# 大模型在推荐系统中的应用与进展

## 大模型架构创新

当今工业级推荐系统开始引入大模型（大语言模型及多模态大模型）的先进架构，以提升模型容量和表现。**Transformer结构变种**在推荐场景中得到探索，包括改进的Tokenizer、位置编码、注意力机制、前馈网络和归一化方法等。例如，大模型常采用子词级别的分词器（如SentencePiece）来处理物品或内容文本，这甚至被用于生成**“语义ID”**序列以表示物品。这种基于内容的离散ID能够将**相似物品映射到相近表示**，缓解传统ID冷启动问题。在位置编码上，引入RoPE旋转位置嵌入等新方法使模型更好地泛化长序列。在注意力机制方面，工业界尝试了**稀疏激活的Mixture-of-Experts (MoE)**架构来扩展参数规模，比如在多模态模型中引入专家路由以实现**万亿级参数**的稀疏激活。这类架构扩展方法被视为unlock模型性能上限的关键。Meta最近提出的ExFM框架即采用**外部大模型+蒸馏**的思路，将一个超大教师模型的知识外部蒸馏给多个下游推荐学生模型，从而在不大幅增加推理成本的情况下获得性能提升。ExFM的**教师模型规模达数万亿参数**，服务于广告和推荐等多个垂类学生模型，以摊薄训练成本；同时通过辅助头和适配器缓解数据流动态分布造成的偏差。这一方法在内部大规模应用和公共数据集上**显著提升了推荐效果**。此外，**Stacked Factorization Machines**等新颖网络结构也被提出以实现推荐模型的可扩展性规律。例如Meta的Wukong架构纯粹堆叠因子分解机单元并联合宽和深度扩展模型，被证实可以建立**推荐领域的Scaling Law**：随着模型更“高、更宽”，性能持续提升并超越SOTA模型。Wukong在内部百亿规模数据集上的实验显示，其在模型复杂度扩大两个数量级（>100 GFLOPs/样本）时依然保持性能优越且遵循可预测的性能增长曲线。

大模型架构的各组件也针对推荐任务进行优化调整。例如**激活函数**从ReLU改为GELU/SwiGLU等以捕获非线性特征，**归一化**层采用LayerNorm对长序列稳定训练。此外模型也在探索**多任务与多目标架构**：通过设计统一的输入/输出格式，以一个大模型**同时支持召回、粗排、精排甚至多种交互预估任务**。研究表明，大语言模型具备作为**“通用推荐学习器”**的潜力，可在统一框架下处理多种推荐任务而无需为每个任务定制模型。为提升这类通用模型的性能，学者提出了**多模态融合模块**来整合物品的文本、图像等信息，以及“**序列输入-集合输出**”的生成方式来高效地产出候选集合。在某头部电商的工业数据上，这种LLM式推荐模型已达到与精调的任务专家模型相当的效果。不过，由于文本生成式推荐效率和准确率仍有挑战，实际应用往往结合**多任务学习**策略，在一个模型中**共享和区分参数**来兼顾不同目标。例如在新闻推荐中，学者提出**统一模型 UniRec**同时执行召回和排序，用单一的用户表示来生成候选集并做细排，从而**大幅降低了维护多模型的开销**。UniRec通过从用户历史点击学习用户embedding用于排序，再基于该embedding对预训练的**基础兴趣向量集合**做注意力组合，生成兼顾广度的召回embedding。在线下数据上，这种统一方法在效率和效果上均取得提升，被认为有望简化线上多阶段链路。总体而言，大模型架构的引入为推荐系统提供了**更强大的表示学习和建模能力**，从底层组件到整体范式都在演进，以适配推荐领域的规模和实时性需求。

## 融合大模型的世界知识

引入大模型的一个重要价值在于**融入世界知识**，即利用模型对内容和用户的丰富理解来改进推荐。**物品侧**，大模型可以基于物品的跨模态内容（如标题、描述、图像、音频等）学习到**语义丰富的表示**，弥补了传统ID向量的冷启动弱点。例如，在YouTube视频推荐中，Google团队用视频的文本描述和其他内容生成“**语义ID**”，代替随机ID嵌入。通过对每个视频的内容向量进行**残差量化编码**，得到离散代码序列（语义ID），并用LLM常用的SentencePiece对其分片处理。这样，每个视频被表示为若干语义token，使得**相似内容的视频具有相似token序列**。在线实验表明，用语义ID取代原始视频ID后，新上传和长尾视频的泛化推荐能力显著提升，而整体效果不下降。这证明了融合跨模态内容所蕴含的知识可以**提升对冷启动内容的召回和匹配**。另一方向，大模型可以根据商品或媒体内容自动生成**标签、摘要或embedding**，用于丰富物品画像和索引。例如Pinterest等平台使用图像-文本预训练模型理解Pins的视觉内容，赋予语义标签，增强了推荐的准确性（相关工作报道在A/B测试中提高了冷启动Pin的点击率）。同时，多模态大模型（如Vision-LLM融合模型）能够实现**跨模态推理**，理解商品图像与评论文本的一致性，从而过滤不符的推荐。前沿工作SPHINX就展示了通过**融合多种视觉编码器权重和视觉嵌入**到LLM中，可提升模型对不同视觉任务的泛化能力。这预示未来的推荐系统或将拥有一个对图文声多模态内容都深刻理解的“通识大模型”来统一表示物品。

在**用户侧**，大模型的世界知识体现为对用户行为的高层次理解与推理能力。传统推荐系统主要用ID向量或简单统计特征表示用户，而引入LLM则可以对用户历史行为序列进行**自然语言式的总结和解释**。例如，研究者尝试把用户看过的商品名转换成一段描述用户偏好的文本，然后用LLM来生成推荐。这种方法曾在小规模实验中展示了可行性，但也发现**直接将行为转成文本输入LLM往往缺乏判别力**，且长序列逐词生成在实时性上不切实际。为解决这些问题，Google提出了**“User-LLM”框架**：先用一个Transformer Encoder将用户的海量历史交互压缩为一个紧凑的**用户embedding**，该embedding融合了多种行为（浏览、点击、评分等）的模式和偏好，然后在LLM推理时通过**Cross-Attention或软Prompt**把此用户embedding融合到LLM中。这样，LLM不需要处理长行为序列文本，而是直接获得浓缩的用户画像。例如在美食推荐场景下，User-LLM把用户过去一年的餐厅访问记录编码成embedding提供给LLM，结果LLM能更准确地理解用户口味并生成个性化回应。**多数据集的大量实验**验证了这种方法的有效性：在需要深刻用户理解和长行为序列建模的任务上，注入用户embedding使LLM性能**大幅提升**。特别是在下一个物品预测这类任务中，User-LLM相较不使用LLM的方法取得**最高的Recall@1**，并在偏好类别预测和评论生成等任务上表现出色。更难能可贵的是，这种融合方法**计算开销可控**，因为embedding长度远小于原始序列长度，使LLM推理开销降低，同时也**保持了LLM原有的知识**。除了embedding融合，LLM还可用于**推理用户意图**：例如电商场景下输入用户最近浏览和搜索记录，让大模型生成对用户当前需求的推测（如“正在为家庭影院寻找性价比高的投影仪”），再据此匹配商品。此类对话式推荐代理已在一些研究中出现雏形。总的来说，借助大模型对内容和用户的丰富知识，推荐系统开始具备**理解语义、推理意图、解释行为**的能力，从“相关性匹配”走向“智能决策”。线上实验已经表明，将大模型世界知识融入推荐可带来切实收益，例如YouTube用内容语义ID增强模型在长尾视频上的表现，又如某头部外卖平台通过大模型微调技术捕捉区域时段变化的用户兴趣，实现首页推荐CTR和交易额提升。

## 训练范式的创新

在训练范式上，引入大模型催生了**预训练+微调+强化学习**的三段式流程在推荐领域的探索。首先，大模型可以通过大规模无监督预训练（如语言模型预测下一个物品、掩码预测属性等）获取通用知识；然后经过有监督微调（SFT）学习特定推荐任务；最后可能结合强化学习（RL）进一步优化业务目标。近期研究揭示了直接用传统有监督方式微调LLM做推荐的局限：例如在下一地点预测中，用SFT训练LLM去输出**唯一的下一个POI**存在不匹配，因为每个训练样本只有一个实际下一个地点，无法训练模型生成**Top-K列表**。对此，清华等提出了**Refine-POI**框架，使用强化学习微调LLM的生成策略。他们为LLM设计了**与推荐指标挂钩的奖励函数**（如对生成的Top-K列表计算命中率类奖励），让模型在只有单一点真实反馈的情况下学会产出优化整体排名的列表。实验显示，该方法使LLM在Top-K推荐准确率上达到新的SOTA水平。这表明RL（类RLHF）的引入可以弥补LLM生成式推荐和精排目标的不一致，提高列表质量。在工业实践中，也有公司尝试利用**用户行为反馈作为奖励**，通过策略梯度方法优化推荐模型的长期指标。例如字节跳动曾报道用策略价值网络优化用户时长和交互多样性，美团等也探索过强化学习来均衡短期点击和长期留存。

另一大方向是**模型蒸馏和知识迁移**，以应对大模型训练和部署的高成本。Meta的ExFM就是工业界的成功案例：它将超大规模的“外部基础模型”作为教师，持续蒸馏提升线上较小的广告推荐模型。这样，学生模型在保持低延迟的同时吸收了教师大模型在多轮训练中获得的新知识。ExFM还包含**数据增强服务**，持续将新出现的query/ad加入训练，结合辅助头和适配器解决流式数据分布漂移。在Meta内部广告系统的测试中，ExFM带来了**显著性能增益**，并成功通过了WWW2025工业组A/B测试的验证。类似地，外卖平台Ele.me（饿了么）在WWW2025上报告了一种信息瓶颈理论指导的**预训练推荐模型微调方法**（IAK）。他们将工业推荐系统看作预训练模型，提出在微调阶段显式分两步“知识压缩”和“知识匹配”，用特殊的核函数适配微调。该方法已在数亿用户规模的外卖首页上线数月，取得了**可观的业务利润提升**。这证明即使不额外扩增参数，仅通过**精巧的训练范式**设计，也能大幅提升成熟推荐系统的性能。

**高效微调**也是训练范式创新的重点。在面对百亿级参数的基础模型时，全量fine-tune代价高昂且可能过拟合。因而业界大量采用参数高效微调(PEFT)技术，如**LoRA (Low-Rank Adaptation)**、Prompt Tuning等。LoRA通过**冻结原模型权重，仅低秩更新部分权重矩阵**来训练，大幅减少了需要更新的参数。在推荐场景，将LoRA应用于预训练LLM，可以在数小时内完成对新业务数据的微调，并方便地部署在现有大模型推理服务上。业内报告显示，相比全量微调，LoRA能以不到10%的训练成本达到近似的效果，是在推荐系统中**定制大模型**的实用策略。除了微调，近期出现的**模型权重融合**方法也吸引关注。研究者发现，可以直接对不同任务或数据上训练的模型进行**权重平均或合并**，获得兼具多方能力的模型而无需额外训练。例如Solar项目中，将两个在不同数据集上instruction-tuning的中型LLM按权重平均融合，得到的新模型在综合基准上超过了各自单独微调的性能。这一现象也被用于推荐场景：如把一个注重用户短期兴趣微调的模型与一个注重长期偏好微调的模型合并，可望同时顾及短长期效果。据报告，简单的**权重平均**就使某些评测指标提升了1-2个百分点。当然，更复杂的融合（如按层或按模块选择性融合）也在探索中。多模态模型SPHINX进一步展示了**融合不同训练阶段的模型权重与视觉特征**，可以提升模型综合能力。总之，模型微调范式正变得更加灵活高效，通过**蒸馏、迁移、低秩微调和模型融合**等手段，让引入大模型的推荐系统在可用成本内不断逼近性能上限。

值得一提的是，业界也开始尝试**联合训练推荐大模型与现有召回/排序模型**，形成端到端优化的新范式。一种思路是将大模型作为**上游生成器**，产出候选内容，再由下游精排模型打分，整个过程通过可微分的近似目标联合训练，以最大化最终评价指标。另一种思路是利用大模型生成的伪数据或监督信号来训练现有模型（如用LLM判断推荐结果好坏作为标签，训练一个轻量模型实时判断)。LinkedIn 的“eBadMatch”就是这样一个案例：先用GPT-4评估求职推荐匹配的好坏，再训练小模型模仿GPT判断，以过滤低质推荐，结果线上A/B显著降低了用户拒绝率并提高了申请转化。这类**人机协同优化**的训练范式为推荐系统注入了新的活力。

## 推理范式的创新

在推理阶段，引入大模型也要求对**在线推理范式**进行创新，以平衡性能与效率。首先是**缓存和复用**机制的引入。由于大模型推理开销高，**KV Cache**（键值缓存）技术被广泛应用于自回归生成任务，以避免重复计算已有内容的表示。当多个推理请求拥有相同的前缀时（例如相同的系统提示或公共上下文），服务架构可以复用已计算的KV缓存，加速后续生成。例如在对话推荐场景，系统提示和用户长历史可作为静态前缀缓存，针对不同的当前query复用，大幅降低首个推荐结果的延迟。NVIDIA的TensorRT-LLM和vLLM等开源加速器都支持**前缀缓存重用**，据报道开启该功能后**首字节延迟**显著降低，而生成结果完全一致。为了进一步利用缓存，有研究提出将输入拆分为**Prompt（静态上下文）和Dynamic部分**分离编码：静态部分（如长历史、说明文档）提前离线编码，其KV缓存独立存储；仅动态部分（如实时query）在线编码并与静态缓存拼接。这种“前缀-动态”分离让大段上下文的代价摊薄，使得推荐Agent可以高效利用**长会话历史和知识文档**而不显著增加每次响应时间。

其次，在**生成解码加速**方面，业界引入了多种创新技术，借鉴自NLP领域以缓解逐token生成的延迟问题。**Beam Search**等方法可以一次探索生成多个可能序列，取最优者输出，从而在给定步数内获得更佳结果（在推荐场景常用于生成多个备选推荐列表以供选择）。更具突破性的是**Speculative Decoding**和**多token并行解码**等方法。在Speculative Decoding中，引入一个小型“草稿模型”并行运行：它快速生成若干候选下一步tokens，大模型则对这些候选并行验证。如果草稿tokens通过验证，就批量接受，从而**一次大模型计算得到多个输出**；不通过则丢弃重采但也不浪费过多算力。OpenAI报道这种**“先赌后证”**的方法可在不损失回答质量的前提下将GPT系列模型推理速度提升2-3倍。多模态场景下也有类似工作将Speculative Decoding用于图文生成，加速效果同样显著。另一方向，Apple最新研究表明大模型其实**隐含预测多个后续token的能力**。他们设计了**多token并行生成框架**（MTP）：通过在Prompt中插入特殊占位符，让模型一次性填充多个词，再逐一验证纠正。在一个8B参数模型上实验显示，同时生成8个token可平均获得2-3倍加速，在代码和数学这类可预测性高的任务上甚至达到5倍提速，而且生成质量**没有任何下降**。这是通过一种称为**“门控LoRA适配”**的策略来保证多token猜测不影响精度。有了这些技术，大模型用于实时推荐的延迟瓶颈有望被极大缓解——例如用户刷新一个推荐Feed时，LLM可以**成批地产生下一批推荐**而不是逐条生成，从而在严格毫秒级延迟内完成个性化生成推荐列表。

第三，**线上推荐链路的整合与覆盖**成为新的探索方向。传统推荐系统往往分为多阶：召回、粗排、精排、混排，每层模型各司其职。引入大模型后，一种趋势是让大模型**在单次推理中完成多阶段工作**，或者**作为其中某阶段的统一接口**。例如，大模型可根据用户当前意图直接**用自然语言检索**出候选（相当于召回作用），再对这些候选进行顺序排列输出（相当于排序），实现“一步到位”的生成式推荐。在学术研究中，已有工作让LLM通过Prompt得到用户历史和意图描述，然后要求模型输出Top-N物品ID列表。这种生成式推荐可以看作LLM隐式地执行了召回与排序。在实践中，全盘交给LLM可能难以满足效率要求，因此也有探索**混合链路**：LLM与向量召回、DNN排序协同工作。例如Spotify的查询推荐系统中，LLM被用于生成**拓展查询**来补充传统精确匹配结果，再由下游ranker综合排序，两者结合带来了用户更长查询和更多探索性意图的提升。又如Amazon音乐的歌单搜索，LLM生成歌单描述和合成训练查询用于改进召回，然后再用轻量模型排序，最终大幅提高了冷门歌单的召回率。可以预见，未来推荐系统的线上链路将更加灵活：对于热门head内容，仍可用高效的向量索引召回；对于长尾或复杂意图，则调动LLM进行**语义级的理解和生成**；最后由精排模型或LLM自身（结合评分策略）给出最终结果。这种**多智能体/多模型共存**的混合链路在最近一些探索中已现雏形，其目标是在严格的延迟预算内，将大模型智能与传统精细模型的长处结合，全面覆盖推荐流程中的各个环节。

## 探索方向与未来趋势

展望未来，大模型在推荐领域还有诸多令人兴奋的探索方向：

* **多智能体协同推荐**：利用多个LLM代理协作完成推荐任务，例如一个“用户代理”模拟用户与另一个“推荐代理”对话，用户代理提供真实的偏好反馈，推荐代理据此调整推荐策略。这种**多Agent对话框架**被认为有助于更真实地模拟用户-系统交互，解决复杂的偏好获取和多轮推荐问题。一项多智能体推荐研究（MACRec）提出让若干智能体分别扮演推荐系统的不同组件（召回、解释、策略等），通过消息传递共同决定最终推荐。初步结果表明，LLM代理能够通过自然语言协商来改进推荐结果，未来或可用于复杂场景下的**自动推荐策略优化**。
* **强化的推理与链式思维**：大模型天然具备一定的**Reasoning**能力，如Chain-of-Thought链式推理，这在推荐中可用于更好地理解用户需求背后的原因，从而给出有依据的推荐。未来的推荐LLM可能会在生成答案前列出自己的思考路径，例如“用户最近浏览了登山背包和徒步鞋 -> 推断TA可能计划登山 -> 推荐登山相关物品”。这种**可解释推荐**不仅提升用户信任，也有助于模型本身校准。LLM提供**可解释的原因描述**已在对话推荐和评论摘要中有所尝试，一项调查指出LLM Agent有望通过**可解释的自然语言交互**彻底改变推荐系统。
* **工具调用与检索增强生成 (RAG)**：由于单纯的LLM无法包含最新的世界知识，将**工具使用**能力赋予推荐大模型成为趋势。具体体现在：LLM在推理过程中可以自动调用检索模块、知识库查询、计算器等工具以获取即时信息。例如在电影推荐场景，LLM可调用搜索API获取最新上映电影列表或查询某演员的作品，然后综合结果给出推荐。这类似于在对话系统中的Toolformer思想。检索增强生成（RAG）技术可以帮助LLM克服知识时效和领域专精问题。相比反复fine-tune大模型去记忆业务最新数据，RAG允许**实时查询更新的数据库**，不会遇到遗忘旧知识的问题。一项研究将**推荐系统技术用于提高LLM检索提示的多样性**，把挑选少数示例提示的过程视为推荐问题以改进LLM在Few-shot学习中的效果。可以预见，未来的推荐大模型将像“管家”一样，掌握一系列工具：向量检索库、知识图谱、内容生成器等，并通过推理决定何时调用何种工具，为用户提供准确而丰富的推荐服务。
* **全模态生成与沉浸式推荐**：随着多模态大模型的成熟，推荐系统将超越传统的列表式、文字式呈现，走向**沉浸式、多模态**的体验。未来的系统也许不仅告诉用户“这本书你可能喜欢”，还**生成一段视频**摘要这本书的精彩内容，或者在AR眼镜中直接呈现用户房间摆放新家具的虚拟效果。这需要推荐模型能够**生成图像、音频、视频等内容**。当前GPT-4V等多模态模型已经展现了根据文本描述生成图像、分析视觉内容的能力，研究者也在尝试将其用于推荐场景，比如根据用户衣橱照片生成搭配建议的图像。虽然这些仍在早期，但技术趋势非常明确：推荐将从单纯的信息过滤，扩展为**内容创造与场景融合**。通过大模型强大的生成能力，用户能够“体验”推荐——如试听一段音乐、试读几页电子书、甚至让虚拟形象试穿衣服。这种沉浸式推荐有望极大提高用户参与度和决策信心。近期的一些多模态模型（如SPHINX-X）已经在**跨领域多模态理解和推理**上取得突破。它们展现出在不同模态和任务上**规模化提升性能**的趋势，这为全模态生成式推荐打下了模型基础。

综上，大模型+推荐系统正处于蓬勃发展的起点。从模型结构、训练推理范式到未来方向，都出现了丰富的创新。工业界的实践（Google、Meta、Pinterest、抖音字节、阿里、百度、美团、快手等）和顶会论文共同推动着这一浪潮。例如，YouTube语义ID、Meta ExFM、饿了么IAK微调、LinkedIn LLM过滤、Spotify/亚马逊的LLM增强检索等**实战方案**已经通过线上A/B测试验证了效果。可以预见，随着基础模型能力的提升和针对性改进，**大模型将在推荐系统中扮演越来越核心的角色**：不再仅仅是一个召回或排序模型，而是成为贯穿用户交互、内容理解和决策反馈的智能中枢，为用户带来更贴心、更聪明的推荐服务。

**参考文献：**

1. Zhenhao Jiang et al. *“Pre-train and Fine-tune: Recommenders as Large Models.”* WWW 2025.
2. Mingfu Liang et al. *“External Large Foundation Model (ExFM): How to Efficiently Serve Trillions of Parameters for Online Ads Recommendation.”* WWW 2025 (Industrial Track).
3. Buyun Zhang et al. *“Wukong: Towards a Scaling Law for Large-Scale Recommendation.”* ICML 2024.
4. Anima Singh et al. *“Better Generalization with Semantic IDs: A Case Study in Ranking for Recommendations.”* arXiv 2024 (YouTube team).
5. Xiangcheng Cao et al. *“TIGER: Transformer Index for Generative Recommenders.”* NeurIPS 2023.
6. Luyang Liu & Lin Ning. *“USER-LLM: Efficient LLM Contextualization with User Embeddings.”* Google Research Blog, May 2024.
7. Peibo Li et al. *“Refine-POI: Reinforcement Fine-Tuned LLMs for Next POI Recommendation.”* arXiv Jun 2025.
8. Luv Bansal. *“Speculative Decoding — Make LLM Inference Faster.”* Medium blog, Apr 2024.
9. Marcus Mendes. *“Apple taught an LLM to predict tokens up to 5x faster in math and coding tasks.”* 9to5Mac, Aug 2025 (summarizing Apple ARXIV work)
10. Chuhan Wu et al. *“Two Birds with One Stone: Unified Model Learning for Both Recall and Ranking in News Recommendation (UniRec).”* ACL 2022.
11. Qiyao Peng et al. *“A Survey on LLM-powered Agents for Recommender Systems.”* arXiv Feb 2025.
12. Emile Contal & Garrin McGoldrick. *“RAGSys: Item-Cold-Start Recommender as a RAG System.”* SIGIR 2024 Workshop on IR+RAG.
13. Eugene Yan. *“Improving Recommendation Systems & Search in the Age of LLMs.”* Blog, Oct 2023 (summarizing multiple industry case studies)