智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

# 生成式推荐系统综述

石磊, 赵雨秋, 袁瑞萍, 钟岩, 刘艳超

引用本文:
石磊, 赵雨秋, 袁瑞萍, 等. 生成式推荐系统综述[J]. 智能系统学报, 2026, 21(1): 19-40.
SHI Lei, ZHAO Yuqiu, YUAN Ruiping, et al. A survey of generative recommender systems[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(1): 19-40.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202505006

**DOI**: 10.11992/tis.202505006
**网络出版地址**: https://link.cnki.net/urlid/23.1538.TP.20250813.1631.002

# 生成式推荐系统综述

石磊[1], 赵雨秋[1], 袁瑞萍[2], 钟岩[3], 刘艳超[1]
(1. 中国传媒大学 媒体融合与传播国家重点实验室, 北京 100024; 2. 北京物资学院 计算机与人工智能学院, 北京 101149; 3. 北京大学 数学科学学院, 北京 100871)

**摘　要:** 随着社交媒体内容规模的急剧增长, 传统协同过滤推荐系统在数据稀疏性和冷启动等方面的局限性日益凸显。近年来, 生成式模型强大的数据特征分析与内容生成能力, 为推荐系统带来新的发展机遇。本文系统性地综述了生成式推荐系统的技术框架与研究进展, 重点阐述了生成式推荐系统的特征标记方法、核心模型架构、主流评估方案以及典型的应用场景。通过对比分析与文献研究, 论证了生成式推荐系统在推荐准确性、个性化和场景适应性等方面的显著优势。最后, 本文深入探讨了当前研究面临的关键挑战, 包括计算资源消耗、隐私安全风险以及评估标准统一性等问题, 并对未来研究方向提出建设性展望, 为突破生成式推荐系统的认知瓶颈提出了创新性视角。

**关键词:** 推荐系统; 生成式模型; 大语言模型; 特征标记; 表示学习; 模型架构; 协同信息; 评估方法
**中图分类号:** TP301　　**文献标志码:** A　　**文章编号:** 1673−4785(2026)01−0019−22

**中文引用格式: 石磊, 赵雨秋, 袁瑞萍, 等. 生成式推荐系统综述 [J]. 智能系统学报, 2026, 21(1): 19–40.**
**英文引用格式: SHI Lei, ZHAO Yuqiu, YUAN Ruiping, et al. A survey of generative recommender systems[J]. CAAI transactions on intelligent systems, 2026, 21(1): 19–40.**

# A survey of generative recommender systems

SHI Lei[1], ZHAO Yuqiu[1], YUAN Ruiping[2], ZHONG Yan[3], LIU Yanchao[1]
(1. State Key Laboratory of Media Integration and Communication, Communication University of China, Beijing 100024, China; 2. School of Computer and Artificial Intelligence, Beijing Wuzi University, Beijing 101149, China; 3. School of Mathematical Sciences, Peking University, Beijing 100871, China)

**Abstract**: With the rapid growth of social media content scale, traditional collaborative filtering recommender systems increasingly exhibit limitations in data sparsity and cold start problems. In recent years, the powerful data feature analysis and content generation capabilities of generative models have brought new development opportunities for recommender systems. This paper systematically reviews the technical frameworks and research progress in generative recommender systems, focusing on five key aspects: feature tokenization methods, core model architectural designs, mainstream evaluation protocols and typical application scenarios. Through comparative analysis and literature review, we demonstrate that generative recommender systems significantly outperform conventional approaches in recommendation accuracy, personality, and scenario adaptability. The study further identifies critical challenges including computational overhead, privacy risks, and standardization of evaluation metrics. Practical solutions and future research directions are proposed to address these challenges, breaking the cognitive bottleneck of generative recommender systems.

**Keywords**: recommender system; generative model; large language model; feature tokenization; representation learning; model architecture; collaborative information; evaluation method

随着互联网和人工智能技术的飞速发展, 推荐系统已成为连接用户与海量信息的核心工具, 在电商和社交媒体等领域发挥着关键作用。传统推荐系统通常采用协同过滤和多级过滤范式, 通过逐一对候选项目进行评分并排序生成推荐结果。然而, 这种方法面在对大规模项目库时的计算复杂度较高, 难以实时响应用户的需求。此

**收稿日期:** 2025−05−14.　**网络出版日期:** 2025−08−13.
**基金项目:** 北京物资学院系统科学研究院开放课题 (BWUISS35); 国家重点研发计划项目 (2022YFC3302103).
**通信作者:** 袁瑞萍. E-mail: yuanruiping@bwu.edu.cn.

外，传统推荐方法在用户兴趣精准捕捉、文本信息充分利用和多样化场景适应能力等方面的不足，进一步限制了其在复杂场景下的性能表现。

近年来，研究人员开始积极探索生成式模型与推荐系统的融合方式，形成了多种创新范式：将推荐任务重新定义为生成式问题，或将生成式模型作为特征编码器以学习用户和项目的语义表示。与此同时，大语言模型 (large language models, LLMs) 凭借其卓越的语言理解、生成、推理和知识迁移能力，正在重塑人工智能领域的技术格局，为推荐系统的革新提供了新的思路和可能。生成式推荐作为一种新兴范式，通过将项目编码为标记序列并采用生成式模型预测可能产生的交互项目，正吸引着学术界和工业界的广泛关注。

随着传统生成式模型和 LLM 技术的不断发展与完善，生成式推荐系统研究呈现出多元化和创新性的特点。现有部分文献[1-7] 总结了 LLM 与推荐系统结合的主要方式，但其只局限在讨论 LLM 中的推荐方法。目前有相关的研究工作[8-9] 总结了生成式推荐与传统推荐在范式上的区别，涵盖了 LLM 以外的生成式方法在推荐系统中的结合实现，但其只简要讨论了推荐系统中应用的生成模型。Huang 等[10] 总结了基于基础模型的推荐方法及其下游任务，吴国栋等[11] 总结了 LLM 在个性化推荐中的主要研究。然而，目前的研究对生成式推荐领域的整体流程总结和深入分析仍然不足。针对这一研究缺口，本文聚焦于生成式推荐系统，查阅了近 10 年来的发展现状，按照特征标记、模型训练、性能评估和工业应用 4 个阶段系统梳理关键技术路线，并深入探讨当前面临的主要挑战与未来发展方向。旨在为该领域研究者提供一个全面把握生成式推荐整体结构和前沿进展的理论框架，促进生成式推荐系统在学术研究和工业应用中的创新与突破。总的来说，本文梳理了生成式推荐系统的特征标记方法，从语义描述、向量量化、残差量化和 LLM 标记 4 个主要的研究方向进行概述。针对生成式推荐系统的主要模型架构，重点分类综述了 5 种主流的生成式模型在推荐系统中的应用。此外，本文介绍了生成式推荐系统的主要评估方式，包括评估基准和评估指标。最后，本文对生成式推荐系统的主要应用场景进行概述，并探讨了当前的挑战和未来的研究方向。

## 1　生成式推荐系统的特征标记方法

传统推荐的特征标记方法[12-13] 主要依赖唯一的数字 ID 来表示用户和项目，这种方法虽然能够有效捕捉特定项目之间的关系，但是缺少对项目本身语义信息进行捕捉学习，在冷启动和长尾场景特征方面存在明显局限。而语言模型能够理解复杂的语义关系，同时提供了持续进化的能力，这也为生成式推荐方法的发展提供了契机。此外，在现实场景的大规模推荐系统中，分配唯一的符号标记会产生海量的用户和项目 ID，在生成式推荐中会给大语言模型引入过多的词表信息。因此，在生成式推荐系统中，如何将推荐项目表示为适合生成式模型处理的标记序列是一个关键挑战。

特征标记是将推荐项目转化为模型可处理的离散或连续标识符的过程，是连接传统推荐系统与生成式模型的桥梁。标记方法的选取直接影响模型对项目语义的理解能力。本部分系统地介绍了生成式推荐中 4 种主要的特征标记方法：基于语义描述的标识符、基于向量量化的标识符、基于残差量化的标识符和基于 LLM 的标识符。

### 1.1　基于语义描述的标识符

基于语义描述的标识符通过显式提取项目的文本、属性等特征构建语义感知的离散符号作为项目的标识符，将包含语义描述的标识符直接作为输入提供给模型。这种方法充分利用了语言模型在自然语言处理方面的优势，能够捕捉项目的丰富语义信息。模型可以理解项目内容的细微差别，并基于语义相似性进行推荐。语义描述标识符的最大优势在于改善了冷启动问题，新项目只需提供文本描述即可快速更新至推荐系统。

由于仅基于内容的嵌入替换 ID 特征会因记忆能力降低而导致质量下降，Hua 等[14] 使用基于内容的语义索引从项目元数据来构建目标 ID，根据项目的类别组成一个书层次结构，每个非叶节点代表一个类别，每个叶节点代表一个项目，连接从根到叶的路径创建项目索引，同时包含了类别的粗细粒度。Hou 等[15] 提出了第一种上下文动作感知的标记化方法，根据项目的特征序列在单个集合和相邻集合中的共现频率进行动态感知，合并为新标记来构建词表作为语义标记。

为了将推荐中的项目特征空间与语言特征空间对齐，Lin 等[16] 提出一种多层面标识符，同时融合数字 ID、项目标题和项目属性进行项目索引，兼顾项目的区分性和语义表达，并设计了一种专门的数据结构确保仅生成有效的标识符，利用生成的标识符对语料库内项目进行排名。Qiu 等[17] 提出一种对比提示词学习框架 ControlRec，将文本语义信息与 ID 对齐融合，创建一个共享

的表示空间, 以降低 LLM 和项目表示之间的语义差距。Jin 等[18] 使用文本分词器对每个项目的文本描述进行分词化, 将生成的文本分词连接起来以形成输入作序列, 使模型自监督学习生成项目的语义 ID 表示。Zhai 等[19] 则把用户的交互行为当作模型的每一步行动, 将<项目, 行为>组成的 token 标记交错放置在每个交互中用于模型训练。

### 1.2　基于向量量化的标识符

基于向量量化 (vector quantization) 的标识符是一种融合了数字 ID 和语义信息的混合表示方法。该方法首先将项目映射到连续的向量空间, 然后通过向量量化将高维连续表示压缩为离散标记序列。在保持语义信息的同时适配 LLM 处理离散标记的特性, 相较于完整语义描述能提供更紧凑的表示形式, 大幅减少表示特征的维度。这种压缩表示不仅保留了项目间的相似关系, 还提高了模型效率。

由于部分推荐任务不能直接利用语言模型生成的连续型语义嵌入表示, 需通过向量量化技术将高维连续空间映射为离散符号序列。Hou 等[20] 在可迁移序列推荐任务中提出一种向量量化项目表示的新方法 VQ(vector-quantized)-Rec, 通过预训练语言模型 (pre-trained language models, PLMs) 将项目的描述性文本编码, 然后基于优化的点积量化构建文本编码和离散代码之间的映射, 使用这些映射索引查找代码嵌入表生成新项目的表示, 从而消除了项目和项目文本间的耦合绑定。

为了解决推荐系统中的用户和项目数量庞大带来的标记难题, Qu 等[21] 提出一种掩码向量量化分词器 MQ(masked vector)-Tokenizer, 将从协同过滤中学到的用户–项目掩码表示量化为离散标记, 使用少量的 LLM 词表即可表示大量的用户与项目, 从而增强了生成式推荐的分词泛化能力。

为了解决基于 ID 的上下文个性化推荐中项目类别信息可用性不一致的问题, Liu 等[22] 提出了一个可微分向量量化框架 CAGE, 通过级联向量量化构建具有不同粒度级别的类别树, 是首个在基于 ID 的推荐中引入向量量化学习分类知识的方法。

### 1.3　基于残差量化的标识符

基于残差量化 (residual quantization) 的标识符是一种采用多阶段分层压缩的混合表征方法。该方法通过迭代式残差逼近策略, 将高维向量空间分解为多个低维子空间: 首先对原始嵌入进行初始量化编码, 接着对量化残差进行递归分解和再量化, 最终通过级联码本生成层次化的离散复合标识符。这种分层量化机制在保证特征解耦性的同时, 实现了比单阶段量化更高的重构精度。相较于传统向量量化方法, 残差量化标识符在压缩效率与保持表征之间达到更优平衡, 为生成式推荐系统提供了高扩展性的特征标记方案。

当前的研究通常使用残差向量–变分自动编码器 (residual quantization variational auto-encoder, RQ-VAE) 实现特征的残差量化过程。Rajput 等[23] 提出的 TIGER(transformer index for generative recommenders) 在生成式推荐系统中首次引入 RQ-VAE 获取项目的语义标记, 通过分层残差编码结构, 将原始特征向量分解为多个可解释的语义子空间, 基于生成的语义序列在 Transformer 模型上训练并预测项目的语义 ID 表示。在每一级量化过程中, 模型通过变分推断学习具有最小重构误差的码本, 同时利用概率分布约束增强编码的鲁棒性。残差量化过程可形式化表述为

$$\begin{cases} C_l = \arg\min\limits_i \|\boldsymbol{r}_{l-1} - \boldsymbol{e}_i\|^2, \boldsymbol{e}_i \in C_l \\ \boldsymbol{r}_l = \boldsymbol{r}_{l-1} - \boldsymbol{e}_{C_l} \end{cases} \tag{1}$$

式中: $C_l$ 是第 $l$ 层码本分配的索引, $\boldsymbol{r}_{l-1}$ 是最后一层码本的残差语义, $\boldsymbol{e}_i$ 是码本中的嵌入向量。这种分层离散表示的上层标记捕获全局语义, 下层标记细化局部特征差异。在序列推荐任务中与传统模型相比展现了生成式推荐方法的潜力。

由于项目的语义标记对生成式推荐模型的性能至关重要, Zhu 等[24] 提出一种基于对比量化的语义标记化方法 CoST(contrastive quantization based semantic tokenization), 与 RQ-VAE 不同的是, CoST 在学习项目的语义信息的同时使用对比学习完善项目间的邻域关系, 以此对语义嵌入进行编码。保留了输入嵌入与其重建对应项之间的邻域信息, 同时增强了与其他项的差异性。

在生成式推荐场景中, RQ-VAE 虽然通过残差量化和重构损失函数实现了数据压缩, 但其对各项目的语义嵌入进行独立编码与离散化处理的机制, 可能导致生成的语义符号空间出现分布偏差, 容易引发高热度项目在隐空间中的符号表征过度重叠。这种码本坍缩现象使得密集区域的相似项目往往获得高度趋同的离散化表示, 从而削弱了模型对细粒度特征差异的辨别能力。近期一些研究发现了这一现象并提出了改进思路, Zheng 等[25] 提出了一种提出自我改进项目标记化方法 LC(language and collaborative)-Rec, 使用具有统一语义映射的 RQ-VAE 离散化生成多级表征, 为项目分配有意义且无冲突的 ID 映射。使用 LLaMA (large language model meta AI)-7B 微调后直接从整

个项目集合中生成项目进行推荐，而不依赖于候选项目。Wang 等[26] 提出了一种可学习的分词器 LETTER(learnable tokenizer for generative recommendation)，在利用 RQ-VAE 生成标识符的过程中集成层次化语义、协同信号和代码分配多样性正则化，并在重建语义向量的阶段增添了语义正则化。在 LC-Rec[25] 和 TIGER[23] 的基础上实现 LETTER 的实验结果证明了该方法在学习项目分词表示上的有效性。Lin 等[27] 提出的基于 LLM 的生成推荐系统 (SETRec) 引入一种新的标识符范式，整合协同过滤信息和项目本身的语义信息来获得多维项目信息，同时将每个项目表示为一组顺序无关的标记以缓解局部最优问题，提高生成式推荐中标识符 token 的生成效率。Liu 等[28] 改进了 RQ-VAE 中的分层量化码本生成过程，实现更加平衡的标识符分配。Singh 等[29] 提出对 ID 序列的子片段进行哈希处理，从冻结的内容嵌入中学习项目的紧凑离散表示，通过广泛的实验证明了量化生成的 ID 在保持记忆能力的同时可以提高新项目和长尾项目的泛化能力。

### 1.4　基于 LLM 的标记方法

随着 LLM 在语义理解与生成任务中的突破性进展，生成式推荐系统的项目标记方法正面临从“特征工程驱动”到“语义认知驱动”的范式迁移。LLM 凭借其强大的语义理解能力、生成式推理特性、开放域知识的拓扑关联能力，以及基于上下文推理的细粒度意图建模优势，为特征标记贡献新的力量。基于 LLM 的标记方法主要可以分为基于 LLM 的标记增强和基于 LLM 的标记生成两种方法。

#### 1.4.1　基于 LLM 的标记增强

在生成式推荐中，相关工作[30] 表明利用 LLM 的文本描述功能可以将项目和用户的异构特征转化为具有增强语义的辅助表征，从而生成包含隐式关联的语义描述标识符以提高推荐质量。因此，可以将推荐任务直接定义为自然语言序列理解和生成任务，Geng 等[31] 直接将用户和项目 ID 转换为纯文本表示的标识符作为 LLM 的输入，但是这种方式没有考虑两种信息表示的特征空间差距。为了提升在特定领域下的推荐系统特征表示能力，Li 等[32] 通过微调 LLM 将用户和项目的特殊标记集成到 LLM 的词表中以提高推荐的性能。部分文献[33-34] 则利用 LLM 内在的知识能力，根据交互信息直接生成检索模型需要的查询，作为辅助文本特征进行数据增强。

#### 1.4.2　基于 LLM 的标记生成

现有的推荐方法在处理未交互过的项目时会产生表征稀疏的问题，通过 PLM 提取通用的项目表示作为标记信息，能够充分利用 PLM 的内在知识完善项目特征。Tan 等[35] 提出的 IDGenRec 将 LLM 作为语义标识符生成器，根据项目的元信息生成对应的语义 ID，使 LLM 学习生成包含文本语义和潜在语义的高价值信息标识符。Li 等[36] 提出的 LLM4Rec 利用 GPT(generative pre-trained transformer)-2 从用户评论中生成个性化术语的隐式表示，并与普通的项目嵌入信息拼接作为特征标记。Xu 等[37] 提出的 ShopperBERT(bidirectional encoder representations from transformers) 根据用户的历史交互信息，利用语言模型直接生成用户的行为表征。Fan 等[38] 提出一种基于商品知识图谱的预训练推荐框架，通过 4 种预训练任务使 PLM 学习生成用户和项目的细粒度特征标记。

### 1.5　相关工作不足之处

4 种主要的项目标记方法汇总如表 1 所示。

**表 1　特征标记方法总结**

**Table 1　Summary of item tokenization methods**

| 标记类型 | 方法 | 数据集 | 表征类型 | 骨干网络 |
|---|---|---|---|---|
| 仅ID | BERT4Rec[12] | Amazon Beauty、Steam、MovieLens | ID | BERT |
| | SASRec[13] | Amazon、Steam、MovieLens | ID | 自注意力块 |
| 语义描述 | P5-CID[14] | Amazon Sports/Beauty、Yelp | 项目 | T5 |
| | ActionPiece[15] | Amazon Sports/Beauty/CDs | 项目 | Transformers |
| | TransRec[16] | Amazon Beauty/Toys、Yelp | 项目、交互 | BART、LLaMA |
| | ControlRec[17] | Amazon Sports/Beauty/Toys、Yelp | 项目、交互 | T5 |
| | LMIndexer[18] | Amazon Sports/Beauty/Toys | 交互 | Transformers |
| | HSTU[19] | MovieLens-1M/20M、Amazon Books | 项目、交互 | Transformers |
| 向量量化 | VQ-Rec[20] | Amazon、Online Retail | 项目 | Transformers、BERT |
| | MQTokenizer[21] | Amazon、LastFM、MovieLens | 用户、项目 | T5 |
| | CAGE[22] | Zhihu、Spotify、Goodreads、Amazon、MovieLens-1M/100K、MIND | 项目 | — |

续表 1

| 标记类型 | 方法 | 数据集 | 表征类型 | 骨干网络 |
|---|---|---|---|---|
| 残差量化 | TIGER[23] | Amazon Sports/Beauty/Toys | 用户、项目、交互 | T5-X |
| | CoST[24] | MIND、Office-S/L | 项目 | Sentence-T5 |
| | LC-Rec[25] | Amazon Instruments/Arts/Games | 项目、协同信号 | LLaMA-7B |
| | LETTER[26] | Amazon Instruments/Beauty、Yelp | 项目、协同信号 | LLaMA-7B |
| | SETRec[27] | Amazon Sports/Beauty/Toys、Steam | 项目、协同信号 | T5-S、Qwen2.5 |
| | MBGen[28] | Retail、IJCAI | 用户、项目、交互 | Switch Transformers |
| | SPM-SID[29] | YouTube | 项目 | Video-BERT |
| LLM增强 | P5[31] | Amazon Sports/Beauty/Toys、Yelp | 用户、项目、交互 | T0、GPT-2 |
| | EcomGPT[32] | 自定义 | 交互 | BLOOMZ |
| | GPT4Rec[33] | Amazon Beauty/Electronics | 项目 | GPT-2 |
| | MINT[34] | PointRec | 交互 | InstructGPT |
| LLM生成 | IDGenRec[35] | Amazon、Yelp | 用户、项目 | T5-S |
| | LLM4Rec[36] | TripAdvisor、Amazon、Yelp | 用户、项目、方面 | GPT-2 |
| | ShopperBERT[37] | 自定义 | 用户 | Sentence-BERT |
| | MPKG[38] | Xmarket | 项目 | GPT-GNN |

尽管上述项目标记方法在生成式推荐系统中的语义表示能力、空间压缩效率和层次特征表示方面展现出优势，但同时也存在一些局限性和不足之处。

基于语义描述的标识符主要应用于文本内容推荐场景，通过自然语言描述实现用户和项目的特征表示，然而其提取的语义标签难以捕捉用户行为的隐喻性关联。其次，静态语义无法适应动态演化的用户兴趣，会导致语义漂移问题。此外，来自不同领域的编码在统一的语义空间中存在领域对齐差距，限制其在跨域推荐中的应用。

基于向量量化的标识符方法通过离散编码实现特征压缩，其局限性在于量化过程的信息损失与空间划分的刚性约束。首先，在规模数量级较大的推荐系统中，降维生成的低维离散空间难以保留稀疏行为的细微差异。此外，预定义码本的不可变性导致模型无法自适应数据分布变化，对行为序列的时序动态性建模存在表征瓶颈。

基于残差量化的标识符方法通过多阶段残差学习提升层次化表征能力，适用于电商平台等高维稀疏数据场景，捕捉商品属性的细粒度特征差异，其局限性集中在层级误差累积与计算效率瓶颈。首先，多层级残差量化过程会使潜在空间的重建误差逐级放大，产生累积效应。此外，由于在训练过程中仅有少量码本向量被激活使用，可能导致潜在空间表征能力退化，从而产生码本坍缩问题。

基于 LLM 增强的标记方法利用 LLM 的上下文语义理解能力提取特征，适用于多模态或语义模糊项目的标记，虽在语义理解上取得突破，但在事实性和多特征集成上存在一定的局限性。首先，LLM 的概率生成特性存在事实性幻觉，使得生成的特征产生偏差信息，影响个性化推荐的性能，在领域微调中对词表添加的特殊标记也会影响模型的通用能力。此外，直接利用 LLM 提取得到的项目表征会损失 ID 特征中的唯一性和离散属性表示，在多特征集成建模的场景下能力受限。

上述标记方法提供了高效数据表示和强化语义理解的方法，以独特的方式增强了推荐系统的性能，为不同模型架构下的生成式推荐方法奠定了特征标记基础。

## 2　生成式推荐系统的主要模型架构

生成式推荐系统的核心目标是对用户行为、内容特征及上下文信息的生成式建模，实现动态、多样且可解释的推荐结果生成。相较于传统判别式推荐，生成式架构在以下 3 方面展现独特优势：其一，通过隐式建模捕捉用户与项目交互的潜在分布，有效缓解数据稀疏性问题；其二，利用序列生成等范式，实现端到端的推荐列表生成；其三，结合大语言模型的语义理解和生成能力，突破传统协同过滤的显式行为依赖局限。通过整合标记化提取的特征信息，根据任务需求采用不同的生成式模型进行训练，最终应用于下游场景，生成式推荐系统的主要流程如图 1 所示。本章将系统梳理生成式推荐系统的五大主要架构——自动编码器模型、自回归模型、生成对抗模型、扩散模型与大语言模型。

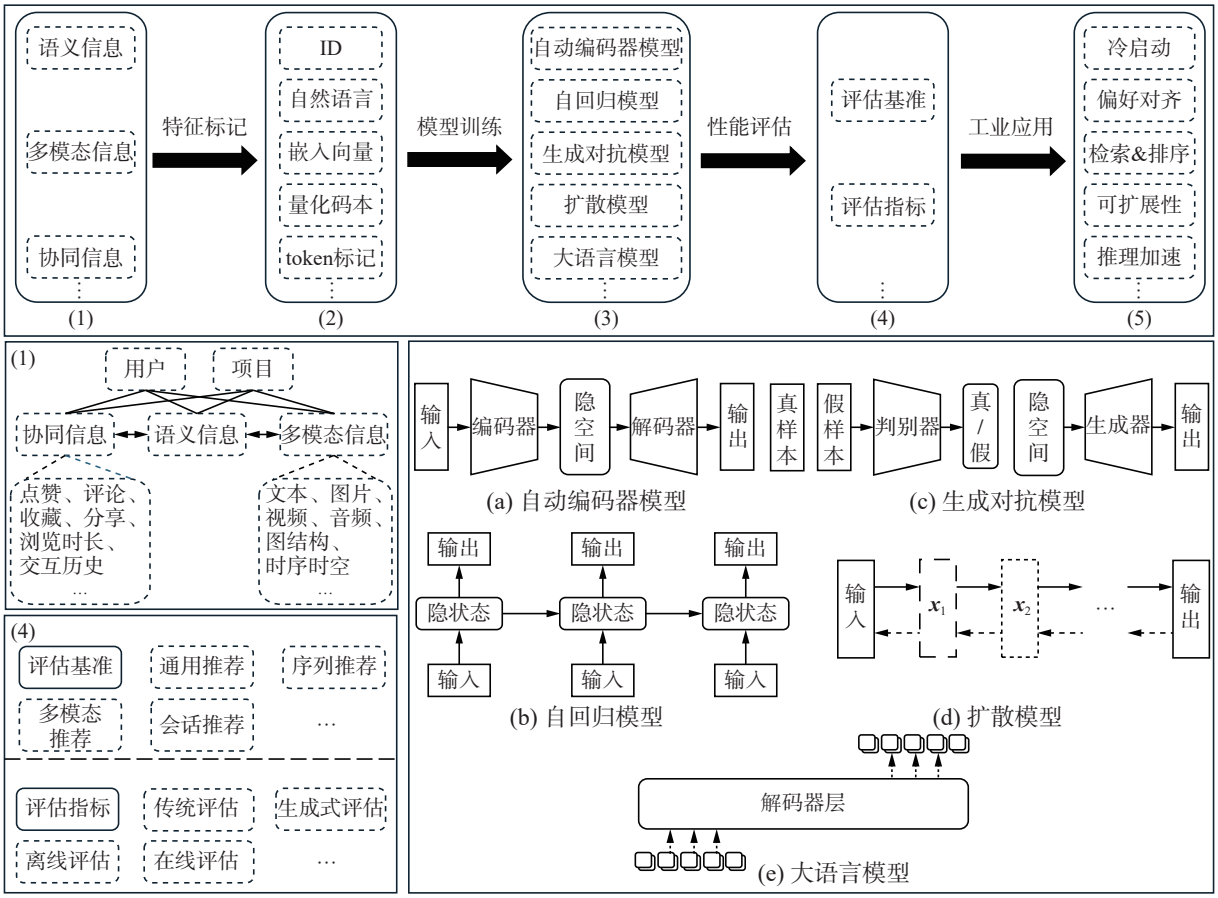


**图 1　生成式推荐系统的主要流程**

**Fig. 1　Main pipeline of generative recommender systems**

### 2.1　自动编码器模型

自动编码器 (auto-encoder models, AE) 是一种通过编码器−解码器架构实现数据重建的无监督神经网络，其核心思想是学习输入数据的低维非线性特征表示。令输入表示为 $\boldsymbol{x}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，编码器将 $x$ 编码为低维隐藏层表示 $\boldsymbol{h}=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n\}$，重建的输入表示为 $\boldsymbol{x}'=\{\boldsymbol{x}'_1,\boldsymbol{x}'_2,\cdots,\boldsymbol{x}'_n\}$。编码与解码过程可形式化表述为

$$\begin{cases}\boldsymbol{h}=f(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{b})\\ \boldsymbol{x}'=g(\boldsymbol{W}'\boldsymbol{h}+\boldsymbol{b}')\end{cases}\tag{2}$$

式中：$f$ 为编码器的激活函数，$g$ 为解码器的激活函数，$\boldsymbol{W}$、$\boldsymbol{b}$ 和 $\boldsymbol{W}'$、$\boldsymbol{b}'$ 分别为编码器和解码器中的权重矩阵与偏置向量。

Sedhain 等[39] 提出的 AutoRec 直接将用户和项目的评分向量作为输入数据，并在输出层获取重建后的评分。Yi 等[40] 则利用堆叠自动编码器提取输入的特征后重建输入信息，实现电影的自动化推荐。

#### 2.1.1　去噪自动编码器

去噪自动编码器 (denoising auto-encoder, DAE)[41] 是一组从输入的损坏版本中学习恢复原始输入的模型，通过向输入数据主动注入噪声，强制模型从部分损坏或扰动的观测中重建原始信号。Wu 等[42] 提出协作去噪自动编码器，将 DAE 引入协同过滤模型，在 Top-K 推荐任务上实现了更好的泛化性能。部分文献[43-44] 使用堆叠去噪自动编码器，将多个 DAE 堆叠在一起以获得多层级的特征表示，通过重建项目内容信息预测点击率。

#### 2.1.2　变分自动编码器

变分自动编码器 (variational auto-encoder, VAE)[45] 是一种可以学习数据分布的自动编码器，通过引入概率空间，从学习到的分布中采样生成新的数据样本。VAE 由编码器和解码器两部分组成，编码器将输入嵌入映射到低维潜在空间，解码器从潜在表示空间中采样并重建原始输入。与普通自动编码器简单重构输入的方式不同，VAE 通过变分推断在潜空间中采样邻近点的概率分布生成输出。

在生成式推荐系统中，VAE 能够基于自动编码器结构建模交互行为的隐式分布，解决数据稀疏性问题。部分文献[46-47] 在协同过滤 (collaborative filtering, CF) 方法基础上结合 VAE，通过建模隐式反馈中的交互关系，更精准地挖掘高维稀疏矩阵中的用户偏好。一些研究则直接利用 VAE

提取并生成推荐信息, Vo 等[48] 从商家的角度出发, 利用 VAE 学习现有项目特征分布并生成预期获得较高评价的项目内容。Xie 等[49] 针对不同交互数据的特点优化 VAE 的分布似然来生成更优的项目推荐。此外, 部分现有研究[50-51] 将 VAE 与基于图的知识表示学习结合, 对图中节点和关系的概率分布建模, 利用图结构约束优化潜在空间, 能够获取复杂拓扑关系的高阶语义。

### 2.2　自回归模型

自回归模型 (auto-regressive models, AR) 通过链式条件概率($p(x_t|x_{<t})$)分解生成推荐序列, 其递推式建模特性与用户行为序列的时序依赖性高度契合。根据上下文编码机制差异, 主流方法可分为递归神经网络和自注意力网络两类。

#### 2.2.1　递归神经网络

基于递归神经网络 (recurrent neural networks, RNNs) 的自回归模型通过隐藏层状态向量 $\boldsymbol{h}_t = f(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t)$迭代更新实现序列信息传递, 其马尔可夫链式特性与用户行为序列的变化适配。然而, 传统 RNN 因梯度消失问题难以建模长程依赖, 为此门控循环单元 (gated recurrent units, GRU) 和长短期记忆网络 (long short-term memory, LSTM) 两种架构被广泛采用。

GRU4Rec[52] 首次将 RNN 应用到基于会话的推荐任务中, 使用堆叠的 GRU 对项目之间的依赖关系进行建模, 完成项目到项目的推荐。Quadrana 等[53] 利用改进后的 GRU 实现序列推荐, 实现对用户偏好的时序动态跟踪。Yu 等[54] 利用 RNN 学习用户的动态兴趣表示和全局顺序特征, 证明 RNN 能够捕获用户在推荐系统中的交互顺序信息。为了捕获用户不同时间跨度下的兴趣, Wang 等[55] 将改进后的 LSTM 应用于下一个兴趣点 (point of interest, POI) 推荐任务, 针对时空上下文信息进行建模, 实现跨时间步的信息持久化存储, 提升用户兴趣周期性波动场景下的推荐性能。Zhu 等[56] 提出的 time-LSTM 利用时间门对用户的行为时间序列建模, 更好地捕捉了用户的短期和长期兴趣。

针对序列推荐中用户行为的长程依赖性建模难题, 部分研究基于选择性状态空间模型 (state space model, SSM) 提出结构化状态张量更新机制, 通过动态调节状态转移矩阵的时序关注权重, 有效捕捉跨会话的全局兴趣关联。Mamba4Rec[57] 和 RecMamba[58] 利用 SSM 平衡状态压缩与细粒度依赖建模能力, 为长行为序列推荐的高效处理提供了更具可扩展性的架构。

#### 2.2.2　自注意力网络

自注意力 (self-attention) 作为 Transformer[59] 架构的核心组件, 通过动态权重分配实现序列内任意位置的信息交互, 克服了传统 RNN 架构因串行计算特性难以捕捉长距离依赖的问题, 适用于用户行为序列中的非连续兴趣关联挖掘。其核心公式可表述为

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{3}$$

式中: $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别代表查询矩阵、键矩阵和值矩阵, 均由输入序列线性投影生成; $d_k$ 为缩放因子以避免梯度饱和。

SASRec(self-attentive sequential recommendation)[13] 首次将单向 Transformer 引入序列推荐任务, 通过堆叠多头自注意力层捕获用户长期和短期交互行为的多粒度特征来预测用户偏好。部分研究[60-61] 探索将时间跨度中的用户个性化引入基于 Transformer 的序列推荐, 部分文献[62-63] 利用自注意力对用户的多种行为特征和上下文共同建模, 进一步提高了推荐的准确性。

近年来, 以 Transformer 架构为基础的 LLM 发展迅速并大量应用于推荐系统, 以 LLM 为主的生成式推荐系统架构将在 2.5 节中单独详细介绍。

### 2.3　生成对抗模型

生成对抗网络 (generative adversarial networks, GANs)[64] 通过生成器与判别器的对抗博弈逼近真实数据分布, 实现生成与判别能力的协同优化。在推荐系统中, 生成器合成用户潜在交互行为, 判别器则基于用户真实反馈区分生成样本与观测数据。

作为生成对抗网络在推荐领域的里程碑式探索, Wang 等[65] 提出的 IRGAN(information retrieval generative adversarial network) 首次将对抗训练范式引入协同过滤的推荐任务, 通过生成器合成负样本与判别器的对抗训练, 提升长尾项目的召回率, 其目标函数可表述为

$$\begin{aligned}\min_{\theta}\max_{\phi} L(g_\theta, D_\phi) = E_{i\sim P_{x(i|u)}}\log f_\phi(i|u) + \\ E_{i\sim P_{\theta(\hat{i}|u)}}\log(1-f_\phi(\hat{i}|u))\end{aligned} \tag{4}$$

式中: $i$ 是用户 $u$ 交互的项目, $\hat{i}$是生成的项目。

此外, 一些研究探索将 GAN 用于不同领域的推荐任务, 包括协同过滤[66-67]、兴趣点推荐[68] 以及跨域推荐[69-70]。Yu 等[71] 还将 GAN 与 VAE 结合, 来解决潜在特征的有限表达问题, 从而增强对抗生成样本的鲁棒性, 减少推荐中的偏差信息。

### 2.4　扩散模型

扩散模型 (diffusion models, DM)[72] 通过前向

扩散$q(x_t|x_{t-1})$与逆向生成$P_\theta(x_{t-1}|x_t)$逐步修正生成结果，相较于 GAN 和 VAE 更擅长交互特征的重建。其加噪与去噪过程的核心公式可分别表述为

$$\begin{cases} q(x_t|x_{t-1}) = N(x_t;\sqrt{1-\beta_t}x_{t-1},\beta_t I) \\ P_\theta(x_{t-1}|x_t) = N(x_{t-1};\mu_\theta(x_t,t),\sigma_\theta(x_t,t)I) \end{cases} \tag{5}$$

式中：$t$ 为前向扩散的时间步；$x_{t-1}$为前一步的数据；$x_t$为当前步的数据，其分布是一个高斯分布。而反向生成过程由可学习的参数 $\theta$ 控制，在时间步 $t$ 时给定当前的噪声数据 $x_t$、模型需要预测前一步的数据分布 $x_{t-1}$。推荐系统是根据用户交互的正负样本预测推荐物品的交互概率，这与扩散模型的重建目标基本一致。基于此，推荐领域逐步探索扩散模型的适配方式。

Wang 等[73] 提出一种扩散推荐模型 DiffRec，以去噪的方式学习用户交互的生成过程。具体来说，首先在前向过程中注入高斯噪声逐渐破坏用户的交互历史，然后通过参数化的神经网络迭代地从被破坏的交互中恢复原始状态。此外，部分文献[74-75] 将扩散模型引入协同过滤，将交互历史视为潜在扩散变量，通过反向去噪恢复预测用户真实的偏好分布。部分相关研究[76-77] 利用扩散模型实现序列推荐任务，将用户交互序列的表示作为反向去噪过程的指导信息，进行基于条件扩散的生成式推荐。部分文献[78-79] 利用扩散模型的生成特性，根据文本提示生成对应图像作为推荐数据的补充，来优化点击率等推荐效果。

### 2.5　大语言模型

大语言模型凭借其海量预训练知识库、强大的语义理解与生成能力，为生成式推荐提供了新的范式。LLM 能够将用户和项目信息以及交互信号统一编码为自然语言序列，并通过自回归生成等机制输出推荐列表或交互式响应。生成式推荐系统中，基于 LLM 的架构可系统划分为非参数调优与参数调优两大范式，其中包含的具体结构如图 2 所示。

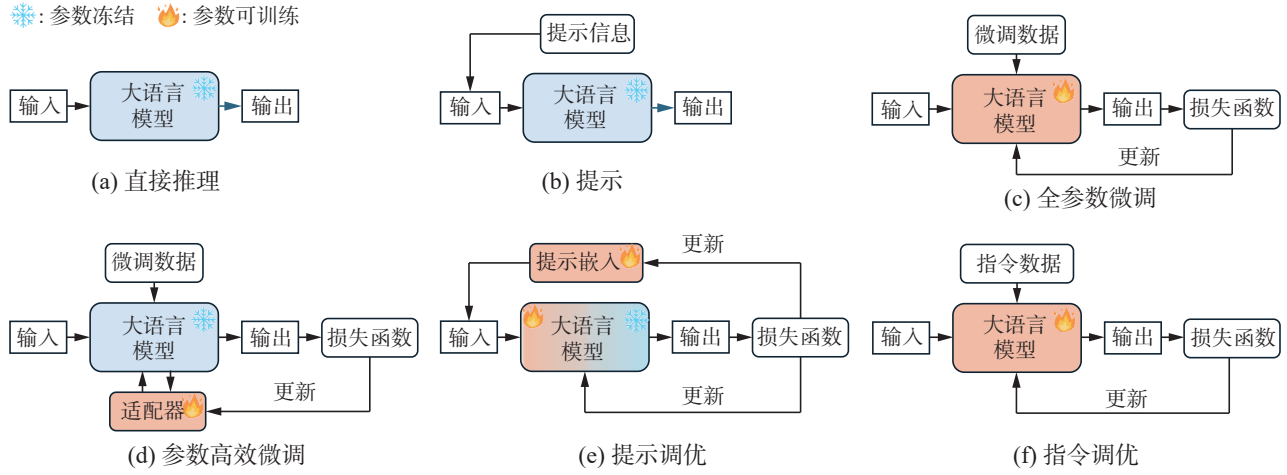


**图 2　生成式推荐系统中的大语言模型范式**

**Fig. 2　LLM's paradigms for generative recommender systems**

#### 2.5.1　非参数调优

非参数调优直接利用 LLM 的预训练能力，以输入设计为核心驱动推荐生成，无需更新模型参数。此类方法凭借低计算成本，在少样本、零样本推荐及资源敏感场景中展现出显著优势。其核心逻辑在于通过输入激发 LLM 内嵌的语义推理与知识泛化能力，但其性能受限于预训练语料的领域覆盖度与输入设计质量。根据输入引导策略的差异，非参数调优可进一步细化为直接推理和提示。

直接推理利用 LLM 的预训练参数生成推荐结果，无需任何参数调整或指令引导，基于 LLM 架构的生成式推荐模型通常采用 LLM 本身的预训练方式。Wu 等[80] 提出的 PTUM(pre-training user model) 引入行为掩码预测和接下来 $k$ 个行为预测这两种预训练任务，利用大规模未标记的用户行为数据进行用户模型预训练。Geng 等[31] 提出的 P5 使用多任务掩码语言建模，将多种推荐任务数据集混合预训练并应用至下游任务场景。Ngo 等[81] 提出的 RecGPT 是首个基于文本推荐进行领域适应的 LLM，在收集的混合推荐领域数据集上进行了全面预训练。Cui 等[82] 提出的 M6-Rec 引入文本填充和自回归生成两种预训练任务，遵循 LLM 预训练中的预测下一个 token 范式，自回归解码目标项目序列。

提示 (Prompting) 是指冻结 LLM 的预训练参数，通过人工设计或规则生成的提示词模板显式引导模型在下游任务中的生成能力。根据提示内

容的不同, 可以分为普通提示词、上下文学习 (in-context learning, ICL) 和思维链 (chain-of-thought, CoT)3 种提示方式。

普通提示词通过任务描述直接触发模型的生成能力, Wang 等[8] 利用 LLM 根据现有数据和设计好的提示信息生成新的数据, 用于领域推荐任务模型的训练。Sanner 等[83] 利用包括项目信息和用户偏好在内的 3 种组合提示方式, 验证了不同提示词对 LLM 推荐能力的增强效果。然而, 普通提示词对复杂用户意图的表征建模能力有限。

上下文学习通过向输入中注入少量任务相关的示例, 使模型从示例的输入–输出映射中隐式学习推荐逻辑, 从而提升模型在下游任务下的情境感知能力。利用少样本 (few-shot) 或零样本 (zero-shot) 的 LLM 可用于多种领域的推荐任务, 例如会话推荐[84]、序列推荐[85] 和个性化推荐[86], 但其性能会受到样本质量与分布一致性的影响。

思维链提示要求模型显式拆解推荐决策的中间推理步骤, 以分阶段生成的方式约束生成方向。根据不同的推荐任务场景和目标, 需要设计对应的思维链推理步骤。部分研究将 CoT 提示应用于序列推荐[87] 和个性化推荐[88]。然而, 思维链对提示逻辑的严谨性要求更高, 可能因推理链复杂度增加导致生成效率降低。

#### 2.5.2　参数调优

参数调优通过调整 LLM 的权重参数实现推荐任务的深度适配, 其核心目标是在计算效率与模型性能间寻求最优平衡, 同时增强对复杂推荐逻辑的建模能力。根据参数调整范围与优化策略的差异, 参数调优可分为以下 4 类方法。

全参数微调 (full fine-tuning) 对 LLM 所有权重参数进行端到端更新, 深度捕捉领域数据的分布特征。Shen 等[89] 利用专用的推荐领域数据集对预训练 LLM 进行全参数微调, 实现对下游推荐任务的适配。然而, 直接进行全参数微调会导致高昂的训练成本、灾难性遗忘和偏见问题。

参数高效微调 (parameter-efficient fine-tuning, PEFT) 通过局部参数更新或引入额外的可学习参数, 实现低成本达到与全参数微调相近的效果。近期部分研究利用 PEFT 降低推荐任务领域 LLM 微调成本, 主要实现方式包括低秩适应调优 (low-rank adaptation tuning)[16,90]、适配器调优 (adapter tuning)[91] 和前缀调优 (prefix tuning)[92], 这使得 LLM 与推荐任务高效对齐。

提示调优 (prompt tuning) 通过将可学习的提示向量与输入序列拼接训练, 隐式调整模型生成能力。与直接提示静态语义引导不同的是, 提示调优向 LLM 中添加新的提示标记实现动态表示优化, 自动学习提示语义。部分相关研究[93-96] 利用硬提示 (hard prompt) 或软提示 (soft prompt) 调优指导 LLM 在下游推荐任务中生成预期的输出内容。

指令调优 (instruction tuning) 通过显式任务指令数据微调模型, 强化其对复杂推荐逻辑的遵循能力。与专注于提升单一任务性能的提示调优方法不同的是, 指令调优通过指令生成和模型调优两个阶段, 增强了推荐模型对指令遵循和多样化任务的零样本泛化能力。部分文献[14] 通过指令调优提升 LLM 在特定推荐需求下的生成能力。

### 2.6　其他模型

除了上述几种主流模型架构外, 一些工作还常利用其他的生成模型实现对传统推荐模型的改进。InDGRM(interpretable deep generative recommendation model)[97] 是一种可解释的深度生成推荐模型, 用于建模用户间偏好相似性和用户内偏好多样性, 同时将学习到的表征从观察层和潜在层中分离。DGR[98] 通过深度生成排名模型同时考虑逐点匹配和成对排名, 针对稀疏反馈数据进行个性化推荐。GFN4Rec(generative flow network)[99] 使用流式生成网络确保列表生成概率与其奖励之间的一致性, 为用户生成足够多样化的商品列表, 同时保持较高的推荐质量。未来研究可探索多种生成范式协同架构, 针对不同任务进行联合优化建模, 实现更好的推荐性能。

### 2.7　相关工作不足之处

尽管多种生成式模型架构的改进提升了推荐系统的性能, 但生成式推荐系统的模型架构层面仍面临多重技术挑战。

自动编码器模型的优势在于高效的特征压缩能力和隐式反馈建模, 但存在梯度消失导致推荐多样性不足的问题。去噪自动编码器的性能易受人工噪声策略的影响, 不当的噪声设计可能导致模型过拟合或欠拟合。现有基于 VAE 的推荐模型通常是在全局假设下通过结合简单先验来正则化潜在变量, 当用户的偏好高度多样化时 VAE 的策略会不适用。由于近似后验分布的下界与真实数据分布之间存在差距, VAE 通常无法捕捉多模态分布, 且生成结果偏向平均化, 削弱长尾覆盖效果。

自回归模型中, 基于 RNN 的架构受限于序列化计算效率与长程依赖建模能力, 难以适配超长行为序列的实时推理需求。此外, RNN 方案没有考虑用户邻居动作之间的时间间隔, 而这些时间间隔对于捕捉用户动作的关系很重要。自注意力

网络虽突破序列长度限制，但其二次计算复杂度与位置编码偏差显著增加内存开销与动态场景建模难度。此外，用户行为的严格时序性要求精准的位置编码且兴趣强度随时间非线性变化，而自注意力中存在位置感知偏差，导致推荐泛化性能存在瓶颈。

生成对抗网络因生成器与判别器的博弈均衡难以维持，容易引发模式坍缩问题，表现为生成样本多样性衰减与热门商品重复推荐；同时，判别器与生成器的交替优化机制易导致训练过程不稳定，增加模型收敛的不确定性。

扩散模型虽通过渐进式去噪实现高保真度的推荐生成，但其马尔可夫链式迭代机制引发计算效率瓶颈，多步采样过程导致的线性时间开销与实时推荐的低延迟需求形成根本性冲突；此外，模型对噪声调度策略的敏感使得细微的参数偏差可能引发推荐质量的非线性衰减。

大语言模型的非参数微调具有提示敏感性、领域泛化不足等局限性，而全参数微调的训练成本高昂，模型可能丢失预训练阶段学习的通用语义能力，导致跨域推荐泛化性下降。提示调优性能受初始提示分布影响显著。指令调优依赖高质量指令数据集，且对模糊指令的容忍度较低，容易导致生成偏差。

## 3　生成式推荐系统的评估

生成式推荐系统的评估需构建与其生成特性深度适配的验证体系，涵盖数据表征指标设计的多维性以及场景落地的可迁移性。当前研究主要依赖经典的推荐数据集，结合传统推荐指标与生成式任务特有的评估维度，形成全面的评估范式，但其在生成式性能验证和多任务场景等层面仍面临挑战。本节系统梳理主流数据集与评价指标的分类、特性及应用场景，为生成式推荐模型的性能验证与对比分析提供方法论支撑。

### 3.1　评估基准

当前主流的推荐系统评估数据集主要围绕推荐场景的差异化需求展开，主要可分为通用推荐、序列推荐、多模态推荐与会话推荐等，这些数据集也可用于生成式推荐系统的性能评估。表 2 给出了当前生成式推荐任务的主流评估数据集及其信息。

**表 2　评估基准总结**

**Table 2　Summary of evaluate benchmark**

| 任务场景 | 数据集名称 | 数据集规模 | 任务类型 |
|---|---|---|---|
| 通用推荐 | Yelp Challenge[100] | 47万条评论、15.6万个商业实体、20万张图片 | 商户 |
| | MIND[101] | 100万名用户和16万篇新闻 | 新闻 |
| | Tenrec[102] | 超过500万名用户和1.4亿次交互行为 | 多场景 |
| 序列推荐 | Movielens-1M[103] | 6 040名用户、3 900部电影、1 000 209条评分数 | 电影 |
| | Amazon Review[104] | 2.331亿条评论、18种商品类别 | 商品 |
| | Steam[105] | 超过85 000款游戏详细信息及评价 | 游戏 |
| | Douban[106] | 14万部电影、63万名用户、416万条评分和442万条影评 | 电影 |
| 多模态推荐 | PixelRec[107] | 3 000万用户、2亿次交互行为和40万高质量视频封面 | 封面图像 |
| | MicroLens[108] | 3 400万名用户、10亿次交互行为和100万条短视频 | 短视频 |
| | KuaiSAR[109] | 25 877名用户及交互行为 | 短视频 |
| | KuaiRec[110] | 1 411名用户对3 327个短视频的交互行为 | 短视频 |
| 会话推荐 | AmazonM2[111] | 6个地区1 410 675件商品的用户会话和交互 | 商品 |
| | ReDial[112] | 超过1万条用户对话 | 电影 |
| | U-need[113] | 7 698条对话，333 879条交互行为和332 148个商品 | 商品 |
| | INSPIRED[114] | 1 001条用户对话 | 电影 |
| 可迁移推荐 | NiceRec[115] | 200万名用户、3 000万次交互行为和40万个项目 | 短视频 |

通用推荐数据集主要聚焦用户−物品基础交互的静态建模；序列推荐数据集主要记录用户和项目各自的信息以及交互行为的时序标记，为序列生成提供验证和基准；多模态数据集突破了传统协同信号的局限，融合了图文描述、音视频等内容；会话推荐数据集通过会话下的短期交互，支持生成式模型对兴趣捕捉与会话预测能力的评估。最近，Zhang 等[115] 提出一种用于评估可迁移模型的基准 NiceRec，构建了一个大规模、高质量的迁移学习推荐数据集和基准测试集，为评估生成式推荐系统做出了贡献。

然而，生成式推荐系统评估与传统数据集的设计范式间存在一定的局限性。首先，工业级推荐场景通常需处理较大数量级的信息，而部分数

据集的交互密度远低于真实场景, 导致基于 LLM 的推荐模型的长上下文理解能力难以充分验证。其次, 现有数据集多集中于电商、社交媒体等场景, 缺少医疗、教育等垂直领域的数据, 而以 LLM 为主的推荐模型在预训练阶段对专业知识的学习不足可能导致生成结果的事实性错误。此外, 现有部分 LLM 的预训练语料中可能已经包含推荐领域数据集内容, 容易干扰推荐评估的可信度。最后, 主流数据集的用户行为隐含地域、性别等群体偏好, 模型的生成过程中可能放大此类偏差。综上所述, 未来需要构建更加完善的可信评估基准, 以满足生成式推荐系统的发展需求。

### 3.2　评估指标

生成式推荐系统的评估需突破传统指标的单一维度局限, 建立覆盖推荐效果、生成质量、计算效率等的复合评估体系。

生成式模型应用于传统的推荐任务时, 可以复用传统推荐任务中的一些评估指标。推荐效果评估以准确性为核心, 准确率 (precision)、召回率 (recall)、归一化折扣累积增益 (normalized discounted cumulative gain, NDCG)、平均倒数秩 (mean reciprocal rank, MRR)、F1 分数 (F1-score)、ROC(receiver operating characteristic) 曲线、ROC 曲线下面积 (area under the ROC curve, AUC) 等指标用于衡量推荐类别匹配准确度。平均绝对误差 (mean absolute error, MAE)、平均平方误差 (mean square error, MSE)、均方根误差 (root mean square error, RMSE) 等指标用于衡量推荐预测分数准确度。点击通过率 (click through rate, CTR)、转化率 (conversion rate, CVR)、投资回报率 (return on investment, ROI) 等指标衡量商业推荐流量转化能力。除了离线评估外, 还可以采用在线 A/B 测试, 衡量用户与推荐商品的实际交互效果。

传统指标侧重交互行为的匹配精度, 难以量化生成式任务中文本逻辑性。为了更好地评估以 LLM 为核心的生成式推荐方法, 需要引入 LLM 中的常用评估方法。生成质量层面, 常采用自然语言处理中的评估指标 BLEU(bilingual evaluation understudy) 评估文本生成结果与参考文本之间的 n-gram 匹配精度; ROUGE(recall-oriented understudy for gisting evaluation) 通过计算召回率衡量生成内容对参考文本关键信息的覆盖程度; 困惑度 (perplexity) 用于量化语言模型预测文本的不确定性。表 3 给出了当前生成式推荐任务常用的评估指标。

**表 3　评估指标总结**

**Table 3　Summary of metrics**

| 任务场景 | 评估指标名称 | 说明 |
|---|---|---|
| 分类准确度 | 准确率 | 衡量推荐结果中用户真正感兴趣的项目所占的比例 |
| | 召回率 | 衡量所有用户真正感兴趣的项目中被成功推荐出来的比例 |
| | NDCG | 衡量推荐列表整体质量, 考虑推荐顺序和相关性程度 |
| | MRR | 衡量推荐列表中首个正确推荐项位置的倒数平均值 |
| | F1-score | 精确率与召回率的调和平均数, 反映模型的整体性能 |
| | ROC | 不同阈值下正类识别能力的曲线图 |
| | AUC | 衡量ROC曲线下的面积, 代表模型整体的排序能力 |
| 预测准确度 | MAE | 衡量预测值与真实值之间的平均绝对误差, 反映模型预测的准确性 |
| | MSE | 衡量预测值与真实值之间的平均平方误差, 对较大误差更加敏感 |
| | RMSE | MSE的平方根, 衡量预测值与真实值之间的均方根误差 |
| 流量转化 | CTR | 用户点击推荐内容的比例, 衡量推荐内容的吸引力和用户兴趣匹配程度 |
| | CVR | 用户完成购买等目标行为的比例, 衡量推荐系统的转化效果 |
| | ROI | 衡量推荐系统带来的收益与投入成本的比值 |
| 生成质量 | BLEU | 衡量生成文本与参考文本的相似度 |
| | ROUGE | 衡量生成文本与参考文本的重叠度 |
| 不确定性 | 困惑度 | 衡量语言模型生成文本的不确定性, 数值越低表示模型对数据的拟合越好 |

计算效率评估主要关注 LLM 在训练和推理过程中的时间和资源开销, 其核心目标在于平衡模型性能与工业落地的可行性。训练阶段需量化 GPU(graphic processing unit) 资源使用量以及模型参数规模; 推理阶段则需关注 token 解码生成速率、检索效率与实时性能等。

然而, 当前大多数的生成式推荐方法在评估阶段只采用传统推荐指标, 仅关注推荐结果与用户历史行为的匹配度, 但生成式模型可能因幻觉问题产生不存在或错误的描述。尽管现有方法可以利用 LLM 中的幻觉评价指标, 但其与推荐评估仍然是独立的, 尚无标准化指标能够联合量化生

成式推荐的推荐幻觉问题。此外，现有评估依赖自然语言处理任务中的文本相似度指标，而 LLM 在思考推理过程中生成的解释信息难以通过传统指标评估。最后，生成式推荐常需同时完成推荐、解释生成与总结等多任务，但现有评估体系仍以单一任务指标为主，在跨场景与多任务迁移中缺乏统一的评估标准。

## 4　生成式推荐系统的应用场景

生成式推荐系统作为一种新型推荐范式，已在多种场景中展现出其独特优势。本章探讨生成式推荐的 3 个主要应用场景：冷启动问题、用户偏好对齐及检索与排序优化。在冷启动场景中，生成式推荐通过其强大的特征理解和生成能力，有效缓解了新增用户和项目特征稀疏的问题。在偏好对齐场景中，通过将用户的显式和隐式偏好与模型对齐，实现更加个性化的推荐效果。在检索与排序场景中，通过重新定义传统推荐流程，提供了一种统一的端到端解决方案。这 3 个场景相互关联，共同构成了生成式推荐系统的核心应用领域。

### 4.1　冷启动

冷启动问题一直是推荐系统面临的主要挑战之一，包括用户冷启动和项目冷启动两个方面。生成式推荐系统凭借其强大的特征理解和内容生成能力，为冷启动问题提供了新的解决思路。例如，针对电商平台中新注册的用户，通过生成式模型分析用户注册信息、兴趣范围和浏览行为，快速构建初始兴趣画像并生成个性化推荐列表；在社交平台的新内容发布场景下，利用生成式模型提取视频或图文的多模态特征，结合相似内容的交互分布生成兴趣标签并匹配相似的用户浏览群体和推荐范围。

在项目冷启动方面，生成式模型主要将训练阶段学习到的语义特征保存在模型参数中，因此在推理阶段难以直接和未交互过的新项目匹配。在动态偏好感知场景中，采用周期性全参数重训练机制来维持知识库时效性并不现实。Zhao 等[1]使用启发式策略将新项目混合到推荐列表中，从而增强模型对未知项目的拟合能力。Ding 等[116]提出的 SpecGR(generative recommendation) 使生成式推荐模型能够在归纳偏置中推荐新物品，根据候选项目可能成为输入序列目标的概率来判断接受或拒绝，提升了生成式模型对新样本的归纳学习推荐能力。Xu 等[117]提出了一种生成式自约束框架 $GS^2$-RS(recommender system)，通过生成用户细粒度兴趣与满意度偏好，构造虚拟可信邻居偏好，并引入低兴趣高满意度的“意外项”逆向注入评分矩阵，同时缓解冷启动和过滤气泡问题。

在用户冷启动方面，新用户或低活跃用户由于缺少有效的消费行为数据，使得模型在开启推荐策略时预测准确率降低。为了更好的改善用户冷启动问题，Bai 等[118]提出一种提出了一个基于多媒体项目的生成式冷启动推荐框架 Gorec，利用条件变分自动编码器构建预热物品潜在分布，通过条件概率建模从学习到的潜在分布中生成新物品的预热表征。Huang 等[119]提出一种用于在线推荐的耦合漏斗状 LLM 框架，使用经过训练的耦合过滤器有效地将候选用户数量从数十亿减少到数百，从而允许 LLM 在过滤集上高效运行。

基于大规模语料预训练的生成模型具有强大的知识迁移能力，能够从稀疏的用户数据中挖掘潜在语义信息，生成用户兴趣的初始表示，但是直接利用 LLM 完成训练需要大量的资源成本。Kusano 等[120]设计提示词将 LLM 作为数据增强手段，在降低推理成本的同时改进训练数据不足的冷启动问题。Wu 等[121]提出的 PromptRec 将推荐任务转换为包含用户和项目的自然语言的情感分析任务，通过构建用于模型预训练的精细语料库和提示词模板，使得增强后的小语言模型实现与大型模型相当的冷启动推荐性能。Jiang 等[122]认为将正反馈信号作为提示信息能够更好地降低语义差距，并提出利用逐项个性化提示生成器对正反馈进行编码，以减轻正反馈优势问题对模型偏差的影响。

### 4.2　偏好对齐

偏好对齐是指推荐系统对用户兴趣偏好的精准捕捉和匹配过程，直接影响推荐的个性化程度和用户满意度。在生成式推荐系统中引入偏好对齐，能够处理更复杂、更细粒度的用户偏好表达。例如，在电商推荐中，通过捕捉用户对商品属性的不同粒度偏好，实现多维度的精准匹配；在内容推荐平台上，利用偏好对齐机制理解用户对视频主题、时长和评论观点等复合偏好，生成个性化内容序列。

在显式偏好对齐方面，生成式推荐系统能够理解和处理用户以自然语言表达的需求和偏好。与传统推荐系统仅依赖评分或点击等简单反馈不同，生成式推荐可以处理用户的详细描述、多维度偏好表达以及偏好的条件约束。Fang 等[123]提出一种推理驱动框架 Reason4Rec，设置 3 个具有推理能力的专家学习实现偏好蒸馏、偏好匹配和

反馈预测, 在 LLM 做出预测之前进行推理, 将推理过程与用户的真实偏好保持一致。由于现有方法大多数获取的是共有性偏好, 缺少对每个用户的独立偏好判断, Liao 等[124] 提出的个性化偏好对齐框架 PosePO(recommendation with smoothing personalized preference optimization) 从模型误判中采样构建负样本, 选择高相似性负例进行对抗语义相似性误导, 并根据项目的受欢迎程度选择被拒绝的样本以修正流行度偏差, 从而实现用户深层偏好与项目特征的对齐优化。Shao 等[125] 提出一种个性化用户索引机制 ULM-Rec(user-centric large language model for sequential recommendation), 在生成包含个性化信息的用户索引后通过指令微调设计偏好对齐任务, 将用户级别的个性化信息注入 LLM。

在隐式偏好挖掘方面, 生成式推荐系统可以通过分析用户的历史行为序列, 推断用户未明确表达但可能存在的潜在偏好。基于 LLM 理解能力的方法能够捕获用户行为背后的潜在意图和长期兴趣, 有效挖掘用户的隐式偏好, 提高推荐的多样性和新颖性。一些研究探索利用直接偏好优化 (direct preference optimization, DPO) 使生成式推荐模型和人类偏好保持一致。Deng 等[126] 提出一种迭代偏好对齐策略 IPA(iterative preference alignment), 通过预训练奖励模型提供的分数对样本响应进行排名模拟用户生成, 使用有限数量的 DPO 样本即可调整用户的兴趣偏好。Gao 等[127] 则认为 DPO 会使模型偏向于少数得分较高的项目, 因此提出一种自我迭代的学习方法 SP(self-play)Rec, 对 DPO 损失函数进行重新加权, 自适应地抑制存在偏差的项目, 减少过度推荐并提高公平性。

### 4.3　检索与排序

在推荐系统中, 检索方法作为召回−粗排−精排级联架构的首层核心组件, 面临着海量候选集与极短响应时间的双重挑战, 传统基于协同过滤或向量内积的检索范式难以平衡覆盖度与计算效率。Bin 等[128] 基于变分自动编码器提出一种流式向量量化检索器, 实时为项目分配可以更新和分布更平衡的索引, 通过动态调整检索策略, 根据用户当前状态和系统反馈自适应地生成候选物品, 有效提升大规模工业推荐系统的检索效率。

当前的研究主要依赖项目的 ID 特征进行建模, 而放弃了传统嵌入中的细粒度信息。Zhai 等[19] 根据用户的交互行为来统一生成式推荐系统中的检索和排序任务。Zhou 等[129] 讨论了结合生成检索和密集检索的必要性, 认为仅使用生成式检索会导致标识符生成存在偏差。为了优化这个问题, Yang 等[130] 提出的 COBRA(cascaded organized bi-represented generative retrieval) 集成了稀疏 ID 和密集嵌入分别作为粗粒度和细粒度语义, 将两种检索方式的分数加权融合生成预测结果的排序。Penha 等[131] 则验证了将检索与推荐集成在生成式联合训练任务中能规范项目的流行度估计和潜在表示, 相较于单任务训练能够提升效果。

在排序阶段, 生成式推荐系统可以将多特征的排序问题转化为生成问题, 通过综合考虑用户特征、物品特征、上下文信息等, 直接生成排序得分或排序列表, 能够自然融合多种异构特征, 无需手动设计特征交互方式; 另一方面, 还能够处理多目标排序问题, 平衡相关性、多样性等多种优化目标。Chen 等[132] 提出的 Rankformer 利用 Transformer 结构根据所有用户和项目的全局信息生成更丰富的特征, 并将计算降低为线性复杂度, 以提高排序性能。为减少 LLM 在通用任务和下游任务之间的能力差异, Luo 等[96] 提出的 Recranker 采用自适应用户采样构建指令微调数据集, 并在提示词中提出位置转移策略来减轻 LLM 中的位置偏差, 最后将各种排名任务集成后调整 LLM 的指令以提升多任务的排序性能。Wang 等[133] 提出一种基于 LLM 的知识约束生成式重排序框架 KC(knowledge-constrained)-GenRe, 通过知识引导的交互式训练和知识增强的约束推理方法, 指导 LLM 的生成候选项目的有效排名。

### 4.4　其他应用场景

除了前文讨论的应用场景外, 生成式推荐系统在其他多个工业领域也展现出了广阔的应用前景。这些应用场景进一步展现了生成式推荐的潜力, 为解决传统推荐系统面临的挑战提供了新思路。

在工业级推荐系统的实际部署中, 资源分配与模型效能的最优平衡是核心挑战之一。为挖掘 LLM 在推荐任务中的扩展潜力, Zhai 等[19] 提出一种层次化序列直推式单元编码器架构 HSTU (hierarchical sequential transduction units), 在工业界首次实现万亿级别参数的生成式推荐系统, 并展示了 LLM 中的缩放定律 (Scaling Law) 同样适用于大规模推荐系统。Yan 等[134] 也探索了基于 LLM 的工业推荐系统中的缩放定律, 验证了随着训练数据规模与 LLM 参数量的同步增长, 模型对用户兴趣的捕捉能力会持续提升。通过动态扩展数据与模型的协同规模, 可在保证推理效率的前提下实现推荐效果的持续优化。

尽管生成式推荐系统规模化扩展可显著提升性能，工业场景中较高数量级的模型参数使得直接训练 LLM 存在局限性。为此，Jia 等[135] 使用预训练 LLM 作为项目编码器并冻结 LLM 参数以优化运算效率，避免模型灾难性遗忘并保留开放世界知识。Zhao 等[136] 提出的 LLM-KERec(knowledge enhanced recommendation) 将传统推荐模块与基于 LLM 的互补知识增强模块结合，在保证计算成本的同时充分利用 LLM 强大的推理能力。Hu 等[137] 提出的 SAID(semantic alignment for item descriptions) 框架使 LLM 显示学习项目 ID 嵌入中保留的文本语义信息，从而减少特征提取中的冗余 token 序列，降低了工业场景所需的资源消耗。

### 4.5　相关工作不足之处

尽管生成式推荐系统为几种重要的工业应用场景提供了新的架构范式与解决思路，但从理论优势向工程实效的转化过程仍然存在一些局限之处。

冷启动场景下，生成式推荐面临数据稀疏性与生成可信度双重挑战。新用户或新项目的交互数据匮乏导致模型难以捕捉潜在特征分布，基于生成增强数据的推荐可能偏离真实用户意图。此外，跨域知识迁移受限于语义鸿沟，难以实现精准领域适配。

偏好对齐旨在弥合用户显式反馈与隐式行为间的语义鸿沟，但生成式模型在此任务中面临多目标优化冲突与动态演化失配。点赞、评论等显式反馈与浏览时长等隐式行为的内在逻辑差异导致联合表征学习时梯度信号相互干扰，模型可能偏向高曝光物品的点击偏差而忽略长尾偏好。同时，用户兴趣的动态漂移特性要求生成式模型具备在线自适应能力，但生成过程难以实时捕捉偏好演化轨迹，导致推荐结果滞后于真实需求。

检索阶段中，生成式模型需从亿级候选池中筛选潜在正样本，但其全局生成机制导致计算复杂度与候选集规模呈线性增长，难以满足工业级实时响应需求。排序阶段中，生成式模型倾向于为头部商品分配过高置信度，而长尾物品因生成概率分布衰减被忽略，加剧推荐同质化。

## 5　挑战与展望

### 5.1　特征标记优化

生成式推荐需要捕捉用户偏好和项目属性的复杂关系，现有方法大多局限于文本的特征标记，而多模态特征增加了标记的复杂程度，仅使用特殊位置标记区分不同模态可能会丢失多模态中的粗细粒度特征。此外，现有工作大多将特征标记视为生成式推荐训练的预处理步骤，导致模型优化过程中特征标记和自回归生成完全解耦[138]。未来应进一步优化生成式推荐的特征标记策略，以提升特征表示的语义完整性和鲁棒性。具体而言，应探索解决标记过程中的语义碰撞问题，通过动态语义码本更新机制，基于用户反馈和行为数据持续优化标记映射关系。此外，应细化多模态特征的标记方式，根据用户的行为序列和上下文模态信息动态调整特征权重分配，充分整合不同粒度的特征信息，在多媒体场景下提升对用户偏好的理解能力。还应探索端到端联合优化特征标记与生成式模型训练过程，同时平衡表征建模能力和计算效率，以适应大规模推荐场景。

### 5.2　轻量化生成式架构设计

当前的生成式模型尤其是基于 LLM 的推荐系统普遍面临算力成本高的问题，限制了其在资源受限的工业场景下部署和应用[139]。传统的大规模预训练模型在推荐任务中需要的高昂计算开销，显著增加了在线系统的响应延迟。未来的研究可以聚焦于设计更加轻量的生成式推荐架构。具体而言，通过知识蒸馏技术将教师模型的海量知识参数迁移至轻量化的学生模型，同时采用量化技术降低模型权重的存储与计算开销。针对特定领域的推荐任务，可以探索专用的参数剪枝策略，保留关键的特征表达能力，同时最大限度地降低模型复杂度。此外，可以利用混合专家的动态路由机制，实现更加精细和高效的模型计算。

### 5.3　新型评估体系构建

生成式推荐系统的评估需要超越传统指标，特别是生成的推荐内容的质量、多样性和个性化等方面。传统指标往往无法评估生成内容的用户满意度，也难以检测生成内容中的潜在偏见或虚假信息。在新闻和电商等领域，评估生成内容的真实性和公平性尤为重要[9]。因此，如何开发能够综合考虑传统推荐指标和生成式推荐能力的评估框架，是当前的主要挑战。未来的研究需构建更全面的评估框架，以适应生成式推荐系统的发展需求。通过引入多样性生成式评价方式来补充传统指标，同时应评估生成内容中的不实信息，以确保推荐系统的可信度。此外，还应评估推荐系统对用户认知偏差和信息茧房的影响。通过建立跨领域、跨场景的新型综合评估体系，为生成式推荐系统提供可靠的基准。

### 5.4　隐私保护与推荐安全

当前的生成式推荐系统在隐私保护与推荐安

全领域面临双重挑战: 一方面, 模型训练依赖海量用户数据如个人资料与交互历史的深度挖掘, 但现有隐私保护技术在平衡数据效用与匿名化强度时仍存在不足[140]。另一方面, 生成式模型可能存在提示词注入攻击、训练数据污染以及生成歧视与偏见内容的风险[141]。未来的研究可以考虑在隐私保护层面借助联邦学习与隐私保护算法, 实现数据可用不可见。在推荐安全层面可以融合因果去偏技术, 设计实时内容审核引擎, 从特征表示、生成过程和输出结果三阶段实施动态监测。此外, 应建立伦理审核与对齐机制, 通过人类价值观对齐和不良推荐结果的筛选过滤机制, 确保推荐系统在隐私安全和内容安全方面发挥更大的潜力。

### 5.5　推荐与强化学习结合

当前的生成式推荐系统在实时交互和长期用户兴趣演变建模方面存在局限, 这源于传统生成式模型对静态历史数据的过度依赖以及短期奖励驱动的优化机制。部分文献[126]尝试引入基于反馈信息的强化学习进行偏好调整, 但大多局限于近端策略优化 (proximal policy optimization, PPO) 或 DPO 的离线调整, 使得面对海量的动态新增数据时建模能力有限。近期深度求索 (DeepSeek) 团队提出的 DeepSeek-R1[142] 首次验证无需依赖预置的指令微调数据, 仅通过纯强化学习范式即可驱动大语言模型自主学习长思维链推理和自我反思能力。因此, 结合强化学习的探索−利用机制将成为生成式推荐系统能力提升的重要发展方向。未来研究可以设计构建能够自适应学习用户长期偏好演化的奖励机制, 并基于反馈循环持续优化推理路径。通过强化学习和思维链推理技术, 实现推荐系统对复杂用户行为模式的精准捕捉和预测。

## 6　结束语

本文对生成式推荐系统的主要发展和最新研究现状进行了综述, 包括生成式推荐系统的特征标记方法、主要模型架构、评估方法以及主要应用场景, 论证了生成式推荐相比传统的基于协同过滤的推荐方法在个性化、准确性和泛化性等方面的优势。

然而, 当前的生成式推荐从理论到实践的转化仍存在关键挑战: 首先, 多模态特征提取不足和训练优化过程割裂现象限制了表征建模的能力上限; 其次, 生成式模型对计算资源的需求与推理延迟问题制约了工业推荐场景实时部署; 此外, 生成式推荐性能的准确评估、用户的隐私安全保护和强化思考能力提升也是需要解决的问题。

针对生成式推荐方法, 仍存在许多值得进一步探索的方向: 1) 优化生成式推荐的特征标记过程; 2) 探索设计轻量级生成式推荐模型架构; 3) 完善构建全面的生成式推荐评估体系; 4) 增强用户隐私保护和推荐安全性能; 5) 拓展生成式推荐系统与强化学习结合的思考推理能力。

生成式推荐系统作为一种全新的推荐范式, 随着生成式人工智能和大语言模型的快速发展, 将持续推荐并产生更多创新性的解决方案, 致力于为用户提供更加高效、安全和个性化的推荐服务。

## 参考文献:


[1] ZHAO Zihuai, FAN Wenqi, LI Jiatong, et al. Recommender systems in the era of large language models (LLMs)[J]. IEEE transactions on knowledge & data engineering, 2024(1): 1−20.

[2] LI Lei, ZHANG Yongfeng, LIU Dugang, et al. Large language models for generative recommendation: A survey and visionary discussions[EB/OL]. (2024−05−23)[2025−04−20]. http://arxiv.org/abs/2309.01157.

[3] LIN Jianghao, DAI Xinyi, XI Yunjia, et al. How can recommender systems benefit from large language models: a survey[J]. ACM transactions on information systems, 2025, 43(2): 1−47.

[4] VATS A, JAIN V, RAJA R, et al. Exploring the impact of large language models on recommender systems: An extensive review[EB/OL]. (2024−05−19)[2025−04−20]. http://arxiv.org/abs/2402.18590.

[5] WU Likang, ZHENG Zhi, QIU Zhaopeng, et al. A survey on large language models for recommendation[J]. World wide web, 2024, 27(5): 60.

[6] WANG Qi, LI Jindong, WANG Shiqi, et al. Towards next-generation llm-based recommender systems: A survey and beyond[EB/OL]. (2024−10−10)[2025−04−20]. http://arxiv.org/abs/2410.19744.

[7] 卡祖铭, 赵鹏, 张波, 等. 面向大语言模型的推荐系统综述[J]. 计算机科学, 2024, 51(S2): 11−21.
KA Zuming, ZHAO Peng, ZHANG Bo, et al. Survey of recommender systems for large language models[J]. Computer science, 2024, 51(S2): 11−21.

[8] WANG Wenjie, LIN Xinyu, FENG Fuli, et al. Generative recommendation: Towards next-generation recommender paradigm[EB/OL]. (2024−02−25)[2025−04−20]. http://arxiv.org/abs/2304.03516.

[9] DELDJOO Y, HE Zhankui, MCAULEY J, et al. A review of modern recommender systems using generative

models (gen-recsys)[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Barcelona: ACM, 2024: 6448−6458.

[10] HUANG Chengkai, YU Tong, XIE Kaige, et al. Foundation models for recommender systems: A survey and new perspectives[EB/OL]. (2024−02−17)[2025−04−20]. http://arxiv.org/abs/2402.11143.

[11] 吴国栋, 秦辉, 胡全兴, 等. 大语言模型及其个性化推荐研究[J]. 智能系统学报, 2024, 19(6): 1351−1365.
WU Guodong, QIN Hui, HU Quanxing, et al. Research on large language models and personalized recommendation[J]. CAAI transactions on intelligent systems, 2024, 19(6): 1351−1365.

[12] SUN Fei, LIU Jun, WU Jian, et al. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing: ACM, 2019: 1441−1450.

[13] KANG W C, MCAULEY J. Self-attentive sequential recommendation [C]//2018 IEEE International Conference on Data Mining. Singapore: IEEE, 2018: 197−206.

[14] HUA Wenyue, XU Shuyuan, GE Yingqiang, et al. How to index item ids for recommendation foundation models[C]//Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region. Beijing: ACM, 2023: 195−204.

[15] HOU Yupeng, NI Jianmo, HE Zhankui, et al. ActionPiece: Contextually tokenizing action sequences for generative recommendation[EB/OL]. (2025−02−19)[2025−04−20]. http://arxiv.org/abs/2502.13581.

[16] LIN Xinyu, WANG Wenjie, LI Yongqi, et al. Bridging items and language: A transition paradigm for large language model-based recommendation[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Barcelona: ACM, 2024: 1816−1826.

[17] QIU Junyan, WANG Haitao, HONG Zhaolin, et al. ControlRec: Bridging the semantic gap between language model and personalized recommendation[EB/OL]. (2023−11−28)[2025−04−20]. http://arxiv.org/abs/2311.16441.

[18] JIN Bowen, ZENG Hansi, WANG Guoyin, et al. Language models as semantic indexers[C]//Proceedings of the 41st International Conference on Machine Learning. Vienna: ACM, 2024: 22244−22259.

[19] ZHAI Jiaqi, LIAO L, LIU Xing, et al. Actions speak louder than words: trillion-parameter sequential transducers forgenerative recommendations[C]//Proceedings of the 41st International Conference on Machine Learning. Vienna: ACM, 2024: 58484−58509.

[20] HOU Yupeng, HE Zhankui, MCAULEY J, et al. Learning vector-quantized item representation for transferable sequential recommenders[C]//Proceedings of the ACM Web Conference 2023. Austin: ACM, 2023: 1162−1171.

[21] QU Haohao, FAN Wenqi, ZHAO Zihuai, et al. Tokenrec: learning to tokenize id for llm-based generative recommendation[EB/OL]. (2024−08−18)[2025−04−20]. http://arxiv.org/abs/2406.10450.

[22] LIU Qijiong, XIAO Jiaren, FAN Llu, et al. Learning category trees for ID-vased recommendation: Exploring the power of differentiable vector quantization[C]//Proceedings of the ACM Web Conference 2024. Singapore: ACM, 2024: 3521−3532.

[23] RAJPUT S, MEHTA N, SINGH A, et al. Recommender systems with generative retrieval[J]. Advances in neural information processing systems, 2023, 36: 10299−10315.

[24] ZHU Jieming, JIN Mengqun, LIU Qijiong, et al. CoST: contrastive quantization based semantic tokenization for generative recommendation[C]//Proceedings of the 18th ACM Conference on Recommender Systems. Bari: ACM, 2024: 969−974.

[25] ZHENG Bowen, HOU Yupeng, LU Hongyu, et al. Adapting large language models by integrating collaborative semantics for recommendation[C]//2024 IEEE 40th International Conference on Data Engineering. Utrecht: IEEE, 2024: 1435−1448.

[26] WANG Wenjie, BAO Honghui, LIN Xinyu, et al. Learnable item tokenization for generative recommendation[C]//Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. Boise: ACM, 2024: 2400−2409.

[27] LIN Xinyu, SHI Haihan, WANG Wenjie, et al. Order-agnostic identifier for large language model-based generative recommendation[C]//Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. Padua: ACM, 2025: 1923−1933.

[28] LIU Zihan, HOU Yupeng, MCAULEY J. Multi-behavior generative recommendation[C]//Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. Boise: ACM, 2024: 1575−1585.

[29] SINGH A, VU T, MEHTA N, et al. Better generalization with semantic ids: A case study in ranking for recommendations[C]//Proceedings of the 18th ACM Conference on Recommender Systems. Bari: ACM, 2024:

1039−1044.

[30] XI Yunjia, LIU Weiwen, LIN Jianghao, et al. Towards open-world recommendation with knowledge augmentation from large language models[C]//Proceedings of the 18th ACM Conference on Recommender Systems. Bari: ACM, 2024: 12−22.

[31] GENG Shijie, LIU Shuchang, FU Zuohui, et al. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5)[C]//Proceedings of the 16th ACM Conference on Recommender Systems. Seattle: ACM, 2022: 299−315.

[32] LI Yangning, MA Shirong, WANG Xiaobin, et al. EcomGPT: instruction-tuning large language models with chain-of-task tasks for E-commerce[C]//Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. Vancouver: ACM, 2024: 18582−18590.

[33] LI Jinming, ZHANG Wentao, WANG Tian, et al. GPT4Rec: A generative framework for personalized recommendation and user interests interpretation[EB/OL]. (2023−04−19)[2025−04−20]. http://arxiv.org/abs/2304.03879.

[34] MYSORE S, MCCALLUM A, ZAMANI H. Large language model augmented narrative driven recommendations[C]//Proceedings of the 17th ACM Conference on Recommender Systems. Singapore: ACM, 2023: 777−783.

[35] TAN Juntao, XU Shuyuan, HUA Wenyue, et al. Idgenrec: Llm-recsys alignment with textual id learning[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. Washington DC: ACM, 2024: 355−364.

[36] LI Pan, WANG Yuyan, CHI E H, et al. Prompt tuning large language models on personalized aspect extraction for recommendations[EB/OL]. (2023−06−02)[2025−04−20]. http://arxiv.org/abs/2306.01475.

[37] XU Da, RUAN Chuanwei, KORPEOGLU E, et al. Product knowledge graph embedding for e-commerce[C]//Proceedings of the 13th International Conference on Web Search and Data Mining. Houston: ACM, 2020: 672−680.

[38] FAN Ziwei, LIU Zhiwei, HEINECKE S, et al. Zero-shot item-based recommendation via multi-task product knowledge graph pre-training[C]//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. Birmingham: ACM, 2023: 483−493.

[39] SEDHAIN S, MENON A K, SANNER S, et al. Autorec: Autoencoders meet collaborative filtering[C]//Proceedings of the 24th International Conference on World Wide Web. Florence: ACM, 2015: 111−112.

[40] YI Baolin, SHEN Xiaoxuan, ZHANG Zhaoli, et al. Expanded autoencoder recommendation framework and its application in movie recommendation[C]//2016 10th International Conference on Software, Knowledge, Information Management & Applications. Chengdu: IEEE, 2016: 298−303.

[41] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[C]//Proceedings of the 25th International Conference on Machine learning. Helsinki: ACM, 2008: 1096−1103.

[42] WU Yao, DUBOIS C, ZHENG A X, et al. Collaborative denoising auto-encoders for top-n recommender systems[C]//Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. San Francisco: ACM, 2016: 153−162.

[43] WANG Hao, SHI Xingjian, YEUNG D Y. Relational stacked denoising autoencoder for tag recommendation[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin: ACM, 2015: 3052−3058.

[44] WANG Hao, WANG Naiyan, YEUNG D Y. Collaborative deep learning for recommender systems[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney: ACM, 2015: 1235−1244.

[45] KINGMA D P, WELLING M. Auto-encoding variational bayes[EB/OL]. (2013−12−20)[2025−04−20]. http://arxiv.org/abs/1312.6114.

[46] LIANG Dawen, KRISHNAN R G, HOFFMAN M D, et al. Variational autoencoders for collaborative filtering[C]//Proceedings of the 2018 World Wide Web Conference. Lyon: ACM, 2018: 689−698.

[47] WU Ga, BOUADJENEK M R, SANNER S. One-class collaborative filtering with the queryable variational autoencoder[C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Paris: ACM, 2019: 921−924.

[48] VO T V, SOH H. Generation meets recommendation: proposing novel items for groups of users[C]//Proceedings of the 12th ACM Conference on Recommender Systems. Vancouver: ACM, 2018: 145−153.

[49] XIE Zhe, LIU Chengxuan, ZHANG Yichi, et al. Adversarial and contrastive variational autoencoder for sequential recommendation[C]//Proceedings of the Web Conference 2021. Ljubljana: ACM, 2021: 449−459.

[50] YI Jing, CHEN Zhenzhong. Multi-modal variational graph auto-encoder for recommendation systems[J]. IEEE transactions on multimedia, 2021, 24: 1067−1079.

[51] ZHANG Yi, ZHANG Yiwen, YAN D, et al. Revisiting graph-based recommender systems from the perspective of variational auto-encoder[J]. ACM transactions on information systems., 2023, 41(3): 1−28.

[52] HIDASI B, KARATZOGLOU A, BALTRUNAS L, et al. Session-based recommendations with recurrent neural networks[C]. Proceedings of the 11th ACM Conference on Recommender Systems. Como: ACM, 2017: 130–137.

[53] QUADRANA M, KARATZOGLOU A, HIDASI B, et al. Personalizing session-based recommendations with hierarchical recurrent neural networks[C]//Proceedings of the 11th ACM Conference on Recommender Systems. Como: ACM, 2017: 130−137.

[54] YU Feng, LIU Qiang, WU Shu, et al. A dynamic recurrent model for next basket recommendation[C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. Pisa: ACM, 2016: 729−732.

[55] WANG Chen, YUAN Mengting, YANG Yang, et al. Revisiting long-and short-term preference learning for next POI recommendation with hierarchical LSTM[J]. IEEE transactions on mobile computing, 2024.

[56] ZHU Yu, LI Hao, LIAO Yikang, et al. What to do next: modeling user behaviors by time-LSTM[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne: ACM, 2017: 3602−3608.

[57] LIU Chengkai, LIN Jianghao, WANG Jianling, et al. Mamba4rec: Towards efficient sequential recommendation with selective state space models[EB/OL]. (2024−06−29)[2025−04−20]. http://arxiv.org/abs/2403.03900.

[58] YANG Jiyuan, LI Yuanzi, ZHAO Jingyu, et al. Uncovering selective state space model's capabilities in lifelong sequential recommendation[EB/OL]. (2024−05−25)[2025−04−20]. http://arxiv.org/abs/2403.16371.

[59] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017−06−12)[2025−04−20]. http://arxiv.org/abs/1706.03762.

[60] WU Liwei, LI Shuqing, HSIEH C J, et al. SSE-PT: Sequential recommendation via personalized transformer[C]//Proceedings of the 14th ACM Conference on Recommender Systems. Virtual Event: ACM, 2020: 328−337.

[61] De SOUZA P M G, RABHI S, LEE J M, et al. Transformers4rec: Bridging the gap between nlp and sequential/session-based recommendation[C]//Proceedings of the 15th ACM Conference on Recommender Systems. Amsterdam: ACM, 2021: 143−153.

[62] YUAN Enming, GUO Wei, HE Zhicheng, et al. Multi-behavior sequential transformer recommender[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. Madrid: ACM, 2022: 1642−1652.

[63] HUANG Xiaowen, QIAN Shengsheng, FANG Quan, et al. Csan: Contextual self-attention network for user sequential recommendation[C]//Proceedings of the 26th ACM International Conference on Multimedia. Seoul: ACM, 2018: 447−455.

[64] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[65] WANG Jun, YU Lantao, ZHANG Weinian, et al. Irgan: a minimax game for unifying generative and discriminative information retrieval models[C]//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. Shinjuku: ACM, 2017: 515−524.

[66] CHAE D K, KANG J S, KIM S W, et al. CFGAN: a generic collaborative filtering framework based on generative adversarial networks[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. Torino: ACM, 2018: 137−146.

[67] WU Qiong, LIU Yong, MIAO Chunyan, et al. PD-GAN: adversarial learning for personalized diversity-promoting recommendation[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao: ACM, 2019: 3870−3876.

[68] ZHOU Fan, YIN Ruiyang, ZHANG Kunpeng, et al. Adversarial point-of-interest recommendation[C]//The World Wide Web Conference. San Francisco: ACM, 2019: 3462−34618.

[69] FAN Wenqi, DERR T, MA Yao, et al. Deep adversarial social recommendation[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao: ACM, 2019: 1351−1357.

[70] PERERA D, ZIMMERMANN R. CnGAN: Generative Adversarial Networks for Cross-network user preference generation for non-overlapped users[C]//The World Wide Web Conference. San Francisco: ACM, 2019: 3144−3150.

[71] YU Xianwen, ZHANG Xiaoning, CAO Yang, et al. VAEGAN: a collaborative filtering framework based on adversarial variational autoencoders[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao: ACM, 2019: 4206−4212.

[72] SOHL-DICKSTEIN J, WEISS E A, MAHESWARANATHAN N, et al. Deep unsupervised learning us-

ing nonequilibrium thermodynamics[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille: ACM, 2015: 2256−2265.

[73] WANG Wenjie, XU Yiyan, FENG Fuli, et al. Diffusion recommender model[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. Taipei: ACM, 2023: 832−841.

[74] HOU Yu, PARK J D, SHIN W Y. Collaborative filtering based on diffusion models: Unveiling the potential of high-order connectivity[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. Washington DC: ACM, 2024: 1360−1369.

[75] ZHU Yunqin, WANG Chao, ZHANG Qi, et al. Graph signal diffusion model for collaborative filtering[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. Washington DC: ACM, 2024: 1380−1390.

[76] YANG Zhengyi, WU Jiancan, WANG Zhicai, et al. Generate what you prefer: reshaping sequential recommendation via guided diffusion[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans: ACM, 2023: 24247−24261.

[77] LI Zihao, SUN Aixin, LI Chenliang. Diffurec: A diffusion model for sequential recommendation[J]. ACM transactions on information systems, 2023, 42(3): 1−28.

[78] YANG Hao, YUAN Jianxin, YANG Shuai, et al. A new creative generation pipeline for click-through rate with stable diffusion model[C]//Companion Proceedings of the ACM Web Conference 2024. Singapore: ACM, 2024: 180−189.

[79] CZAPP Á T, JANI M, DOMIÁN B, et al. Dynamic product image generation and recommendation at scale for personalized ecommerce[C]//Proceedings of the 18th ACM Conference on Recommender Systems. Bari: ACM, 2024: 768−770.

[80] WU Chuhan, WU Fangzhao, QI Tao, et al. PTUM: Pre-training user model from unlabeled user behaviors via self-supervision[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Online: ACL, 2020: 1939−1944.

[81] NGO H, NGUYEN D Q. RecGPT: Generative pre-training for text-based recommendation[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok: ACL, 2024: 302−313.

[82] CUI Zeyu, MA Jianxin, ZHOU Chang, et al. M6-rec: Generative pretrained language models are open-ended recommender systems[EB/OL]. (2022−05−19)[2025−04−20]. http://arxiv.org/abs/2205.08084.

[83] SANNER S, BALOG K, RADLINSKi F, et al. Large language models are competitive near cold-start recommenders for language-and item-based preferences[C]//Proceedings of the 17th ACM Conference on Recommender Systems. Singapore: ACM, 2023: 890−896.

[84] WANG Xiaolei, TANG Xinyu, ZHAO Xin, et al. Rethinking the evaluation for conversational recommendation in the era of large language models[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: ACL 2023: 10052−10065.

[85] WANG Lei, LIM E P. The whole is better than the sum: Using aggregated demonstrations in in-context learning for sequential recommendation[C]//Findings of the Association for Computational Linguistics: NAACL 2024. Mexico City: ACL, 2024: 876−895.

[86] LIU Dairui, YANG Boming, DU Honghui, et al. RecPrompt: a self-tuning prompting framework for news recommendation using large language models[C]//Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. Boise: ACM, 2024: 3902−3906.

[87] WANG Yuling, TIAN Changxin, HU Binbin, et al. Can small language models be good reasoners for sequential recommendation?[C]//Proceedings of the ACM Web Conference 2024. Singapore: ACM, 2024: 3876−3887.

[88] TSAI A, KRAFT A, JIN Long, et al. Leveraging LLM reasoning enhances personalized recommender systems[C]//Findings of the Association for Computational Linguistics ACL 2024. Bangkok: ACL, 2024: 13176−13188.

[89] SHEN Tianshu, LI Jiaru, BOUADJENEK M R, et al. Towards understanding and mitigating unintended biases in language model-driven conversational recommendation[J]. Information processing and management, 2023, 60(1): 103139.

[90] ZHANG Yang, FENG Fuli, ZHANG Jizhi, et al. CoLLM: Integrating collaborative embeddings into large language models for recommendation[J]. IEEE transactions on knowledge and data engineering, 2025(1): 1−12.

[91] GENG Shijie, TAN Juntao, LIU Shuchang, et al. VIP5: Towards multimodal foundation models for recommendation[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: ACL, 2023: 9606−9620.

[92] JIANG Junzhe, QU Shang, CHENG Mingyue, et al. Reformulating sequential recommendation: learning dynamic user interest with content-enriched language mod-

eling[C]//International Conference on Database Systems for Advanced Applications. Gifu: ACM, 2024: 353−362.

[93] ZHANG Zizhuo, WANG Bang. Prompt learning for news recommendation[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. Taipei: ACM, 2023: 227−237.

[94] WANG Xiaolei, ZHOU Kun, WEN Jirong, et al. Towards unified conversational recommender systems via knowledge-enhanced prompt learning[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Washington DC: ACM, 2022: 1929−1937.

[95] WEI Wei, TANG Jiabin, XIA Lianghao, et al. Prompt-mm: Multi-modal knowledge distillation for recommendation with prompt-tuning[C]//Proceedings of the ACM Web Conference 2024. Singapore: ACM, 2024: 3217−3228.

[96] LUO Sichun, HE Bowei, ZHAO Haohan, et al. Recranker: Instruction tuning large language model as ranker for top-k recommendation[EB/OL]. (2023−12−26)[2025−12−23]. https://arxiv.org/abs/2312.16018.

[97] LIU Huafeng, JING Liping, WEN Jingxuan, et al. Interpretable deep generative recommendation models[J]. Journal of machine learning research, 2021, 22(202): 1−54.

[98] LIU Huafeng, WEN Jingxuan, JING Liping, et al. Deep generative ranking for personalized recommendation[C]//Proceedings of the 13th ACM Conference on Recommender Systems. Copenhagen: ACM, 2019: 34−42.

[99] LIU Shuchang, CAI Qingpeng, HE Zhankui, et al. Generative flow network for listwise recommendation[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Long Beach: ACM, 2023: 1524−1534.

[100] YELP I.Yelp Dataset[EB/OL]. (2014)[2025−12−23]. https://www.yelp.com/dataset.

[101] WU Fangzhao, QIAO Ying, CHEN J H, et al. Mind: a large-scale dataset for news recommendation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: ACL, 2020: 3597−3606.

[102] YUAN Guanghu, YUAN Fajie, LI Yudong, et al. Tenrec: a large-scale multipurpose benchmark dataset for recommender systems[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: ACM, 2022: 11480−11493.

[103] HARPER F M, KONSTAN J A. The movielens datasets: History and context[J]. ACM transactions on interactive intelligent systems, 2015, 5(4): 1−19.

[104] NI Jianmo, LI Jiacheng, MCAULEY J. Justifying recommendations using distantly-labeled reviews and fine-grained aspects[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: ACL, 2019: 188−197.

[105] WAN Mengting, MCAULEY J. Item recommendation on monotonic behavior chains[C]//Proceedings of the 12th ACM Conference on Recommender Systems. Vancouver: ACM, 2018: 86−94.

[106] WU Yu, WU Wei, XING Chen, et al. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based Chatbots[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver: ACL, 2017: 496−505.

[107] CHENG Yu, PAN Yunzhu, ZHANG Jiaqi, et al. An image dataset for benchmarking recommender systems with raw pixels[C]//Proceedings of the 2024 SIAM International Conference on Data Mining (SDM). Texas: SIAM, 2024: 418−426.

[108] NI Yongxin, CHENG Yu, Liu Xiangyan, et al. A content-driven micro-video recommendation dataset at scale [EB/OL]. (2023−09−27)[2025−04−20]. http://arxiv.org/abs/2309.15379.

[109] SUN Zhongxiang, SI Zihua, ZANG Xiaoxue, et al. KuaiSar: a unified search and recommendation dataset [C]//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. Birmingham: ACM, 2023: 5407−5411.

[110] GAO Chongming, LI Shijun, LEI Wenqiang, et al. KuaiRec: a fully-observed dataset and insights for evaluating recommender systems[C]//Proceedings of the 31st ACM International Conference on Information & Knowledge Management. Atlanta: ACM, 2022: 540−550.

[111] JIN Wei, MAO Haitao, LI Zheng, et al. Amazon-M2: a multilingual multi-locale shopping session dataset for recommendation and text generation[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans: ACM, 2023: 8006−8026.

[112] LI R, EBRAHIMI KAHOU S, SCHULZ H, et al. Towards deep conversational recommendations[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal: ACM, 2018: 9748−9758.

[113] LIU Yuanxing, ZHANG Weinan, DONG Baohua, et al.

U-need: A fine-grained dataset for user needs-centric e-commerce conversational recommendation[C]//Proceedings of the 46th international ACM SIGIR Conference on Research and Development in Information Retrieval. Taipei: ACM, 2023: 2723−2732.

[114] HAYATI S A, KANG D, ZHU Q, et al. INSPIRED: Toward sociable recommendation dialog systems[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Online: ACL, 2020: 8142−8152.

[115] ZHANG Jiaqi, CHENG Yu, NI Yongxin, et al. NineRec: A benchmark dataset suite for evaluating transferable recommendation[J]. IEEE transactions on pattern analysis and machine intelligence, 2024(1): 1−12.

[116] DING Yijie, HOU Yupeng, LI Jiacheng, et al. Inductive generative recommendation via retrieval-based speculation[EB/OL]. (2024−11−03)[2025−04−20]. http://arxiv.org/abs/2410.02939.

[117] XU Yuanbo, WANG En, YANG Yongjian, et al. GS-RS: A generative approach for alleviating cold start and filter bubbles in recommender systems[J]. IEEE transactions on knowledge and data engineering, 2023, 36(2): 668−681.

[118] BAI Haoyue, HOU Min, WU Le, et al. Gorec: a generative cold-start recommendation framework[C]//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM, 2023: 1004−1012.

[119] HUANG Feiran, BEI Yuanchen, YANG Zhenghang, et al. Large language model simulator for cold-start recommendation[C]//Proceedings of the 18th ACM International Conference on Web Search and Data Mining. Hannover: ACM, 2025: 261−270.

[120] KUSANO G. Data Augmentation using reverse prompt for cost-Efficient cold-start recommendation[C]//Proceedings of the 18th ACM Conference on Recommender Systems. Bari: ACM, 2024: 861−865.

[121] WU Xuansheng, ZHOU Huachi, SHI Yucheng, et al. Could small language models serve as recommenders? Towards data-centric cold-start recommendation[C]//Proceedings of the ACM Web Conference 2024. Singapore: ACM, 2024: 3566−3575.

[122] JIANG Yuezihan, CHEN Gaode, ZHANG Wenhan, et al. Prompt tuning for item cold-start recommendation[C]//Proceedings of the 18th ACM Conference on Recommender Systems. Bari: ACM, 2024: 411−421.

[123] FANG Yi, WANG Wenjie, ZHANG Yang, et al. Large language models for recommendation with deliberative user preference alignment[EB/OL]. (2025−02−17)[2025−04−20]. http://arxiv.org/abs/2502.02061.

[124] LIAO Jiayi, HE Xiangnan, XIE Ruobing, et al. RosePO: Aligning LLM-based recommenders with human values[EB/OL]. (2024−10−16)[2025−04−20]. http://arxiv.org/abs/2410.12519.

[125] SHAO Minglai, HUANG Hua, PENG Qiyao, et al. UL-MRec: user-centric large language model for sequential recommendation[EB/OL]. (2024−12−07)[2025−04−20]. http://arxiv.org/abs/2412.05543.

[126] DENG Jiaxin, WANG Shiyao, CAI Kuo, et al. OneRec: Unifying retrieve and rank with generative recommender and iterative preference alignment[EB/OL]. (2025−02−26)[2025−04−20]. http://arxiv.org/abs/2502.18965.

[127] GAO Chongming, CHEN Ruijun, YUAN Shuai, et al. SPRec: Self-play to debias LLM-based recommendation[C]//Proceedings of the ACM on Web Conference 2025. Sydney: ACM, 2025: 5075−5084.

[128] BIN Xingyan, CUI Jianfei, YAN Wujie, et al. Real-time indexing for large-scale recommendation by streaming vector quantization retriever[EB/OL]. (2025−01−15)[2025−04−20]. http://arxiv.org/abs/2501.08695.

[129] ZHOU Yujia, YAO Jing, DOU Zhicheng, et al. ROGER: Ranking-oriented generative retrieval[J]. ACM Transactions on information systems, 2024, 42(6): 1−25.

[130] YANG Yuhao, JI Zhi, LI Zhaopeng, et al. Sparse meets dense: Unified generative recommendations with cascaded sparse-dense representations[EB/OL]. (2025−03−04)[2025−04−20]. http://arxiv.org/abs/2503.02453.

[131] PENHA G, VARDASBI A, PALUMBO E, et al. Bridging search and recommendation in generative retrieval: Does one task help the other?[C]//Proceedings of the 18th ACM Conference on Recommender Systems. Bari: ACM, 2024: 340−349.

[132] CHEN Sirui, HAN Shen, CHEN Jiawei, et al. Rankformer: A graph transformer for recommendation based on ranking objective[C]//Proceedings of the ACM on Web Conference 2025. Sydney: ACM, 2025: 3037−3048.

[133] WANG Yilin, HU Minghao, HUANG Zhen, et al. KC-GenRe: a knowledge-constrained generative re-ranking method based on large language models for knowledge graph completion[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino: ACL, 2024: 9668−9680.

[134] YAN Bencheng, LIU Shilei, ZENG Zhiyuan, et al. Unlocking scaling law in industrial recommendation systems with a three-step paradigm based large user model[EB/OL]. (2025−02−12)[2025−04−20]. http://arx-

iv.org/abs/2502.08309.

[135] JIA J, WANG Y, LI Y, et al. LEARN: knowledge adaptation from large language model to recommendation for practical industrial application[C]//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia: ACM, 2025: 11861−11869.

[136] ZHAO Qian, QIAN Hao, LIU Ziqi, et al. Breaking the barrier: utilizing large language models for industrial recommendation systems through an inferential knowledge graph[C]//Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. Boise: ACM, 2024: 5086−5093.

[137] HU Jun, XIA Wenwen, ZHANG Xiaolu, et al. Enhancing sequential recommendation via llm-based semantic embedding learning[C]//Companion Proceedings of the ACM Web Conference 2024. Singapore: ACM, 2024: 103−111.

[138] LIU Enze, ZHENG Bowen, LING Cheng, et al. End-to-end learnable item tokenization for generative recommendation[EB/OL]. (2025−03−12)[2025−04−20]. http://arxiv.org/abs/2409.05546.

[139] 孟岱. 工业大模型的落地难题[J]. 中国工业和信息化, 2024(4): 26−32.
MENG Dai. The landing challenge of industrial large models[J]. China industry and information technology, 2024(4): 26−32.

[140] LIU Fan, CHENG Zhiyong, CHEN Huilin, et al. Privacy-preserving synthetic data generation for recommendation systems[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. Madrid: ACM, 2022: 1379−1389.

[141] 黄河燕, 李思霖, 兰天伟, 等. 大语言模型安全性: 分类、评估、归因、缓解、展望[J]. 智能系统学报, 2025, 20(1): 2−32.
HUANG Heyan, LI Silin, LAN Tianwei, et al. A survey on the safety of large language model: classification, evaluation, attribution, mitigation and prospect[J]. CAAI transactions on intelligent systems, 2025, 20(1): 2−32.

[142] GUO Daya, YANG Dejian, ZHANG Haowei, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning[EB/OL]. (2025−01−22)[2025−04−20]. http://arxiv.org/abs/2501.12948.

## 作者简介:

石磊, 副研究员, 中国人工智能学会智能服务专委会委员, 主要研究方向为智能信息处理、大数据分析与挖掘、社交网络搜索及人工智能。发表学术论文 40 余篇。E-mail: leiky_shi@cuc.edu.cn。

赵雨秋, 硕士研究生, 主要研究方向为推荐系统与信息检索。E-mail: yuqiuzhao@mails.cuc.edu.cn。

袁瑞萍, 教授, 主要研究方向为复杂物流系统、数据分析与智能决策。主持国家自然科学基金项目 1 项, 发表学术论文 40 余篇。E-mail: yuanruiping@bwu.edu.cn。