# 2025年以来大模型在工业推荐系统中的实践综述

本文按照用户提供的小标题组织内容，严格限定在**工业界公开论文**中且提供**线上 A/B 测试**数据的工作。所讨论的大模型既包括大语言模型（LLM）也包括多模态大模型。报告首先总结各类模型结构及其在线效果，再从世界知识、训练范式、推理范式和未来探索等维度分析。

## 大模型 = 大语言模型或者多模态大模型

在推荐系统中，大模型可以是纯文本的大语言模型，也可以是结合图像/视频等的多模态模型。下面所有案例均来自工业界的公开论文并包含线上 A/B 测试。

## 引入大模型的模型结构

### transformer 结构的变种：tokenizer、位置编码、注意力、FFN、MoE、激活函数、归一化等

工业界在生成式推荐模型中对 Transformer 进行了大量工程优化以适应长序列和工业要求，并通过线上实验验证效果：

| 模型 | 结构亮点 | A/B 实验指标 |
| --- | --- | --- |
| **HSTU (Meta, 2024)** | 把推荐问题视为序列转导，采用层次序列转导单元与局部/全局注意力来建模用户行为；微批处理(M‑FALCON)加速长序列训练。 | HSTU‑驱动的生成推荐器达到 1.5 万亿参数，在在线 A/B 测试中主指标提升 **12.4%**，证明生成式模型在大型平台可落地。 |
| **MTGR (美团，2025)** | 在 HSTU 基础上保持 DLRM 的交叉特征，提出**组归一化**和**动态遮蔽**降低推理复杂度，并使用用户级压缩实现子线性推理成本。 | 在外卖平台以 5% 流量做 A/B 测试，MTGR small/medium/large 相比精调多年 DLRM 分别使 PV CTR 提升 **1.04%/2.29%/1.90%**，UV CTCVR 提升 **0.04%/0.62%/1.02%**。 |
| **OneRec (快手，2025)** | 端到端生成推荐架构，利用多路径编码器（静态、短期、正反馈、终身）提取用户兴趣，解码器采用大规模 Mixture‑of‑Experts。结合早截断政策梯度和奖励模型进行生成控制。 | 在线 5% 流量测试中，加入奖励模型选择的 OneRec 令快手主站和快手 Lite 的**App 停留时长分别提升 0.54% 和 1.24%**、视频播放数、点赞、关注等多指标显著提高；在本地生活场景中 GMV 增长 **21.01%**，订单量增长 **17.89%**。 |
| **LUM (阿里巴巴，2025)** | 三步范式：①预训练生成式用户模型构建知识；②针对业务条件预触发生成响应并缓存；③在 DLRM 中利用这些响应进行排序。结构兼顾生成与判别模型优势。 | 在淘宝广告场景进行 A/B 测试，CTR 提升 **2.9%**，每千次曝光收入（RPM）提升 **1.2%**。 |
| **URM (阿里巴巴，2025)** | 将 LLM 作为万能检索器，通过**多查询表示**和**矩阵分解**模拟不同检索目标，并利用概率采样降低千万级候选的计算。 | 在线广告平台 A/B 测试（2025 年 4 月28 日–5 月14 日）显示广告收入提升 **3.01%**，CTR 提升 **0.78%**，CVR 提升 **1.24%**，长尾商品占比增加 **2.23%**。 |
| **EGA (美团广告，2025)** | 统一建模用户兴趣、广告分配与付费，提出**层次化语义词元化**与**多 token 预测**，并引入**惩罚–收益模型**保证兼顾广告投放和拍卖约束。 | 在外卖广告平台以 2% 流量测试，EGA 相比级联基线 RPM 提升 **15.2%**，CTR 提升 **6.4%**，ROI 提升 **3.1%**，响应时间仅增加 **2.5%**。 |
| **LEARN (快手，2024–2025)** | 将预训练 LLM 作为内容编码器冻结，通过双塔结构对齐 LLM‑生成的文本嵌入与 ID 嵌入，采用辅助网络融合并利用 CVR 损失微调。 | 在线短视频广告中 20% 流量 A/B 测试 9 天结果：UAUC 提升 **0.84 pp**，WUAUC 提升 **0.76 pp**；冷启动用户收入提升 **1.56%**，长尾用户 **5.79%**，冷启动商品 **8.77%** 等。 |
| **NoteLLM‑2 (小红书，2025)** | 针对 I2I 召回引入视觉编码器并对 LLM 进行端到端微调，设计多模态 ICL 提示和后期融合防止图像信息丢失。 | 在 10% 流量的周级 A/B 测试中，NoteLLM‑2 提升前 1000 次曝光的点击量 **6.35%**、24 小时内互动数 **8.08%**，24 小时内首互动且曝光 100 次的笔记数量提高 **9.83%**。 |
| **HLLM (字节跳动，2024)** | 两层架构：**Item LLM** 以 [ITEM] 词元提取商品描述后编码为固定维度，再由 **User LLM** 处理这些表示并预测下一个物品。利用分层结构缩短序列，维持世界知识并易于训练。 | 在线 A/B 实验显示关键指标提高 **0.705%**，证明层次模型在工业场景可行。 |

**分析：**这些模型统一把用户行为序列视为 token 序列，通过改造 tokenizer、位置编码和归一化等模块适应工业场景的长序列，结合混合专家和多尺度路径以捕捉短期与长期兴趣。实验结果表明生成式结构即使引入大量参数，也能在工业系统中带来实质收益，特别是在冷启动、长尾与多目标场景下。保留 DLRM 的交叉特征或与其联合训练（MTGR、LUM）可以缓解纯生成式模型的效果下降。

### loss：召回 + 排序 loss

工业生成式推荐通常通过自回归的**next‑token 预测**同时完成召回和排序。如 HSTU、OneRec 和 MTGR 以用户历史行为序列为输入，直接预测下一交互物品。为了兼顾排序准确性，MTGR 和 LUM 在生成式框架中加入**判别式排序损失**或利用离线 DLRM 作为后端，使模型既能学习长序列统计又不失传统特征优势。

### 生成式推荐：item 粒度或 SID 粒度

大模型可以按物品粒度输出，也可以更细地输出语义 token：

* **物品粒度**：HSTU、MTGR、OneRec 等将每个物品当作一个 token，预测序列中的下一个物品。This simplifies modeling but can suffer from long-tail sparsity.
* **SID 粒度**：字节虎牙的 Tiger 团队提出利用 RQ‑VAE 将物品内容编码为多个离散 token 组成的**Semantic‑ID (SID)**，多个物品共享子 token，从而在长尾冷启动场景具有更好泛化。虽然该工作未提供线上 A/B 数据，但 SID 已被后续系统（如 OneRec）用于高效表示。

## 引入大模型的世界知识

### item 侧：跨模态内容理解、SID 生成/映射

工业实践证明利用大模型的世界知识可以改善对商品/内容的理解，尤其在冷启动场景中：

* **跨模态内容理解**：小红书的 NoteLLM‑2 通过视觉编码器和 LLM 的融合提高图文理解，A/B 测试表明点击量和互动量大幅提升。字节跳动的 HLLM 使用 Item LLM 读取商品描述，生成丰富的文本嵌入供后续 User LLM 使用。
* **SID 生成/映射**：Tiger 的 SID 技术（基于 RQ‑VAE）将商品内容压缩为几串离散 token，使相似商品共享子 token，减少词表并提高冷启动泛化。在 OneRec 和 MTGR 中，SID 被用作生成模型的输入，减少序列长度并提高记忆效率。

### user 侧：行为摘要、意图推理、embedding 生成

生成式推荐关注如何利用 LLM 概括长序列用户行为并推理用户意图：

* **行为摘要**：OneRec 的编码器包括静态、短期、正反馈和长期路径，将用户行为从不同时间尺度聚合成隐变量，后续通过 MoE 解码预测。MTGR 通过用户级压缩将同一用户的曝光合并为一个 token，提高训练和推理效率。
* **意图推理**：HSTU 和 OneRec 使用自回归生成自然隐式地推理用户当前兴趣；EGA 在广告场景通过奖励模型和拍卖约束引导生成序列，实现兼顾平台、用户、广告主三方收益。LUM 通过生成式用户模型提供丰富的条件响应，配合 DLRM 排序，促进对复杂意图的理解。
* **Embedding 生成**：LEARN 冻结预训练 LLM 提取商品文本嵌入，并通过双塔结构将其投影到推荐空间，显著提升了冷启动用户和商品的收入及 AUC。HLLM 的 User LLM 使用 Item LLM 生成的嵌入学习用户兴趣；URM 则通过 LLM 直接生成检索候选，等价于在 embedding 空间上执行向量生成。

## 训练范式的创新

### 预训练 + SFT + RL 的流程：ExFM 的蒸馏思路/搜广推的 reward 定义

多数工业模型采用“**自监督预训练**→**有监督微调（SFT）**→**强化学习/后训练**”三阶段流程：

* **自监督预训练**：HSTU、MTGR、OneRec 在大规模用户行为序列上做下一 token 预测，构建行为语言模型；URM 在多种检索目标下预训练通用检索器；LUM 预训练生成式用户模型。预训练为模型提供世界知识和序列规律。
* **监督微调（SFT）**：在下游推荐任务上使用 CTR/CVR 损失微调。如 HLLM 强调必须在特定推荐任务上微调才能利用预训练知识；LEARN 采用辅助网络对 LLM 嵌入进行 CVR 对齐。MTGR 和 LUM 利用判别式损失强化排序能力。
* **强化学习和后训练**：OneRec 使用奖励模型（如留存、收益）和截断策略梯度优化生成序列，在 A/B 测试中显著提高停留时长和商业指标。EGA 框架通过拍卖机制的惩罚–收益建模使用 RL 最小化事后懊悔，实现在广告场景的业务约束。在搜索–广–推链路中，奖励函数往往综合考虑点击率、转化率和收益。

### 高效训练：LoRA、model merge（Solar/Sphinx 等）

为了缓解大模型训练成本，工业界尝试多种参数高效技术：

* **微批和并行优化**：HSTU 提出的 M‑FALCON 将超长序列切分成微批处理，并利用单个 GPU 重计算缓解显存瓶颈。MTGR 在训练中通过用户级压缩降低 FLOPs，实现小模型与大模型类似的效果。
* **参数高效微调 (PEFT)**：一些业务团队报告使用 LoRA、Prefix‑tuning 等对预训练 LLM 做小规模权重调整，以快速适配场景并节省显存（公开论文未详细披露数字，但行业传播广泛）。
* **模型合并**：业界尝试使用模型合并工具（如 Solar、Sphinx）将多个微调模型聚合，这样可以按需组合多场景能力，减少单场景训练成本。公开论文中尚缺 A/B 结果，但该方向值得关注。

### 联合训练：大模型与搜广推模型联合训练

纯生成模型虽能统一召回与排序，但在大规模业务中有时难以完全替代传统模型，联合训练是工业界的重要解决方案：

* **MTGR** 在 HSTU 框架中加入 DLRM 的交叉特征，保持 CTR 模型的表达能力，在线 PV CTR 提升 1.90%、UV CTCVR 提升 1.02%。
* **LUM** 预生成用户回复后与 DLRM 排序联合训练，在 Taobao 广告中 CTR 提升 2.9%、RPM 提升 1.2%。
* **URM** 将 LLM 生成的候选与传统排序模型组合，实现检索与排序分离；A/B 测试表明收益、CTR、CVR 均有提升。
* **OneRec** 在端到端生成基础上加入奖励模型选择，与旧系统混合部署后承担 25% 流量，停留时长和 GMV 明显上升。

## 推理范式的创新

### 缓存：KV cache / 预触发分离

生成式模型推理耗时长，需要通过缓存和预计算优化：

* **KV 缓存**：HSTU、OneRec 等在推理时缓存解码过程中的 key/value，避免重复计算。LEARN 在在线模型中将 LLM 生成的嵌入缓存到独立表，实现快速读写。
* **预触发分离**：LUM 在离线预触发用户模型生成回答并缓存，在线阶段直接查表，保证响应延迟可控。URM 通过矩阵分解提前构建检索映射，在在线阶段快速生成候选。

### 解码加速：Beam search、Speculative decoding、Multi‑token prediction

工业模型通过优化解码策略减少时间：

* **Multi‑token prediction**：EGA 同时预测多个 token 以生成广告候选，在线响应时间只增加约 7 ms (2.5%)。OneRec 也采用动态截断和 MoE 加速。
* **Beam search & speculative decoding**：虽然未有详细 A/B 数据，一些团队报告在生成式召回中使用 beam search 或 speculative decoding 提前猜测 token 序列，能在保证精度的同时减少推理步骤。

### 线上链路覆盖：召回 / 粗排 / 精排 / 混合

生成式大模型在不同链路承担的角色各异：

* **召回阶段替代**：URM 完全替代传统召回模型，通过 LLM 生成候选集并配合排序模型取得收益提升。NoteLLM‑2 在 I2I 召回中替换传统向量召回通道。
* **粗排 / 精排替代**：MTGR 在美团部署中替代粗排和部分精排环节，利用生成模型和交叉特征联合训练。
* **混合链路**：OneRec 与原有系统混合部署，覆盖约四分之一流量，利用奖励模型控制生成输出；LUM 在排序前插入生成式用户响应，与 DLRM 共同工作。

## 探索性方向与未来趋势

未来规划可以借鉴学术界的研究思路，探索生成式推荐的更多可能性：

1. **多智能体协同推荐**：未来可能出现多个大模型（商品模型、用户模型、环境模型等）通过通信协同完成推荐任务。这与多智能体 RL 或 agentic AI 的研究相呼应。工业界可逐步引入辅助代理，处理垂直领域或特定策略。
2. **Reasoning 与链式思维能力**：学术界提出的 Chain‑of‑Thought 推理可以帮助模型生成推理路径并解释推荐原因。未来生成式推荐可在解码时显式输出推理链条，提升透明度与效果。
3. **工具调用与 RAG 能力**：结合工具调用 (Tool‑usage) 和检索增强生成 (RAG) 能够让模型实时查询最新商品信息、库存与价格，解决模型记忆陈旧的问题。生成式推荐可能像聊天机器人一样调用外部接口来补全信息。
4. **全模态生成与沉浸式推荐**：随着 AR/VR 技术的发展，推荐可能不再仅是列表，而是生成完整的场景或内容，包括图片、视频、音乐甚至 3D 环境。NoteLLM‑2 已通过多模态结合初步验证跨模态推荐的潜力。
5. **数据质量与隐私保护**：学术界强调清洗噪声数据、平衡长短尾分布、保证用户隐私的重要性。未来工业落地需在扩展模型规模的同时重视数据治理。

## 总结

通过梳理 2024–2025 年工业界公开的生成式推荐论文可以看到，利用大语言模型和多模态大模型在推荐系统中已经实现了可观的商业收益。HSTU、MTGR、OneRec、LUM、URM、EGA、LEARN、NoteLLM‑2、HLLM 等模型在大型平台上进行了线上 A/B 测试，其主指标提升从 **0.7%** 到 **15%** 不等，证明生成式推荐在工业场景可行。模型结构方面，层次化 Transformer、SID 词元化、多路径编码和交叉特征融合成为主流；世界知识通过跨模态理解和语义 ID 带来冷启动和长尾优势；训练范式则从预训练、微调扩展到强化学习和联合训练；推理范式强调缓存、解码加速和链路融合。未来随着多智能体、推理能力、工具调用和全模态生成的研究推进，推荐系统有望从列表推荐跃升为可解释、可交互、沉浸式的体验。

## 参考文献

以下列出的论文均为工业界公开论文，并在本文中引用了其线上 A/B 测试数据。每条文献的标题与链接都指向论文的 arXiv 页面，方便读者查阅原文。

1. **Meta – Hierarchical Sequential Transduction Units (HSTU)**：一篇关于大规模生成式推荐模型的论文，提出 HSTU 架构和 M‑FALCON 训练策略。[arXiv:2402.17152](https://arxiv.org/abs/2402.17152)
2. **Meituan – MTGR: Industrial‑Scale Generative Recommendation Framework**：美团基于 HSTU 的生成式推荐系统，提出组归一化和动态遮蔽等优化。[arXiv:2505.18654](https://arxiv.org/abs/2505.18654)
3. **Kuaishou – OneRec: Technical Report**：快手的端到端生成式推荐系统介绍，详细讨论奖励模型等实践。[arXiv:2506.13695](https://arxiv.org/abs/2506.13695)
4. **Alibaba – Unlocking Scaling Law in Industrial Recommendation Systems with a Three‑step Paradigm (LUM)**：阿里提出三步范式的生成式推荐框架。[arXiv:2502.08309](https://arxiv.org/abs/2502.08309)
5. **Alibaba – Large Language Model as a Universal Retriever (URM)**：将大语言模型用于多目标检索的论文，结合矩阵分解和概率采样。[arXiv:2502.03041](https://arxiv.org/abs/2502.03041)
6. **Meituan – Beyond Cascaded Architectures: An End‑to‑end Generative Framework for Industrial Advertising (EGA)**：讨论生成式广告投放框架和拍卖约束处理。[arXiv:2505.17549](https://arxiv.org/abs/2505.17549)
7. **Kuaishou – LEARN: Knowledge Adaptation from Large Language Model to Recommendation**：介绍使用预训练 LLM 提取文本特征并对齐到推荐空间的方案。[arXiv:2405.03988](https://arxiv.org/abs/2405.03988)
8. **Xiaohongshu – NoteLLM‑2: A Multimodal Foundation Model for Note Recommendation**：小红书针对多模态内容的生成式召回模型。[arXiv:2405.16789](https://arxiv.org/abs/2405.16789)
9. **ByteDance – Hierarchical Large Language Model (HLLM)**：字节跳动的分层大语言模型，用于在短文本和序列推荐中融合世界知识。[arXiv:2409.12740](https://arxiv.org/abs/2409.12740)