6%90%9C%E7%B4%A2%E5%B9%BF%E5%91%8A%E7%AD%89%E4%B8%BB%E8%A6%81%E4%B8%9A%E5%8A%A1%E7%BA%BF%E6%88%90%E5%8A%9F%E5%AE%9E%E6%96%BD%E3%80%82%E5%9C%A8%E6%8E%A8%E8%8D%90%E5%B9%BF%E5%91%8A%E6%96%B9%E9%9D%A2%EF%BC%8C%E5%80%9F%E5%8A%A9%E7%94%9F%E6%88%90%E5%BC%8F%E6%A8%A1%E5%9E%8B%E7%9A%84%E5%8F%82%E6%95%B0%E8%A7%84%E6%A8%A1%E5%8F%8A%E8%AF%AD%E4%B9%89%E7%90%86%E8%A7%A3%E4%BC%98%E5%8A%BF%EF%BC%8CAB%20%E5%AE%9E%E9%AA%8C%E7%BB%93%E6%9E%9C%E6%98%BE%E7%A4%BA%E5%95%86%E5%93%81%E7%82%B9%E5%87%BB%E7%8E%87%E4%B8%8E%E6%B6%88%E8%B4%B9%E5%B8%A6%E6%9D%A5%E4%BA%86%E6%98%BE%E8%91%97%E7%9A%84%E6%8F%90%E5%8D%87%E3%80%82%E5%9C%A8%E6%90%9C%E7%B4%A2%E5%B9%BF%E5%91%8A%E6%96%B9%E9%9D%A2%EF%BC%8C%E9%80%9A%E8%BF%87%20LLM%20%E6%89%80%E5%85%B7%E5%A4%87%E7%9A%84%E8%AF%AD%E4%B9%89%E7%90%86%E8%A7%A3%E8%83%BD%E5%8A%9B%EF%BC%8C%E6%98%BE%E8%91%97%E6%8F%90%E5%8D%87%E4%BA%86%E5%AF%B9%E6%9F%A5%E8%AF%A2%E4%B8%8E%E5%95%86%E5%93%81%E7%9A%84%E8%AE%A4%E7%9F%A5%E8%83%BD%E5%8A%9B%EF%BC%8C%E7%89%B9%E5%88%AB%E6%98%AF%E5%9C%A8%E5%A4%84%E7%90%86%E6%90%9C%E7%B4%A2%E4%B8%AD%E7%9A%84%E9%95%BF%E5%B0%BE%E6%9F%A5%E8%AF%A2%E6%97%B6%EF%BC%8C%E5%A1%AB%E5%85%85%E7%8E%87%E6%9C%89%E6%98%8E%E6%98%BE%E6%8F%90%E5%8D%87%EF%BC%8CAB%20%E5%AE%9E%E9%AA%8C%E5%90%8C%E6%A0%B7%E5%8F%96%E5%BE%97%E4%BA%86%E7%82%B9%E5%87%BB%E7%8E%87%E4%B8%8E%E6%B6%88%E8%B4%B9%E5%87%A0%E4%B8%AA%E7%99%BE%E5%88%86%E7%82%B9%E7%9A%84%E6%94%B6%E7%9B%8A%E5%A2%9E%E9%95%BF%E3%80%82)。其中多项工作已在生产环境中通过 A/B 测试验证了效果，如 Kuaishou 和京东的线上实验结果[[13]](https://ar5iv.labs.arxiv.org/html/2506.13695v1#:~:text=statistically%20significant%20improvements%20of%20%2B0.54,with%20implementation%20details%20available%20in)[[8]](https://developer.nvidia.com/zh-cn/blog/nvidia-tensorrt-llm-ad-generation/#:~:text=%E7%9B%AE%E5%89%8D%EF%BC%8C%E7%94%9F%E6%88%90%E5%BC%8F%E5%8F%AC%E5%9B%9E%E4%B8%80%E6%9C%9F%E5%B7%B2%E5%9C%A8%E4%BA%AC%E4%B8%9C%E6%8E%A8%E8%8D%90%E5%B9%BF%E5%91%8A%E5%8F%8A%E6%90%9C%E7%B4%A2%E5%B9%BF%E5%91%8A%E7%AD%89%E4%B8%BB%E8%A6%81%E4%B8%9A%E5%8A%A1%E7%BA%BF%E6%88%90%E5%8A%9F%E5%AE%9E%E6%96%BD%E3%80%82%E5%9C%A8%E6%8E%A8%E8%8D%90%E5%B9%BF%E5%91%8A%E6%96%B9%E9%9D%A2%EF%BC%8C%E5%80%9F%E5%8A%A9%E7%94%9F%E6%88%90%E5%BC%8F%E6%A8%A1%E5%9E%8B%E7%9A%84%E5%8F%82%E6%95%B0%E8%A7%84%E6%A8%A1%E5%8F%8A%E8%AF%AD%E4%B9%89%E7%90%86%E8%A7%A3%E4%BC%98%E5%8A%BF%EF%BC%8CAB%20%E5%AE%9E%E9%AA%8C%E7%BB%93%E6%9E%9C%E6%98%BE%E7%A4%BA%E5%95%86%E5%93%81%E7%82%B9%E5%87%BB%E7%8E%87%E4%B8%8E%E6%B6%88%E8%B4%B9%E5%B8%A6%E6%9D%A5%E4%BA%86%E6%98%BE%E8%91%97%E7%9A%84%E6%8F%90%E5%8D%87%E3%80%82%E5%9C%A8%E6%90%9C%E7%B4%A2%E5%B9%BF%E5%91%8A%E6%96%B9%E9%9D%A2%EF%BC%8C%E9%80%9A%E8%BF%87%20LLM%20%E6%89%80%E5%85%B7%E5%A4%87%E7%9A%84%E8%AF%AD%E4%B9%89%E7%90%86%E8%A7%A3%E8%83%BD%E5%8A%9B%EF%BC%8C%E6%98%BE%E8%91%97%E6%8F%90%E5%8D%87%E4%BA%86%E5%AF%B9%E6%9F%A5%E8%AF%A2%E4%B8%8E%E5%95%86%E5%93%81%E7%9A%84%E8%AE%A4%E7%9F%A5%E8%83%BD%E5%8A%9B%EF%BC%8C%E7%89%B9%E5%88%AB%E6%98%AF%E5%9C%A8%E5%A4%84%E7%90%86%E6%90%9C%E7%B4%A2%E4%B8%AD%E7%9A%84%E9%95%BF%E5%B0%BE%E6%9F%A5%E8%AF%A2%E6%97%B6%EF%BC%8C%E5%A1%AB%E5%85%85%E7%8E%87%E6%9C%89%E6%98%8E%E6%98%BE%E6%8F%90%E5%8D%87%EF%BC%8CAB%20%E5%AE%9E%E9%AA%8C%E5%90%8C%E6%A0%B7%E5%8F%96%E5%BE%97%E4%BA%86%E7%82%B9%E5%87%BB%E7%8E%87%E4%B8%8E%E6%B6%88%E8%B4%B9%E5%87%A0%E4%B8%AA%E7%99%BE%E5%88%86%E7%82%B9%E7%9A%84%E6%94%B6%E7%9B%8A%E5%A2%9E%E9%95%BF%E3%80%82)。未来规划部分结合了学术界的思路和实践经验，例如多智能体推荐框架[[12]](https://arxiv.org/abs/2402.15235#:~:text=%3E%20Abstract%3ALLM,Analyst%2C%20Reflector%2C%20Searcher%2C%20and%20Task)等，为产业界应用提供了有益参考。

[[1]](https://ar5iv.labs.arxiv.org/html/2506.13695v1#:~:text=architecture.%20Built%20as%20an%20encoder,training%2C%20we%20develop%20a) [[6]](https://ar5iv.labs.arxiv.org/html/2506.13695v1#:~:text=Mixture,framework%20to%20refine%20recommendations%20by) [[10]](https://ar5iv.labs.arxiv.org/html/2506.13695v1#:~:text=Infrastructure%20and%20Efficiency%20We%20utilize,we%20achieve%20a%205%20throughput) [[13]](https://ar5iv.labs.arxiv.org/html/2506.13695v1#:~:text=statistically%20significant%20improvements%20of%20%2B0.54,with%20implementation%20details%20available%20in) [2506.13695] OneRec Technical Report

<https://ar5iv.labs.arxiv.org/html/2506.13695v1>

[[2]](https://ar5iv.labs.arxiv.org/html/2305.05065#:~:text=refer%20to%20our%20method%20as,Transformer%20model%20on%20the%20sequential) [2305.05065] Recommender Systems with Generative Retrieval

<https://ar5iv.labs.arxiv.org/html/2305.05065>

[[3]](https://ar5iv.labs.arxiv.org/html/2402.17152#:~:text=Our%20key%20insights%20are%2C%20a,Due%20to) [[4]](https://ar5iv.labs.arxiv.org/html/2402.17152#:~:text=high%20cardinality%2C%20non,More%20importantly) [2402.17152] Actions Speak Louder than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations

<https://ar5iv.labs.arxiv.org/html/2402.17152>

[[5]](https://recsys.acm.org/recsys24/accepted-contributions/#:~:text=specific%20interest%20and%20behavioral%20interest,start%20scenarios) RecSys 2024 - Accepted Contributions - RecSys – RecSys

<https://recsys.acm.org/recsys24/accepted-contributions/>

[[7]](https://ar5iv.labs.arxiv.org/html/2502.17494v7#:~:text=In%20light%20of%20the%20above,Foundation%20Model%20that%20can%20serve) [2502.17494] External Large Foundation Model: How to Efficiently Serve Trillions of Parameters for Online Ads Recommendation

<https://ar5iv.labs.arxiv.org/html/2502.17494v7>

[[8]](https://developer.nvidia.com/zh-cn/blog/nvidia-tensorrt-llm-ad-generation/#:~:text=%E7%9B%AE%E5%89%8D%EF%BC%8C%E7%94%9F%E6%88%90%E5%BC%8F%E5%8F%AC%E5%9B%9E%E4%B8%80%E6%9C%9F%E5%B7%B2%E5%9C%A8%E4%BA%AC%E4%B8%9C%E6%8E%A8%E8%8D%90%E5%B9%BF%E5%91%8A%E5%8F%8A%E6%90%9C%E7%B4%A2%E5%B9%BF%E5%91%8A%E7%AD%89%E4%B8%BB%E8%A6%81%E4%B8%9A%E5%8A%A1%E7%BA%BF%E6%88%90%E5%8A%9F%E5%AE%9E%E6%96%BD%E3%80%82%E5%9C%A8%E6%8E%A8%E8%8D%90%E5%B9%BF%E5%91%8A%E6%96%B9%E9%9D%A2%EF%BC%8C%E5%80%9F%E5%8A%A9%E7%94%9F%E6%88%90%E5%BC%8F%E6%A8%A1%E5%9E%8B%E7%9A%84%E5%8F%82%E6%95%B0%E8%A7%84%E6%A8%A1%E5%8F%8A%E8%AF%AD%E4%B9%89%E7%90%86%E8%A7%A3%E4%BC%98%E5%8A%BF%EF%BC%8CAB%20%E5%AE%9E%E9%AA%8C%E7%BB%93%E6%9E%9C%E6%98%BE%E7%A4%BA%E5%95%86%E5%93%81%E7%82%B9%E5%87%BB%E7%8E%87%E4%B8%8E%E6%B6%88%E8%B4%B9%E5%B8%A6%E6%9D%A5%E4%BA%86%E6%98%BE%E8%91%97%E7%9A%84%E6%8F%90%E5%8D%87%E3%80%82%E5%9C%A8%E6%90%9C%E7%B4%A2%E5%B9%BF%E5%91%8A%E6%96%B9%E9%9D%A2%EF%BC%8C%E9%80%9A%E8%BF%87%20LLM%20%E6%89%80%E5%85%B7%E5%A4%87%E7%9A%84%E8%AF%AD%E4%B9%89%E7%90%86%E8%A7%A3%E8%83%BD%E5%8A%9B%EF%BC%8C%E6%98%BE%E8%91%97%E6%8F%90%E5%8D%87%E4%BA%86%E5%AF%B9%E6%9F%A5%E8%AF%A2%E4%B8%8E%E5%95%86%E5%93%81%E7%9A%84%E8%AE%A4%E7%9F%A5%E8%83%BD%E5%8A%9B%EF%BC%8C%E7%89%B9%E5%88%AB%E6%98%AF%E5%9C%A8%E5%A4%84%E7%90%86%E6%90%9C%E7%B4%A2%E4%B8%AD%E7%9A%84%E9%95%BF%E5%B0%BE%E6%9F%A5%E8%AF%A2%E6%97%B6%EF%BC%8C%E5%A1%AB%E5%85%85%E7%8E%87%E6%9C%89%E6%98%8E%E6%98%BE%E6%8F%90%E5%8D%87%EF%BC%8CAB%20%E5%AE%9E%E9%AA%8C%E5%90%8C%E6%A0%B7%E5%8F%96%E5%BE%97%E4%BA%86%E7%82%B9%E5%87%BB%E7%8E%87%E4%B8%8E%E6%B6%88%E8%B4%B9%E5%87%A0%E4%B8%AA%E7%99%BE%E5%88%86%E7%82%B9%E7%9A%84%E6%94%B6%E7%9B%8A%E5%A2%9E%E9%95%BF%E3%80%82) [[11]](https://developer.nvidia.com/zh-cn/blog/nvidia-tensorrt-llm-ad-generation/#:~:text=%E5%9C%A8%20NVIDIA%20GPU%C2%A0%C2%A0%E4%B8%8A%E8%BF%9B%E8%A1%8C%E7%9A%84%E6%B5%8B%E8%AF%95%E4%B8%AD%EF%BC%8C%E9%80%9A%E8%BF%87%E5%AF%B9%E6%AF%94%20TensorRT,LLM%20%E8%BF%9B%E8%A1%8C%E6%8E%A8%E7%90%86%EF%BC%8C%E5%85%B6%E5%90%9E%E5%90%90%E9%87%8F%E7%9B%B8%E8%BE%83%E4%BA%8Ebaseline%E6%8F%90%E5%8D%87%E4%BA%86%E4%BA%94%E5%80%8D%E4%BB%A5%E4%B8%8A%E3%80%82%E8%BF%99%E7%9B%B8%E5%BD%93%E4%BA%8E%E5%B0%86%E9%83%A8%E7%BD%B2%E6%88%90%E6%9C%AC%E9%99%8D%E8%87%B3%E5%8E%9F%E6%9D%A5%E7%9A%84%E4%BA%94%E5%88%86%E4%B9%8B%E4%B8%80%E3%80%82) NVIDIA TensorRT-LLM 在推荐广告及搜索广告的生成式召回的加速实践 - NVIDIA 技术博客

<https://developer.nvidia.com/zh-cn/blog/nvidia-tensorrt-llm-ad-generation/>

[[9]](https://arxiv.org/abs/2507.00715#:~:text=Extensive%20experiments%20on%20three%20datasets%2C,effectiveness%20gap%20in%20LLMRec%2C%20offering) [2507.00715] EARN: Efficient Inference Acceleration for LLM-based Generative Recommendation by Register Tokens

<https://arxiv.org/abs/2507.00715>

[[12]](https://arxiv.org/abs/2402.15235#:~:text=%3E%20Abstract%3ALLM,Analyst%2C%20Reflector%2C%20Searcher%2C%20and%20Task) [2402.15235] MACRec: a Multi-Agent Collaboration Framework for Recommendation

<https://arxiv.org/abs/2402.15235>