智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于大语言模型的推荐系统综述

谢广明, 白彦冰, 吴子昂, 张艳玲

引用本文:
谢广明, 白彦冰, 吴子昂, 等. 基于大语言模型的推荐系统综述[J]. 智能系统学报, 2025, 20(6): 1520-1533.
XIE Guangming, BAI Yanbing, WU Ziang, et al. Review of LLM-based recommendation systems[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(6): 1520-1533.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202410007

DOI: 10.11992/tis.202410007
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20251009.1818.006

基于大语言模型的推荐系统综述

谢广明[1], 白彦冰[1], 吴子昂[2], 张艳玲[2]
(1. 北京大学 工学院, 北京 100871; 2. 北京科技大学 智能科学与技术学院, 北京 100083)

摘　要: 随着社交网络平台和电子商务平台的崛起, 工业级个性化推荐系统在移动互联网时代的作用日益显著, 对提升用户浏览体验、购物体验以及扩大用户规模起到了不可替代的作用。在推荐系统中, 模型发挥着至关重要的作用。随着算力和数据量的增长, 模型结构呈现复杂化、大型化趋势, 推荐精准度相较于传统推荐模型也有显著提升。以 GPT 和 DeepSeek 为代表的大语言模型(large language model, LLM), 不仅显著改善了语言模型的效果, 而且助推了提示工程等训练范式的发展。LLM 所具备的语义理解和内容生成能力, 使其在工业级推荐系统中的落地应用正处于快速发展阶段。本文对 LLM 和推荐系统的结合点进行调研, 梳理了 LLM 与工业级推荐系统的结合方式, 并提出了对 LLM 和推荐系统结合的展望, 以期利用 LLM 技术提升推荐模型的训练、推理效率和效果。

关键词: 推荐系统; 推荐模型; 大语言模型; 提示工程; 规模效应; 视觉大模型; 序列建模; 深度学习
中图分类号: TP391.3　　**文献标志码:** A　　**文章编号:** 1673−4785(2025)06−1520−14

中文引用格式: 谢广明, 白彦冰, 吴子昂, 等. 基于大语言模型的推荐系统综述 [J]. 智能系统学报, 2025, 20(6): 1520–1533.
英文引用格式: XIE Guangming, BAI Yanbing, WU Ziang, et al. Review of LLM-based recommendation systems[J]. CAAI transactions on intelligent systems, 2025, 20(6): 1520–1533.

Review of LLM-based recommendation systems

XIE Guangming[1], BAI Yanbing[1], WU Ziang[2], ZHANG Yanling[2]
(1. College of Engineering, Peking University, Beijing 100871, China; 2. School of Intelligence Science and Technology, University of Science and Technology Bejing, Beijing 100083, China)

Abstract: As social networking and e-commerce platforms have grown in popularity, industrial recommendation systems have assumed an increasingly significant role in the mobile internet era. The recommendation system is imperative for enhancing user experience, optimizing shopping experience, and promoting user growth. In the domain of recommendation systems, the role of models is paramount. As computing power and data volume have increased, model structures have become increasingly complex. These models have also improved the accuracy of recommendation systems in comparison to traditional models. Represented by GPT and DeepSeek, LLM has been demonstrated to enhance the efficacy of language models and catalyze the evolution of novel model training paradigms, such as prompt engineering. The rapid advancements in large language model (LLM) capabilities, particularly in semantic understanding and content generation, are poised to transform industrial recommendation systems. This paper reviews the connections between LLM and recommendation systems, then outlines the ways in which LLM can be integrated with industrial recommendation systems. The objective of our work is to leverage technologies associated with LLM to enhance the efficiency and efficacy of recommendation models.

Keywords: recommendation system; recommendation models; LLM; prompt engineering; scaling law; visual large model; sequential modeling; deep learning

推荐系统和 LLM 是当今人工智能领域的研究热点。随着以抖音、快手、TikTok、Facebook 为代表的社交网络平台和以淘宝、京东、拼多多、Amazon 为代表的电子商务平台的崛起, 以及大语言模型(large language model, LLM)的发展, 推荐效果和语义生成效果在日趋改善, 而算力消耗也在日趋增加[1-3]。如何将 LLM 与推荐算法进行结合, 使得二者在工业级推荐系统中发挥更大作用, 是当前有待探索的领域。由于算力和效率方面的约束, LLM 在工业级推荐系统的落地面临诸多挑战。本工作旨在对工业级推荐系统和 LLM 的基本现状, 以及二者的结合方式进行综述。

收稿日期: 2024−10−09.　**网络出版日期:** 2025−10−10.
基金项目: 国家自然科学基金项目(62033010, 61603036).
通信作者: 张艳玲. E-mail: yanlzhang@ustb.edu.cn.

1　研究背景

1.1　推荐系统

推荐系统运转的过程是一个协同过滤[1]的过程, 系统根据人与人之间的相似度为用户推荐感兴趣的物品（图 1）。

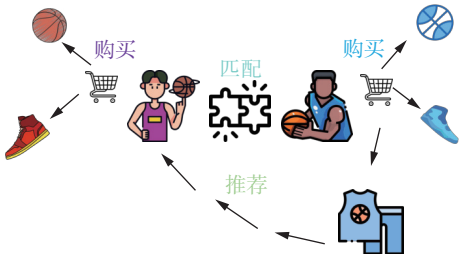


图 1　协同过滤示意

Fig. 1　Schematic of collaborative filtering

个性化推荐[1]是通过用户的兴趣喜好, 为用户进行个性化推荐物品或内容的过程, 如图 2 所示。随着信息技术的发展, 尤其是移动互联网的发展, 推荐已经渗透到人们生活的各个方面。典型的推荐场景包括新闻资讯、视频、直播、商品等。在互联网时代, 推荐更严格的定义是通过数据和算法, 建立人和物品的关联, 辅助用户进行决策。推荐系统也是基于人工智能算法构建的智能体, 通过感知用户意图和兴趣偏好, 为用户输出内容, 收集用户反馈, 从而不断学习和进化。

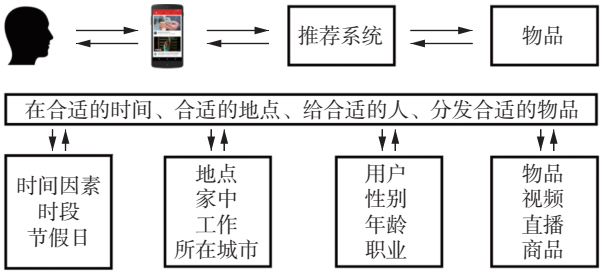


图 2　推荐系统示意

Fig. 2　Schematic of recommendation system

1.2　工业级推荐系统

传统的协同过滤算法可以衍生出用户协同(User-CF)、物品协同(Item-CF)、矩阵分解等算法, 这些算法更多是应用于小规模推荐系统上。现代工业界推荐系统面临的是数以亿计的用户和物品, 以及百亿规模的训练样本。在诸如 YouTube、Facebook、TikTok、快手、抖音[2-3]等社交媒体平台中, 核心的流量分发逻辑建立在深度学习推荐系统上。用户在这些平台进行内容浏览期间, 产生数以百亿计的行为数据。这些行为数据被用于深度学习模型训练, 使得推荐系统学习到用户的短期兴趣和长期兴趣, 推荐出更符合用户偏好的内容, 从而提升平台的用户黏性和商业价值。

表 1 列举了当前互联网亿级用户主流平台的数据规模[4], 根据公开披露的用户规模、物品规模、用户行为密度和推荐自由度, 可以推断出推荐系统所面临的数据规模和复杂度。现代工业级推荐系统已经脱离对协同过滤算法的简单应用, 演化成基于大规模深度学习的复杂系统。

表 1　主流推荐系统的数据规模[4]

Table 1　The scale of mainstream recommendation systems[4]

平台	用户规模(DAU)/亿	物品规模	行为密度	推荐自由度
微信–朋友圈	9	千级	高	低
抖音	6	十亿级	高	高
快手	3	十亿级	高	高
百度	5	十亿级	低	低
Facebook	20	百万级	高	低
X	2.29	百万级	高	低

更大的数据规模意味着更复杂的系统设计。推荐系统的一个重要工作是排序, 即为用户的每一次请求, 对可能的候选进行打分排序。在数据规模和算力的约束下, 推荐分为召回、粗排、精排等多阶段[5]。如图 3 所示, 从上到下每个阶段所处理的候选规模依次降低, 算法复杂度依次增加, 最终选出 Top K (K 一般为 10~20 量级）个候选推送给用户进行展示。在推荐系统的每个阶段, 需要对用户–物品对进行多目标的模型预估。预估用户在物品上的多种行为概率与价值, 将这些预估概率和价值进行融合形成一个综合打分, 用于各个阶段的排序。

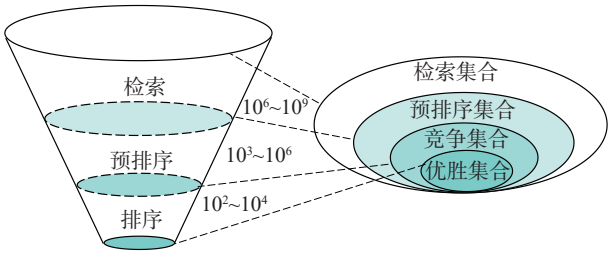


图 3　推荐系统的典型处理漏斗[5]

Fig. 3　Typical processing funnel for recommendation systems[5]

除了数据规模之外, 推荐系统还面临着时效性、多目标、数据噪声等挑战。首先, 现代推荐系统所服务的平台中, 用户和内容都具有高动态性特点, 即用户即来即走, 内容随时上传更新, 尤其对于直播内容来讲, 更属于边生产边分发的过程。这样的高动态性要求推荐系统在数据更新、模型训练、系统响应上具备秒级别甚至毫秒级别的时效能力。其次, 工业界中推荐优化目标是多

样性的，以抖音和快手为例，存在观看时长、有效互动、负向反馈、商品购买、浏览体验等多方面的优化目标。这些目标如何通过模型精准估算，并且进一步更好地融合，是推荐系统持续解决的难题。最后，推荐系统是利用用户的行为优化模型和策略，为用户推荐出适合的内容。然而，用户的行为本身存在噪声，这给推荐系统的学习带来巨大困难。因此，如何利用好深度学习和 LLM 提升推荐精准度，成为工业级推荐系统的关键。

1.3　LLM

语言模型的发展按时间先后分为以下 4 个阶段：统计语言模型、神经网络语言模型、预训练语言模型、LLM[6]。统计语言模型，例如 N-Gram 和马尔可夫模型，主要利用统计方法来计算词语之间的概率关系，从而预测下一个词出现的概率。神经网络模型，例如 LSTM[7]、RNN[8] 和 Transformer[9] 等，使用神经网络来学习词语之间的复杂关系，从而更好地预测下一个词出现的概率。预训练语言模型的早期尝试可以追溯到 2018 年提出的基于 Bi-LSTM 的方法 ELMo(embeddings from language models)[10]。早期的语言模型通常只在一个特定的任务上进行训练，而 ELMo 采用了一种新的训练方法，即“预训练 + 微调”范式：在预训练阶段，语言模型通常在大型语料库上进行训练；在微调阶段，则在特定的任务上进行微调。BERT(bidirectional encoder representations from Transformers)[11]、GPT(generative pretraining Transformer)[12] 等模型提出后，基于 Transformer 的预训练语言模型发展迅速，诸多研究者发现，采用更大规模的参数量能够在下游任务上产生较大效果提升。于是，涌现了拥有 1 750 亿参数量的 GPT-3、54 亿参数量的 PaLM(pathways language mode) 等大语言模型。

LLM 通常可以分为 3 类：Encoder-Decoder、Encoder-Only 和 Decoder-Only 模型。在 LLM 发展的早期阶段，Encoder-Decoder 架构以及 Encoder-Only 架构占据研究的主流，在文本内容分析、分类等任务上表现较好。Encoder-Decoder 的代表模型有 Meta 研发的 BART[13] 以及 Google 研发的 T5[14] 等。Encoder-Only 的代表模型是 BERT，并衍生出 ERNIE(enhanced language representation with informative entities)[15]、RoBERTa(robustly optimized BERT pretraining approach)[16]、ALBERT(a lite BERT for self-supervised learning of language representations)[17] 等模型。这两类模型均属于判别式模型，并且其与训练任务都是预测掩码单词。在 OpenAI 发布 Decoder-Only 的 GPT-3 后，Decoder-Only 模型由于其良好的生成能力和泛化性能逐渐成为 LLM 的主流结构，这里的泛化性能指的是下游任务上的 Zero-Shot 或 Few-Shot 的泛化性能。与其他两种模型的不同之处在于，Decoder-Only 模型属于生成式模型，预训练任务主要是对下一个词进行预测。常见的 Decoder-Only 模型包括 DeepSeek-V3[18]、DeepSeek-R1[19]、GPT-4[20]、GPT-4v、GPT-4o、PaLM、OPT[21]、Bard、Llama[22]、Llama2[23]、Llama3[24] 等。各类别对应的常见模型梳理如图 4 所示。

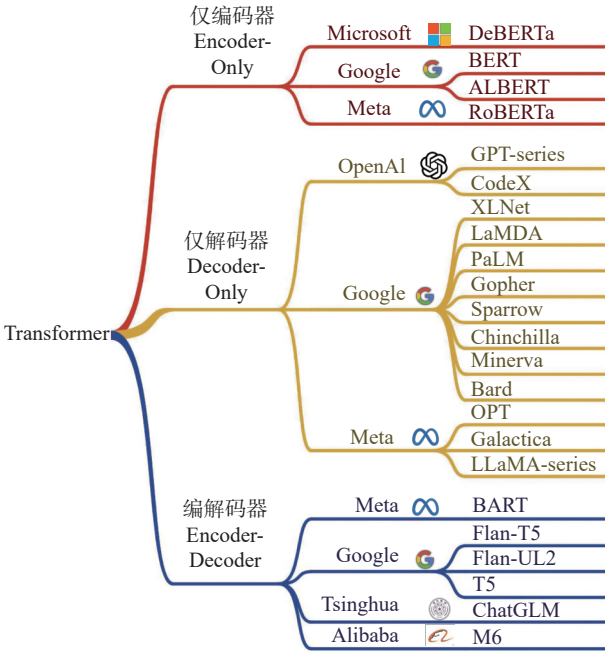


图 4　各个架构对应的常见 LLM 分类

Fig. 4　Common LLM classification of each architecture

Decoder-Only 模型分支中的 GPT 模型自 3.0 版本以来在 LLM 的众多模型中始终处于领跑位置。在此过程中，OpenAI 提出的诸多关键性技术已经逐渐成为 LLM 训练和微调的新范式。GPT-1 主要为 GPT 系列的模型结构打下基础，其参数量约为 1.17 亿，此后的模型都是在此模型核心结构上进行的改动。GPT-2 在 GPT-1 基础上将参数增大量到 1.5×10^9，并在约 40 GB 的数据上进行充分预训练。此外，GPT-2 主要对零样本的多任务学习进行了探索，在 GPT-1 基础上去掉了 Fine-Tuning 层，不再在具体任务上进行微调，在零样本多任务学习上取得了初步成果。但是，此模型与经过微调训练的模型相比，在具体任务上的性能仍有差距。GPT-3 将参数规模大幅提升至 1 750 亿，提出了一种上下文学习的大规模学习范式 ICL(In-Context-Learning)，将下游任务表达为示例和提示的形式作为模型的前置输入，在此基础上让

GPT 执行相关任务。ICL 包含 Zero-Shot、One-Shot、Few-Shot 等类型, 其划分准则是示例的个数。

与 GPT-2 类似, GPT-3 不再需要为每个具体的下游任务微调生成模型, 减少了重复训练, 同时大大增强了预训练模型的泛化能力。事实上, GPT-3 也是同时期参数量最大的 LLM, 证明了参数量的增大可以提升模型能力, 这可视作 LLM 发展的一个重要转折点。在确立了模型架构和训练范式之后, OpenAI 针对 GPT-3 的薄弱领域, 例如代码编写、推理能力和数学能力, 通过微调进行了提升。此外, GPT-3 还采用了强化学习算法 RLHF[24-25] (reinforcement learning from human feedback), 使得 LLM 的回答能够在用户层面进行对齐, 通过显式引入人的偏好提升用户满意度, 该算法主要是通过收集人类对于回答的反馈从而改进模型。GPT-4 的出现则掀起了多模态大语言模型的研究热潮, 但是 GPT-4 的技术报告没有公布多模态能力的具体实现方式。同时期开源多模态大模型 LLaVA(large language and vision assistant)[26] 则利用了 CLIP(contrastive language-image pretraining)[27] 的图像编码器部分生成能够和文本对齐的特征, 辅助 LLM 进行图像理解, 从而在多模态任务上取得了一定成果。

LLM 具备 3 个与规模相关的特点: 1)语料规模, 几乎包含了全部公开的数字化语料; 2)参数规模, 经典的 GPT-3 的参数是 1 750 亿, 具备极强的表征能力; 3)算力规模, 以 NVIDIA 的 A100 型号 GPU 卡为例, 完成 GPT-3 的训练, 在 1 000 张卡的规模上需要训练 20 天以上, 极大提升了 LLM 基座模型训练的门槛。

1.4　LLM 为推荐带来的改变

如图 5 所示, LLM 的出现带来的不仅是效果和 AI 泛化能力的提升, 同时也带来模型范式上的变化。

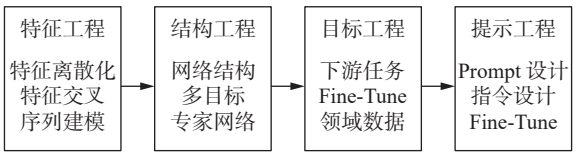


图 5　模型训练演进的不同阶段及其核心工作

Fig. 5　Evolution of model training stages and the associated key works

第 1 阶段, 特征工程。以自然语言处理领域为例, 在机器学习的早期阶段, 自然语言学习的工作以特征工程为主, 主要工作集中在如何构造更有效的特征和表征来提升模型效果。

第 2 阶段, 结构工程。当深度学习开始大规模应用后, 模型训练进入到结构工程阶段, 即通过不断迭代深度学习网络的结构来达到更好的效果。这个阶段涌现出了 LSTM、GRU[28] 等序列神经网络模型, 同时也诞生了 Transformer、BERT 等基于注意力机制的模型。

第 3 阶段, 目标工程。随着模型复杂度的增加, 尤其是 BERT 的诞生, 从零训练单个模型的训练成本变高。但由于不同下游任务存在差异, 因此衍生出预训练加微调模式为基础的任务工程范式, 即采用通用语料训练一个复杂的基础模型, 例如 BERT。在应用到具体任务时, 以具体任务的语料为主, 对部分参数进行神经网络的微调, 从而达到适配效果。

第 4 阶段, 提示工程。以 Instruct GPT[12] 为早期代表的 LLM 出现后, 即便针对具体任务, 也难以有足够资源进行整个模型的微调 (Fine-Tune)。针对这种情况, 进一步衍生出预训练−提示−预测三阶段的提示工程模式。相比于直接微调整个模型, 提示工程不需要对预训练模型进行变更, 而是为其提供有效的提示 (Prompt), 通过提示来达成模型和下游任务之间的适配。如何针对下游任务给出合适提示这个环节, 这一环节被称为提示工程[29]。

以上模型领域的演进, 尤其是在语言模型上的演进, 为推荐系统带来很多有益的启发。现有推荐系统模型的工作, 对应到上面的模型领域的演进过程, 更多是停留在特征工程和结构工程阶段, 主要工作是在定义问题建模的基础上, 挖掘特征、优化表征、改进网络结构以获得更好的效果。一个自然的问题是, 如何推动推荐系统模型更加复杂化, 让模型更好地适配多场景、多任务、多业务, 从而结合 LLM 带来更好的推荐效果。近年来已经有诸多工作对 LLM 在推荐中的应用进行总结[30-32], 本文重点探讨推荐系统如何利用 LLM 技术进行改进。

2　模型在推荐系统中的作用

推荐系统的核心任务是利用用户行为数据, 在给定的时间和场景下, 预估用户对物品的交互概率。整个数据生产到深度学习的过程如图 6 所示。

首先, 用户接收到推荐系统推荐的内容, 在推荐内容上进行点击、观看、购买等反馈; 其次, 这些反馈连同用户、内容、上下文的特征, 被构造成训练样本被系统收集起来。当大量的样本收集到一起则形成多个批次, 通过流式传输送给深度学习平台; 平台利用样本进行训练, 利用梯度下降、

牛顿法、Adam、AdGrad 等数值优化算法学习参数集合。最后，训练好的参数通过参数服务器生效到线上服务，进行点击、观看、购买等概率的推理，系统根据推理结果对内容进行排序，如此循环往复。

推荐系统的核心环节包括召回、粗排、精排 3 个阶段（图 4）。在这 3 个阶段中，机器学习模型均起到核心作用。在现代工业级推荐系统中，模型以深度学习模型为主，浅层模型为辅。从召回到粗排再到精排，所处理的数据量级依次降低，但模型的复杂度依次提升。下面分别介绍模型在排序（包括精排和粗排）和召回中的应用。

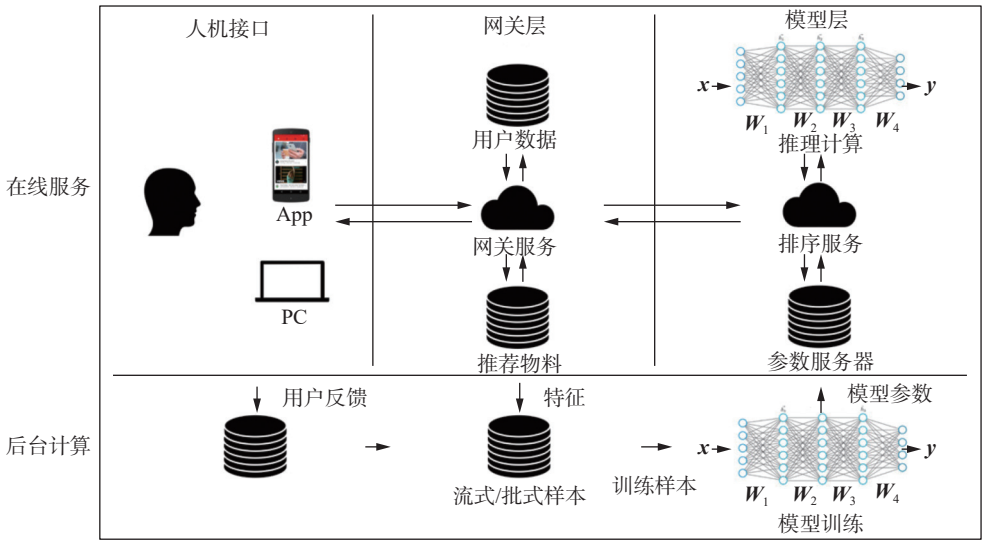


图 6　推荐模型的生命周期

Fig. 6　Life cycle of recommendation model

2.1　排序模型

排序模型在推荐系统中处于核心的地位，具有较高复杂度。图 7 给出了 2016 年 Google 在 YouTube 视频推荐系统[2] 中的精排模型结构，这是视频推荐领域的经典模型。模型整体为多层神经网络结构，最底层为基础特征，包括稠密特征和稀疏特征。稀疏特征通过稀疏编码映射到嵌入式向量的稠密空间。以日期特征和兴趣特征为例，首先通过一个离散向量进行表达：

$$\text{WeekDay} = \text{Mon} \Rightarrow \boldsymbol{x}_{\text{WeekDay,hot}} = [1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]^{\text{T}} \tag{1}$$

$$\text{Topic} = \{\text{Funny}, \text{Pet}\} \Rightarrow \boldsymbol{x}_{\text{Topic,hot}} = [\cdots \quad 0 \quad 1 \quad 0 \quad \cdots] \tag{2}$$

然后，通过一个向量矩阵将原始的稀疏向量压缩到一个低维度（比如 128 维）空间。特征映射之后，稠密特征和稀疏特征也会被进一步处理，通过交叉、高次变换等操作形成高阶的特征表达，最后作为神经网络第 1 层的输入，经过神经网络的前向传递计算得到激活值，再经过激活函数得到预估的点击率、预估观看时长等。

以上是排序模型的经典结构。实际上，随着数据规模的增长，简单的多层感知机神经网络已经不再满足预估精度的要求。因此，网络结构也在不断演化，从最基础的宽度与深度模型，经过演化出现了 DeepFM(a factorization-machine based neural network)[33-34]、DCN(deep and cross network)[35-36]、MMOE(multi-gate mixture of experts)[37]、PLE(progressive layered extraction)[38] 等复杂结构。

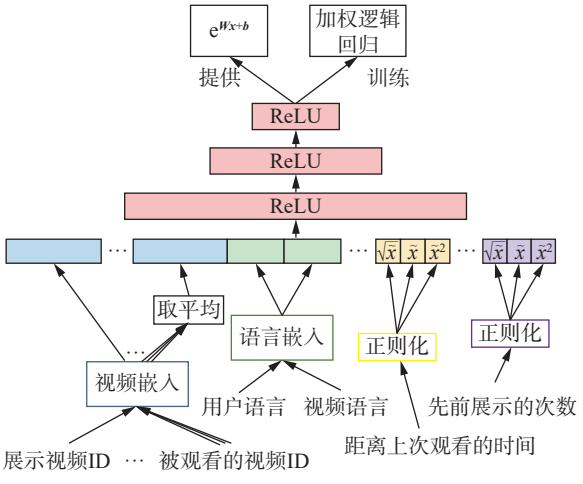


图 7　基于深度神经网络的典型推荐系统

Fig. 7　Typical recommendation system based on deep neural networks

2023 年发表的工作[39] 采用 LHUC (learning hidden unit contribution) 结构[40]，将精排模型的复杂度进一步提升。引入 LHUC 结构后，模型中每个神经元都具备感知个性化的能力，在百亿级别样本的模型训练上带来很大的性能提升。LHUC 的神经元：

$$\delta = \gamma \cdot \text{Sigmoid}(\boldsymbol{x}\boldsymbol{W} + \boldsymbol{b}),\ \delta \in [0,\gamma] \tag{3}$$

式中: $\boldsymbol{x}$ 为特征, $\boldsymbol{W}$ 和 $\boldsymbol{b}$ 为模型参数。经过 Sigmoid 激活函数的作用, 得到个性化激活值, 该激活值与原有神经网络神经元的输出值相乘, 作为下一层神经网络的输入。

模型构建中, 另一个比较重要的课题是用户行为序列的使用。推荐系统感知用户的兴趣, 一方面的感知来自样本的直接反馈信号, 可学习到用户对推荐物品的偏好程度; 另一方面的感知来自用户的行为历史, 例如, 通过用户在短视频或直播方面的观看历史、商品点击和购买历史, 来推断用户接下来对哪类内容和商品存在更强的偏好。将用户行为历史进行建模利用的过程称作序列建模, 这是推荐模型领域一个持续研究的课题。在深度学习模型应用到推荐系统的早期阶段, 由于算力方面的限制, 使用的用户行为序列较短, 一般为 100 个以内。使用方式也比较简单, 即对序列中每个元素单独学习隐式向量表征 (embedding), 并进行向量的加和池化与平均池化操作 (sum pooling or average pooling)。随着系统演进, 以及推荐系统需要从更长的用户行为历史中挖掘有效信息, 序列建模的复杂度也越来越高, 所采用的用户行为序列也在逐步扩大。Deep Interest Network[41] 首先提出用注意力机制来实现序列历史中的差异化权重。DIEN(deep interest evolution network)[42] 提出用 GRU 序列神经网络对序列进行建模。SIM(search-based user interest modeling)[43] 算法采用 1 000以上长度的长序列, 通过两阶段对序列建模, 然而, 随着序列长度的增加, 两阶段建模带来的不一致性会加剧。

2023 年快手提出的 TWIN(two-stage interest network)[44] 工作使用了用户全生命周期的行为序列, 并实现了超长序列的端到端建模能力, 解决了两阶段建模带来的不一致问题。该方法一方面使得建模序列覆盖了用户的全生命周期, 充分挖掘了用户的短期、中期、长期兴趣; 另一方面, 通过 GPU 算子优化, 实现了万级别长度序列上的端到端建模。

2.2　模型在召回阶段的应用

推荐系统的召回阶段负责从海量物品候选中, 快速将用户感兴趣的物品集合进行圈定。相比于精排模型, 召回模型保证了足够的召回率。召回一般分为规则召回和模型召回, 规则召回适合将一定行业先验知识在推荐系统落地, 诸如热门召回、地理位置信息召回、基于标签体系的兴趣召回等。模型召回则面向解决更复杂的召回问题, 在满足召回率基础上大幅提升召回精准度。随着深度学习在排序模型落地方案的成熟, 模型召回开始占据主流, 从简单的用户协同过滤召回、物品协同过滤召回, 再到经典的双塔召回。近几年, 模型整体趋于复杂化, 产生了 TDM 召回[45]、Deep Retrieval[46] 召回等复杂模型。

随着 LLM 训练的兴起, 业界逐渐认识到基于稀疏 ID 推荐系统和基于监督学习的推荐系统的局限性。因此, 涌现出针对生成式召回的研究, 典型的工作是 VQ-VAE(vector quantized variational autoencoder)[47] 和 RQ-VAE(residual quantized variational autoencoder)[48], 它们尝试将召回工作改进成一个带有语义 ID 生成的过程。这样摆脱了对原有系统随机 ID 的依赖, 也为推荐系统进化为基于语义的 LLM 提供了输入。如图 8 所示, 通过维护一个对向量做离散化的 Codebook, 不断根据 Codebook 生成具备语义的 ID, 同时 ID 的生成通过与编码器和解码器的端到端联合训练形成联动。

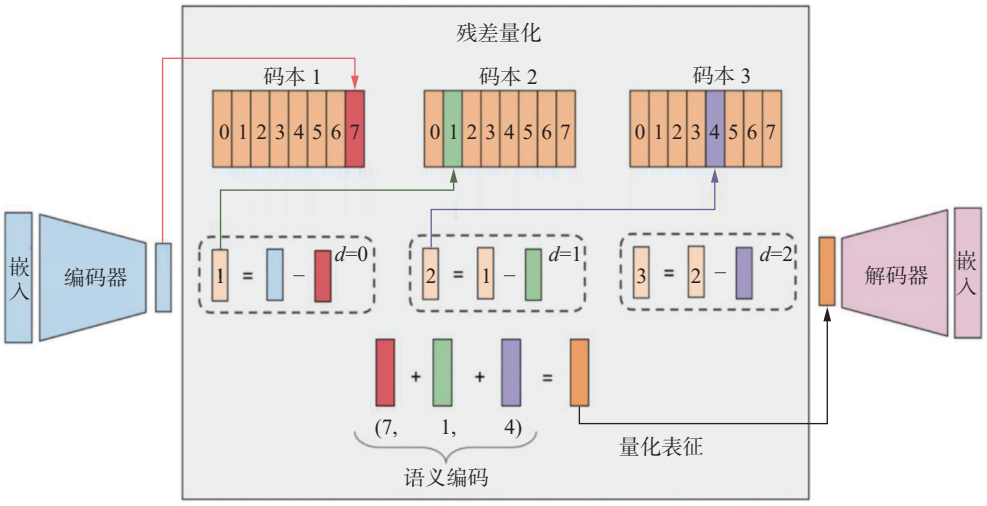


图 8　生成式召回通过 Codebook 的更新和变分推断生成具有语义的 ID[48]

Fig. 8　Generative recall using Codebook updates and variational inference to produce semantically meaningful IDs[48]

2.3　推荐系统模型的局限和挑战

虽然模型训练在推荐系统中起着核心作用，但现有的模型并不能解决推荐系统的全部问题，依然存在巨大的提升空间。

首先，从推荐粒度上，当前推荐系统模型依然是学习用户的行为反馈，包括显式反馈和隐式反馈，而推荐打分依然是针对整个物品（视频、直播、商品、地理位置兴趣点等）。实际上，用户对物品的关注可能存在于诸多方面，比如用户对视频的兴趣点仅在于某个片段，用户对直播的兴趣点可能仅存在某个时刻。这意味着推荐系统的模型会朝着更细的预估粒度演进。当前主流的工业级推荐系统的参数规模为百亿规模，对应亿级别用户规模和亿级别的候选规模。当建模粒度进一步细化，必然带来参数规模的增加和性能开销。这方面带来的性能挑战可以从 LLM 的部署和性能优化上找到具有借鉴意义的经验。

其次，推荐模型本质上是在学习用户的后验行为，而这会导致系统存在信息茧房、样本选择偏差、多样性和体验问题。这些问题需要从样本空间、建模目标、上线规范和指标多方面入手解决。然而，这方面工作的模型化程度还不够，更多依赖人工经验。LLM 领域的 RLHF(reinforcement learning from human feedback)[24-25] 则提供了一个很好的思路，通过与人类反馈对齐的方式指导模型训练。

最后，推荐模型目前的训练迭代主要是在特征工程和深度模型结构上，而且很难达到推荐模型跨场景、跨领域甚至跨产品的通用。LLM 的出现不仅改善了语言交互的效果，同时带来模型学习范式的改变，使得基座模型、微调和提示工程成为模型训练的新范式。这也是推荐模型走向基座化，进而实现跨场景、跨领域轻量级适配建模的演进路径之一。

3　LLM 在推荐系统中的应用

LLM 在自然语言、视觉生成等领域取得突破之后，其在推荐系统领域如何应用的研究也逐步展开。LLM 在推荐系统中的应用具有 6 个切入点[31-32]。1）基于 LLM 对推荐内容进行表征，提升推荐系统对内容的理解能力。2）基于互联网海量数据训练推荐系统基座模型，并采用 Prompt[35] 和 Fine-Tuning 微调技术，以推荐基座模型为基础，针对多场景和多任务进行高效地下游适配。3）基于 LLM 进行交互式推荐。4）多模态大模型改进推荐效果。5）基于工业级场景探索推荐 LLM 结构以及规模效应。6）基于 LLM 的推荐系统部署和推理效率优化。

3.1　基于 LLM 的表征

现代工业推荐系统大量使用基于用户 ID 和物品 ID 的稀疏表征方法，即为每个用户和每个物品学习一个向量表征作为系统的输入。该学习方式的优点在工业大规模推荐系统中得到了放大。在数据量和监督信号量充足的场景下，基于 ID 的表征能很好地从协同过滤角度学习到物品−物品、用户−物品之间的强相关性。在此背景下，基于内容的标签体系及基于内容的表征向量在大规模推荐系统中的表征能力被削弱。而 LLM 的出现使得模型对内容的理解达到了新层次。在 ID 表征方面，原有系统中绝大部分 ID 都是随机生成，ID 之间缺乏语义相关性。基于向量离散化 VQ-VAE[47] 和 RQ-VAE[48] 等方法使得 ID 具备语义含义，为 ID 进入 LLM 提供了可行性。基于具备语义的 ID 和 LLM 对语言的处理能力，业界已经出现基于 GPT-3 结合提示工程的推荐系统，能够基于文本语义协同提供推荐服务。

LLM 语义表征所表达的内容丰富度远高于传统的标签体系和传统推荐深度学习网络。在表征能力方面，LLM 提升了相对稀疏 ID 表达的竞争力。传统的标签体系和深度学习内容表征更多是在冷启动方面帮助系统更好地理解内容。过往工作已证明了基于 BERT 的表征在冷启动阶段和非冷启动阶段均取得了与稀疏特征相匹敌的结果。LLM 的加入能够将表征能力进一步提升。U-BERT[49] 将 BERT 用于构建用户编码器和评论编码器，引入了物品编码器来表示物品，同时提出了一个评论匹配层来捕捉用户和商品评论之间的语义交互，从而成功提升了推荐性能。与 U-BERT 类似，ZESREC(zero-shot recommender systems)[50]、UniSRec[51] 和 VQ-Rec[52]同样采用了 BERT 进行语义表征。ZESREC 主要探索了推荐系统中实现零样本学习的可行性。UniSRec 主要针对现有建模方法过于依赖无语义 ID，难以泛化到新推荐场景的问题，利用 BERT 实现文本编码用于语义表征，将物品的自然语言描述转化为文本嵌入，并且设计了一个包含 MoE[53](Mixture-of-Expert）的物品编码架构来学习通用表示。类似地，VQ-Rec 先将文本通过 BERT 表示成物品代码，然后使用物品代码来查找嵌入表，这些嵌入表使用 OPQ (optimized product quantization)[54] 构建，解决了直接映射文本编码到物品表示过程中性能下降的问题。

3.2　基于基座模型预训练和微调的推荐系统

3.2.1　基于海量数据进行 LLM 推荐系统预训练

模型预训练的主要目的是使得 LLM 能够在

海量未标注的数据中理解语言特征、学习到常识性知识, 从而具备生成连贯且符合语境句子的能力。推荐系统通过在海量的用户历史数据上进行预训练, 能够增强其对推荐任务的理解能力。通常, 不同结构所对应的模型预训练方法有所不同。对于判别式结构的 Encoder-Decoder 以及 Encoder-Only 模型, 其训练方式是随机掩盖序列中的部分词语, 要求 LLM 基于上下文预测被掩盖的部分, 该学习方式使得模型能够学习到双向的语义信息。对于生成式的 Decoder-Only 模型结构, 其训练任务是基于给定一段上下文预测下一个词。现有的 LLM 推荐系统也是根据其模型结构在以上两种方法中进行选择, 虽然并不完全一致, 但在思想上高度重合, 需针对具体的推荐任务进行改进。

Google 发布的 BERT4Rec[55] 预测用户下一个要交互的物品。它采用了第一种方法, 同时利用 BERT 的训练方法, 随机将输入序列的一部分掩盖, 让模型来预测遮盖部分对应的物品。与之不同, PTUM (pretrained user models)[56] 在预训练过程中同时完成了两种任务: 其一是掩码行为预测, 根据用户的其他行为推断用户隐藏的行为, 从而帮助模型捕捉历史用户行为之间的相关性; 其二是预测接下来的 K 个行为, 该任务使得模型能够根据过去的用户行为预测用户未来的 K 个行为, 从而帮助模型捕捉过去和未来行为的相关性。阿里提出的 M6 模型[57] 也采用了以上两个预训练任务, 即文本填充和自回归文本生成。通过随机掩盖包含多个词元的文本序列片段让模型进行预测, 从而提供评估推荐评分任务中文本合理性的能力; 在自回归语言生成任务中, 基于被掩盖的序列来预测未被掩盖的序列。

3.2.2　基于多场景数据进行 LLM 推荐系统微调

工业级推荐系统中经常面临着多任务和多场景推荐的实际问题。以快手的多场景为例, 在推荐页和关注页均会产生用户行为, 也需要推荐模型和算法在这两个场景同时生效。而由于场景数据分布的巨大差异, 实际上很难将这两个场景的数据混合到一起用来训练一个统一的推荐模型, 而分开独立训练又不能很好地将全场景数据进行充分的利用。伴随着 LLM 发展起来的预训练+提示工程的范式, 为解决多场景问题提供了很好的思路[57]。PLATE (prompt learning and tuning enhancement) [58] 算法采用了该种范式来解决多场景推荐问题。该算法设计了两种类型的提示, 分别为 P_{domain} 场景提示和 P_{user} 用户提示。并且, 该工作使用软提示技术使得不同场景的数据通过提示工程的方式, 更好地让基座模型适配到多场景, 同时用户提示的加入保证了足够好的个性化推荐效果。

多场景提示并不是 LLM 在推荐系统上落地的唯一形式, 文献 [32] 对提示技术在推荐场景落地的代表性工作做了全面的总结, 将工作分为 3 类: 第 1 类为 In-Context-Learning, 其为提示模式的直接应用, 即通过提示直接引导基座模型推荐结果; 第 2 类为 Prompt Tuning, 之前提到的 PLATE[56] 算法即属于这类, 在构造提示的基础上, 对基座 LLM 的部分参数和提示参数进行微调, 使得模型更好地适配下游任务的数据; 第 3 类为指令学习 (instruction learning), 与 Prompt Tuning 不同的是, 指令学习是在给定相对固定的提示集合下, 对基座的 LLM 进行微调, 让模型更好地通过提示解决推荐问题。指令学习分为两个阶段: 第 1 个阶段为指令生成阶段, 即提示生成阶段, 一般而言, 相比于自然语言领域, 推荐系统模型的下游任务类型相对固定, 因此在这个阶段可以针对不同的任务类型, 生成数量有限的提示集合; 第 2 个阶段为微调阶段, 结合第 1 阶段所生成的指令集合, 对整个 LLM 进行微调, 使得模型针对指令集合的输出效果得到改善。经过以上两个阶段的处理, 在有限下游类型任务的情况下, 模型能够通过提示给出更优的推荐效果。

3.3　基于 LLM 的交互式推荐系统

传统的推荐系统更多依赖于用户的行为反馈来学习用户的兴趣偏好。这里的反馈包括但不限于用户的点击、点赞、购买、直播观看、评论、搜索等行为。在抖音、快手、微信等主流社区平台上, 积累了大量的用户行为用于内容推荐。然而, 这些内容依然无法直观表达出用户对系统的意图和偏好, 尤其是短期内的意图和偏好。例如, 当用户希望推荐系统推荐出更多美食教学类视频的时候, 无法显式地告知推荐系统, 而是通过点击、点赞、观看等行为反馈给推荐系统。虽然这在技术上是一个可行的方式, 但对于用户来讲可操作性和目的性比较差。LLM 的出现使得用户的意图理解变得更加容易和自然。因此, 基于 LLM 也衍生出一系列方法, 使得用户拥有更加直接的方式向推荐系统反馈自己的偏好。这主要存在以下两类生效路径。

第 1 类是通过交互式推荐, 将 LLM 对人类意图的理解作用于推荐系统。Chat-Rec[59] 工作直接利用 ChatGPT 与用户进行对话式交互, 通过对话理解用户意图, 并提升对推荐结果的可解释性。Google 提出的 RecLLM[60] 也探讨了如何在推荐系统中利用 LLM 与用户进行交互。此工作在 YouTube 场景下使用了 LLM LaMDA 来帮助交互式

推荐系统的构建，该模型的作用包括用户偏好理解、对话管理以及可解释推荐。利用 LaMDA 与用户进行对话并理解用户的对话，从而捕捉用户偏好。在此过程中，该模型还能够处理用户的输入，并在对话过程中进行有效的上下文跟踪，进而完成对话管理。与此同时，还能够利用该模型生成自然语言解释，说明推荐背后的逻辑，达到可解释推荐的目的。整体的交互式推荐流程涉及对话管理、推荐与细化，并且建立用户档案。对话管理模块是核心，负责引导用户通过多轮对话来探索推荐项目，并生成合理的回应。当对话管理模块触发，推荐模块基于用户反馈返回一组高质量、相关性强且多样化的推荐项目展示给用户。RecLLM 采用自然语言档案来代表用户。用户档案是根据用户历史的对话信息以及用户自己手动编辑上传的信息而建立，通过建立用户档案使得系统记住用户偏好的同时，避免用户在搜索时需要反复强调其偏好。

第 2 类是将 LLM 的结果直接作用于推荐系统的融合公式及推荐结果排序打分。典型的应用是将预估任务类比于一个文本分类任务，利用 LLM 进行处理。Prompt4NR[61] 将一个推荐系统的分数估计转化成一个提示生成过程，使得推荐系统任务转换为一个在小空间内的问答式任务。例如，询问推荐结果是否相关、是否推荐得足够精准等。

3.4　基于多模态大模型的推荐系统

随着大规模视觉语言模型（large-scale vision-language model, LVLM）和 LLM 技术的发展，推荐系统的研究也在朝着多模态的方向推进。这些系统通过结合图像、文本等多种模态信息，力求提高推荐的准确性和个性化程度。

一方面，LVLM 面临的挑战是缺乏对特定用户的偏好信息，并且在处理包含噪声、冗余的图像序列时表现不佳。Rec-GPT4V[62] 发现，使用 LVLM 来处理完整视觉信息序列反而比只使用文本信息表现更差。针对此问题，Rec-GPT4V 通过结合用户历史记录生成个性化的推荐，并采用视觉摘要思维方法，即先用 LVLMs 处理单图来进行图片的摘要处理，然后再将摘要重新替换图像并加上系统提示进一步输入给另一 LLM 得到输出，从而显著提高了推荐的精度。

同样地，之后 MLLM-MSR（multimodal large language models for multimodal sequential recommendation）[63] 在序列推荐任务中延续了 Rec-GPT4V 图像摘要思路，通过摘要单图像规避了 LVLM 对于多图处理能力有限的问题。然而该工作发现摘要后的文本信息仍过长，导致后续处理摘要和文本信息的 LLM 输出性能不稳定。为了解决该问题，MLLM-MSR 提出了类似于 RNN 的流程，将每个图像摘要分别输入 LLM，以序列的形式结合历史用户偏好以及当前图像摘要推理出当前用户的偏好，从而解决 LLM 对长输入的处理不稳定问题。此外，该工作还构造了微调数据，对 LLaVA-7b 进行微调，输入是前序用户的偏好，结合用户偏好和推荐候选集合用该微调好的多模态大模型来对候选的匹配度进行预测，实验表明该方法能更好地捕捉和适应用户动态偏好，在序列推荐任务上有更好效果。

此外，在点击率预测任务场景中，MMREC (multi-modal recommender system)[64]) 框架则同样通过利用 LVLM 提取并整合文本与图像摘要信息。但不同的是，最终会用另一较小的语言模型将它们统一编码到潜在空间中，进而通过传统的线性和激活层处理后输出点击率预测。与基线模型相比，结合多模态信息的 MMREC 方法能够显著提高推荐系统的相关性和准确性。

总的来说，多模态大模型推荐系统通过有效整合多种模态的数据，结合用户的历史偏好信息，能够提供更个性化、更精准的推荐，推动了推荐系统技术的发展。

3.5　工业场景下推荐大模型的规模效应

本节主要从推荐模型本身出发，回顾工业场景下对推荐模型自身规模效应进行探索的工作。Meta 团队探索了千亿级参数的生成式推荐系统。该工作将序列推荐的特征工作简化到极致，保留 Token 和时间作为主序列，合并用户历史行为，丢弃过往重度使用的基于用户行为的统计类特征（尤其数值类特征）。此外，该工作提出了 HSTU[65] (hierarchical sequential transduction unit) 结构，与原始 Transformer 相比，采用了名为逐点聚合注意力的注意力计算方法。该方法替换了原有的注意力的 Softmax 算符，解决了序列推荐中 Softmax 无法解决的两个问题：其一，行为历史与推荐目标部分元素的强相关属性会被 Softmax 削弱；其二，在大部分工业场景中，序列推荐因为涉及大量使用动态生成的 Token ID，Softmax 无法适应大规模 Token ID 导致的词表膨胀与稀疏问题。基于以上改进，Meta 团队在工业级场景下验证了推荐模型的缩放定律，模型参数量增长至与 GPT-3 规模 10 倍相近的 1.5 万亿时，推荐效果会有显著提升。而传统的深度学习推荐模型在参数规模达到 2 000 亿后，模型效果会处于饱和状态难以继续提升。

Wukong[66] 也是 Meta 在工业场景下探索推荐大模型结构和规模效应的工作，该工作主要设计了特殊的悟空层结构，从而更好地提升推荐模型

的高阶特征交叉能力。该结构包含因子分解机处理和线性压缩处理, 将两个处理后的结果拼接起来作为输出, 通过多个悟空层叠加形成交互堆栈。该工作也对该结构在工业场景下的缩放定律进行分析, 同样观察到了随着模型参数增长而带来的推荐效果的持续提升。

3.6　推荐大模型的落地部署

工业界推荐系统通常面临每秒万级别以上的请求, 每个请求处理几十到几万级别的推荐物料规模, 叠加 LLM 的复杂度, 为高质量工程部署带来算力上的挑战。为应对该挑战, 目前推荐系统与 LLM 结合方式可分为 3 类, 其中典型工作如下。其一, 把 LLM 的输出作为推荐系统的输入, 利用 LLM 的表征与理解能力进行特征提取并参与训练, 线上推理时模型规模维持原有模型水准。以小红书 NoteLLM[67] 和蚂蚁集团的 LLM-KERec[68]、快手的 LEARN[69] 为例, 这些工作借助 LLM 深度剖析文本、用户行为等数据, 挖掘标签、实体、知识图谱等信息, 从而为精准推荐提供基础。其二, 以 HSTU[65] 为代表, 复现 LLM 的 Scaling Law 使得推荐模型复杂化, 为应对部署困难进行多维度优化: 一方面精简注意力机制以外的神经网络层数, 提升推理效率、削减内存占用, 整合 Layer-Norm 等输出操作避免重复计算, 引入 Row-Wise Adam 算法使内存用量降为 1/6; 另一方面提出 M-FALCON 算法, 并行化处理多候选注意力机制操作、分组利用 Encoder 缓存结构, 保障大规模候选下推理算力规模化, 助力推荐系统筛选最优推荐项。其三, 以 DLLM2Rec[70] 为代表, 该工作将基于 LLM 的推荐模型所蕴含的知识蒸馏到更轻量级的推荐模型, 提高性能的同时减少推理延迟。该工作根据教师模型的置信度和教师−学生一致性来加权, 从而蒸馏出可靠且适合学生模型学习的知识。总之, 以上 3 种方式为推荐系统与 LLM 结合开辟了新的路径, 带来更智能、精准、高效的推荐体验。

此外, LLM 自身在算力效率上的演化, 也使其在工业级推荐上的大规模推理应用更具可能性, 以 Deepseek-V3[18] 为例, 总参数 671 B, 按照 H800 GPUd 的标准, 按 2 美元 GPU 小时的租金计算, 其总训练成本仅为 557.6 万美元。同等参数规模的开源模型 Llama3, 训练则用了 3930 万 H100 GPU 小时, 同时考虑到 H100 GPU 和 H800 GPU 的租金价格差异, DeepSeek-V3 的训练成本仅有 Llama3 模型的 1/20。

4　未来的方向和挑战

在 LLM 驱动下, 推荐系统的研究与实践发展迅速, 尽管取得了显著进步, 仍面临着一系列挑战和机遇。这些挑战不仅涉及技术层面, 如模型训练和推理的资源效率、算法的可解释性与透明度问题和对用户隐私保护的问题, 因此需要在保持模型性能的同时, 寻找解决方案以应对这些挑战。例如, 开发更高效的模型压缩技术和微调技术、探索可解释 AI 的新框架, 以及加强对用户数据处理的伦理规范。此外, 随着技术的进步和应用场景的扩展, 新的研究领域和应用可能会出现, 将推动推荐系统技术向更加智能、可靠和用户友好的方向发展。具体而言, 可以期待以下几个方面的突破。

4.1　更加精细化和个性化推荐

基于 LLM 的推荐系统可以融合多种数据源, 如文本、图像、音频等, 从而更全面地理解用户的兴趣和需求。随着用户对推荐精准度要求的增加和数据规模的快速增长, 推荐系统将进一步精细化和复杂化, 并为传统的推荐带来一系列挑战。首先, 随着 AIGC 的普及, 推荐物料会快速增长, 在新增推荐物料大规模增长的情况下, 如何对推荐物料进行充分地内容理解尤为重要, 而 LLM 的优势则是能够在语义理解、向量表征方面提供较强的能力, 但将内容理解能力和推荐基于行为的表征如何更好地融合, 是需要持续研究的课题。其次, 用户对推荐物料的兴趣点比较多样, 典型的如直播推荐, 用户只对部分直播内容感兴趣并愿意付出时间和精力, 这就使得推荐的数据规模从原有的物料规模扩展到兴趣点规模。要在更大的数据规模下提升推荐精度, 用户行为数据规模、样本数量规模、模型参数规模等几个层面都会带来挑战, 而 LLM 领域的规模效应、模型压缩、模型部署等工作能够为此提供思路和解决方案。事实上, 业界关于规模效应的研究工作还较少, 前文介绍 Meta 的 Wukong[66] 以及 Actions Speak Louder than Words[65] 是目前少有的对规模效应进行探索的工作, 期待能有更多的工作对推荐大模型自身规模效应进行进一步探索。最后, 虽然 LLM 在语言和多模态领域基本实现了基座模型+提示工程的迭代范式, 但在推荐领域, 推荐模型对特定场景的数据依赖依然很强, 如何训练一个通用的基座模型, 并以提示工程的方式提升下游多场景、多任务的高效迭代, 依然面临领域内数据依赖的挑战。

4.2　推荐系统的轻量化和实际落地部署

由于 LLM 本身参数量较大, 在工业场景下进行部署存在巨大挑战, 将模型压缩和高效微调技术应用于基于 LLM 的推荐系统能够使推荐系统更加智能化、轻量化。LLM 剪枝技术通过去除冗

余参数和结构来减小模型的大小，从而降低模型的计算和存储成本，提高模型的推理速度和搜索效率。量化蒸馏技术将浮点数参数转换为定点数参数，从而减少模型存储和计算的需求。模型轻量化微调能够使基于 LLM 的推荐系统的微调过程变得更加高效，以较低的计算成本实现对推荐大模型定制化微调，从而加快推荐系统的迭代速度。事实上，关于推荐系统的轻量化，有如 Lighter and Better[71] 这样比较经典的工作，该工作通过设计低秩分解的注意力模块来使模型设计更加轻量化。随着模型架构自动搜索技术 NAS(neural architecture search)[72]、模型压缩和蒸馏算法以及微调技术的持续改进，推荐系统和搜索引擎可以不断地优化模型结构和参数，从而使模型更加轻量化和高效，能够更准确地捕捉用户的查询意图，并提供更相关、更有用的搜索结果，从而提升用户体验和搜索效率。但是当前将压缩算法等应用于实际推荐大模型的工作仍然较少，期待更多工作在这方面进行尝试，以助力推荐系统大模型的工业界实际落地部署。

4.3　更智能的交互方式

传统推荐系统中，用户只能相对被动地参与到推荐过程中。因此，当遇到信息茧房等问题时，用户较难与推荐模型交互从而主动改善体验。而交互式推荐系统（conversational recommendation system, CRS）带来了一些改变，使推荐系统能够不单一依赖用户行为，采用对话交流向用户推荐，采纳用户反馈来进一步优化推荐内容。现有的结合 LLM 的交互式推荐系统，如先前介绍的 ChatRec[59] 以及 RecLLM[60]，均尝试了通过 LLM 参与用户反馈过程来优化推荐内容。但是这些工作的局限在于用户反馈的方式仍为比较单一的文本输入对话式反馈。这种形式对于用户来说较为麻烦，从而导致用户参与反馈的可能性大大减少。

足够智能化的推荐系统大模型可以与用户进行更深入的对话，以更好地理解用户需求和偏好。这种个性化的对话式推荐系统可以结合 LLM 的能力来理解语境、情感和用户意图，从而提供更准确、个性化的推荐。例如，系统可以不仅通过对话了解用户的特定需求、情感状态或偏好变化，从而实时调整推荐结果，还可以通过反馈按钮[73] 结合 LLM 动态生成交互内容，用户仅需点击进行反馈，从而使反馈过程更加高效。除此之外，LLM 和语音识别等技术的融合将使得未来的推荐和搜索系统更加智能化成为可能。基于 LLM 的推荐系统不仅可以接受文字输入，还可以整合语音、图像等多种模态的输入，使用户可以通过更多样化的方式与系统进行交互。这种多模态交互将进一步提高用户体验，使得用户可以根据实际场景和偏好选择最适合的交互方式，从而更方便地获取个性化推荐。

4.4　数据安全和隐私保护的增强

未来，推荐系统应更加注重用户数据的安全和隐私保护，采取更严格的数据保护措施，确保用户的个人信息不被滥用或泄露，从而增强用户对系统的信任感和使用舒适度。基于 LLM 的推荐系统需要采用相关技术保护用户的个人数据，例如差分隐私技术[74]，通过在数据处理过程中引入噪声或扰动，使得输出结果不会泄露个别用户的信息，从而保护用户的隐私。这种技术可以在不影响模型效果的前提下，有效地防止敏感信息的泄露。在 LLM 的训练过程中可以进一步加强联邦学习技术[73] 的研究，即在不同设备或服务器上进行模型训练，而不需要将原始数据集集中存储于同一地点，避免用户个人数据集中存储和传输过程中的风险，同时保护用户隐私。

此外，LLM 的可解释性问题[74] 一直是学术界的一个热门话题，基于 LLM 的推荐系统可以通过提高模型的透明度和可解释性，让用户了解系统是如何利用他们的数据进行推荐。通过向用户展示模型的运行机制、数据处理过程和推荐依据，增强用户对系统的信任感，使用户更加愿意与系统进行交互。

4.5　LLM 与推荐结合遇到的幻象与偏见问题

LLM 和推荐系统自身存在的问题，在二者结合后会被放大。首先，LLM 在知识理解和生成方面存在一定概率的幻象问题，包括事实性错误、逻辑不一致、无中生有等，这都会通过引入错误的知识表征，使推荐系统得到错误的输入。这类问题的解决依赖于数据清洗、多源数据整合、生成策略调整等优化措施。其次，推荐模型的学习依赖于用户的后验行为，这种模型学习方式会引入较多维度的偏见现象，包括人口统计学偏见、兴趣标签偏见、热门内容偏见等。而 LLM 和推荐模型的联合学习，由于加大了模型复杂度，可能会加剧以上偏见现象，因此在基于 LLM 的推荐系统中，应着重强化人工干预机制的动态调控、算法偏见消除技术的迭代优化以及数据多样性增强策略的系统性实施。

5　结束语

本综述回顾了推荐系统和 LLM 的发展历程，以及推荐系统和 LLM 的研究现状，并且从模型在推荐系统中的作用出发，对推荐系统中运用 LLM 的方式进行梳理，从中总结出了基于 LLM 的推荐

系统发展方向和未来可能面临的挑战。这些挑战不仅包含技术层面的突破, 也包含数据隐私、算法公平性等道德伦理方面的问题。当然, 我们对推荐系统与 LLM 的结合前景充满信心, 相信在实践探索中, 基于 LLM 的推荐系统能够为用户提供更加个性化、更加高效的推荐服务。

参考文献:

[1] 项亮. 推荐系统实践[M]. 北京: 人民邮电出版社, 2012.

[2] COVINGTON P, ADAMS J, SARGIN E. Deep neural networks for YouTube recommendations[C]//Proceedings of the 10th ACM Conference on Recommender Systems. Boston: ACM, 2016: 191−198.

[3] AIVAZOGLOU M, ROUSSOS A O, MARGARIS D, et al. A fine-grained social network recommender system[J]. Social network analysis and mining, 2019, 10(1): 8.

[4] QUESTMOBILE. QuestMobile2023 中国移动互联网年度报告[EB/OL]. (2024−01−30)[2025−02−24]. https://www.questmobile.com.cn/research/report.

[5] GU Siyu, SHENG Xiangrong. On ranking consistency of pre-ranking stage[EB/OL]. (2022−05−03)[2025−02−24]. https://arxiv.org/abs/2205.01289.

[6] ZHAO W X, ZHOU Kun, LI Junyi, et al. A survey oflarge language models[EB/OL]. (2023−05−31)[2025−02−24]. https://arxiv.org/abs/2303.18223.

[7] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735−1780.

[8] ELMAN J. Finding structure in time[J]. Cognitive science, 1990, 14(2): 179−211.

[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems30 (NIPS2017). Long Beach: Curran Associates Inc, 2017: 5998−6008.

[10] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of the 2018 Conference of the North American Chapter Of the Association for Computational Linguistics: Human LanguageTechnologies. New Orleans: USAACL, 2018: 2227−2237.

[11] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minnesota: Association for Computational Linguistics, 2019: 4171−4186.

[12] OUYANG L, WU J, JIANG Xu, et al. Training language models to follow instructions with human feedback[C]//Advances in Neural Information Processing Systems 35 (NeurIPS 2022). Louisiana: Curran Associates Inc, 2022: 27730−27744.

[13] LEWIS M, LIU Yinhan, GOYAL N, et al. Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[EB/OL]. (2019−10−29)[2025−02−24]. https://arxiv.org/abs/1910.13461.

[14] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with aunified text-to-text Transformer[EB/OL]. (2019−10−23)[2025−02−24]. https://artxiv.org/abs/1910.10683.

[15] ZHANG Zhengyan, HAN Xu, LIU Zhiyuan, et al. ERNIE: enhanced language representation with informative entities[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Stroudsburg: USAACL, 2019: 1441−1451.

[16] LIU Yinhan, OTT M, GOYAL N, et al. RoBERTa: A robustly optimized BERT pretraining approach[EB/OL]. (2019−07−26)[2024−05−21]. https://arxiv.org/abs/1907.11692.

[17] LAN Z, CHEN M, GOODMAN S, et al. Albe-rt: a lite BERT for self-supervised learning of language representations[C]//8th International Conference on Learning Representations. [S.l.]: Open Access, 2020.

[18] LIU Aixin, FENG Bei, XUE Bing, et al. Deepseek-v3 technical report[EB/OL]. (2024−12−27)[2025−01−01]. https://arxiv.org/abs/2412.19437.

[19] GUO D, YANG Dejian, ZHANG Haowei, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning[EB/OL]. (2025−01−22)[2025−02−02]. https://arxiv.org/abs/2501.12948.

[20] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report[EB/OL]. (2024−05−04)[2025−02−24]. https://arxiv.org/pdf/2303.08774.

[21] ZHANG Susan, ROLLER S, GOYAL N, et al. OPT: open pre-trained Transformer language models[EB/OL]. (2022−05−02)[2025−02−24]. https://arxiv.org/abs/2205.01068.

[22] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: open and efficient foundation language models[EB/OL]. (2023−02−27)[2025−02−24]. https://arxiv.org/abs/2302.13971.

[23] TOUVRON H, MARTIN L, STONE K, et al. LLaMA 2: open foundation and fine-tuned chat models[EB/OL]. (2023−07−18)[2025−02−24]. https://arxiv.org/abs/2307.09288.

[24] DUBEY A, JAUHRIi A, PANDEY A, et al. The LLaMA 3 herd of models[EB/OL]. (2024−07−31)[2025−02−24]. https://arxiv.org/abs/2407.21783.

[25] BAI Yuntao, JONES A, NDOUSSE K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback[EB/OL]. (2022−04−12)[2025−02−24]. https://arxiv.org/abs/2204.05862.

[26] LIU H, LI C, WU Q, et al. Visual instruction tuning[C]//Advances in Neural Information Processing Systems 36 (NeurIPS2023). Louisiana: Curran Associates Inc, 2023: 34892−34916.

[27] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//The 38thInternationalConference on Machine Learning. [S.l.]: Curran Associates Inc, 2021: 8748−8763.

[28] DEY R, SALEM F M. Gate-variants of gated recurrent unit (GRU) neural networks[C]//2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS). Boston: IEEE, 2017: 1597−1600.

[29] LIU Pengfei, YUAN Weizhe, FU Jinlan, et al. Pre-train,

prompt, and predict: a systematic survey of prompting methods in natural language processing[J]. ACM computing surveys, 2023, 55(9): 1−35.

[30] WANG Qi, LI Jindong, WANG Shiqi, et al. Towards next-generation LLM-based recommender systems: a survey and beyond[EB/OL]. (2024−10−10)[2025−02−24]. https://arxiv.org/abs/2410.19744.

[31] LIN Jianghao, DAI Xinyi, XI Yunjia, et al. How can recommender systems benefit from large language models: a survey[EB/OL]. (2023−07−09)[2025−02−24]. https://arxiv.org/abs/2306.05817.

[32] ZHAO Zihuai, FAN Wenqi, LI Jiatong, et al. Recommender systems in the era of large language models (LLMS)[EB/OL]. (2023−07−05)[2025−02−24]. https://arxiv.org/abs/2307.02046.

[33] GUO Huifeng, TANG Ruiming, YE Yunming, et al. DeepFM: a factorization-machine based neural network for CTR prediction[EB/OL]. (2017−05−13)[2025−02−24]. https://arxiv.org/abs/1703.04247.

[34] LIAN Jianxun, ZHOU Xiaohuan, ZHANG Fuzheng, et al. xDeepFM: combining explicit and implicit feature interactions for recommender systems[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: ACM, 2018: 1754−1763.

[35] WANG R, FU B, FU G, et al. Deep and crossnetwork for adclick predictions[C]//Proceedings of the ADKDD'17. Halifax: Associationfor Computing Machinery, 2017: 1−7.

[36] WANG Ruoxi, SHIVANNA R, CHENG D, et al. DCN V2: improved deep & cross network and practical lessons for web-scale learning to rank systems[C]//Proceedings of the Web Conference 2021. Ljubljana: ACM, 2021: 1785−1797.

[37] MA Jiaqi, ZHAO Zhe, YI Xinyang, et al. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: ACM, 2018: 1930−1939.

[38] TANG Hongyan, LIU Junning, ZHAO Ming, et al. Progressive layered extraction (PLE): a novel multi-task learning (MTL) model for personalized recommendations[C]//Fourteenth ACM Conference on Recommender Systems. Brazil: ACM, 2020: 269−278.

[39] CHANG Jianxin, ZHANG Chenbin, HUI Yiqun, et al. PEPNet: parameter and embedding personalized network for infusing with personalized prior information[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Long Beach: ACM, 2023: 3795−3804.

[40] SWIETOJANSKI P, LI Jinyu, RENALS S. Learning hidden unit contributions for unsupervised acoustic model adaptation[J]. IEEE/ACM transactions on audio, speech, and language processing, 2016, 24(8): 1450−1463.

[41] ZHOU Guorui, ZHU Xiaoqiang, SONG Chenru, et al. Deep interest network for click-through rate prediction [C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: ACM, 2018: 1059−1068.

[42] ZHOU Guorui, MOU Na, FAN Ying, et al. Deep interest evolution network for click-through rate prediction[J]. Proceedings of the AAAI conference on artificial intelligence, 2019, 33(1): 5941−5948.

[43] PI Qi, ZHOU Guorui, ZHANG Yujing, et al. Search-based user interest modeling with lifelong sequential behavior data for click-through rate prediction[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. [S.l.]: ACM, 2020: 2685−2692.

[44] CHANG Jianxin, ZHANG Chenbin, FU Zhiyi, et al. TWIN: two-stage interest network for lifelong user behavior modeling in CTR prediction at Kuaishou[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Long Beach: ACM, 2023: 3785−3794.

[45] ZHU Han, LI Xiang, ZHANG Pengye, et al. Learning tree-based deep model for recommender systems[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: ACM, 2018: 1079−1088.

[46] GAO Weihao, FAN Xiangjun, WANG Chong, et al. Deep retrieval: Learning a retrievable structure for large-scale recommendations[EB/OL]. (2020−07−12)[2025−02−24]. https://arxiv.org/abs/2007.07203.

[47] VAN DEN OORD A, VINYALS O, KAVUKCUOGLU K, et al. Neural discrete representation learning[C]//Advances in Neural Information Processing Systems 30. California: Curran Associates Inc, 2017.

[48] RAJPUT S, MEHTA N, SINGH A, et al. Recommender Systems with Generative Retrieval[C]//Advances in Neural Information Processing Systems 36. Louisiana: Curran Associates Inc, 2023: 10299−10315.

[49] QIU Zhaopeng, WU Xian, GAO Jingyue, et al. U-BERT: pre-training user representations for improved recommendation[J]. Proceedings of the AAAI conference on artificial intelligence, 2021, 35(5): 4320−4327.

[50] DING Hao, MA Yifei, DEORAS A, et al. Zero-shot recommender systems[EB/OL]. (2021−05−18)[2025−02−24]. https://arxiv.org/abs/2105.08318.

[51] HOU Yupeng, MU Shanlei, ZHAO W X, et al. Towards universal sequence representation learning for recommender systems[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Washington DC: ACM, 2022: 585−593.

[52] 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.

[53] SHAZEER N, MIRHOSEINI A, MAZIARZ K, et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer[C]//International Conference on Learning Represent-ations. Toulon: openreview. net, 2017.

[54] GE Tiezheng, HE Kaiming, KE Qifa, et al. Optimized product quantization[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 36(4): 744−755.

[55] SUN Fei, LIU Jun, WU Jian, et al. BERT4Rec: sequen-

tial 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.

[56] WU Chuhan, WU Fangzhao, QI Tao, et al. PTUM: pre-traininguser model fromunlabeled user behaviors via self-supervision[EB/OL]. (2020−10−04)[2025−02−24]. https://arxiv.org/abs/2010.01494.

[57] LIN Junyang, MEN Rui, YANG An, et al. M6: a Chinese multimodal pre-trainer[EB/OL]. (2021−05−01)[2025−02−24]. https://arxiv.org/abs/2103.00823.

[58] WANG Yuhao, ZHAO Xiangyu, CHEN Bo, et al. PLATE: a prompt-enhanced paradigm for multi-scenario recommendations[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. Taipei: ACM, 2023: 1498−1507.

[59] GAO Yunfan, SHENG Tao, XIANG Youlin, et al. Chat-rec: Towards interactive and explainable LLMS-augmented recommender system[EB/OL]. (2023−05−25)[2025−02−24]. https://arxiv.org/abs/2303.14524.

[60] FRIEDMAN L, AHUJA S, ALLEN D, et al. Leveraging large language models in conversational recommender systems[EB/OL]. (2023−05−13)[2025−02−24]. https://arxiv.org/abs/2305.07961.

[61] 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.

[62] LIU Yuqing, WANG Yu, SUN Lichao, et al. Rec-GPT4V: multimodal recommendation with large vision-language models[EB/OL]. (2024−02−13)[2025−02−24]. https://arxiv.org/abs/2402.08670.

[63] YE Yuyang, ZHENG Zhi, SHEN Yishan, et al. Harnessing multimodal large language models for multimodal sequential recommendation[EB/OL]. (2024−08−19)[2025−05−20]. https://arxiv.org/abs/2408.09698.

[64] TIAN Jiahao, ZHAO Jinman, WANG Zhenkai, et al. MMREC: LLM based multi-modal recommender system[EB/OL]. (2024−08−08)[2025−02−24]. https://arxiv.org/abs/2408.04211.

[65] ZHAI Jiaqi, LIAO L, LIU Xing, et al. Actions speak louder than words: trillion-parameter sequential transducers for generative recommendations[EB/OL]. (2024−02−27)[2025−02−24]. https://arxiv.org/abs/2402.17152.

[66] ZHANG Buyun, LUO Liang, CHEN Yuxin, et al. Wukong: towards a scaling law for large-scale recommendation[EB/OL]. (2024−05−04)[2025−02−24]. https://arxiv.org/abs/2403.02545.

[67] ZHANG Chao, WU Shiwei, ZHANG Haoxin, et al. NoteLLM: a retrievable large language model for note recommendation[C]//Companion Proceedings of the ACM Web Conference 2024. Singapore: ACM, 2024: 170−179.

[68] 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.

[69] JIA Jian, WANG Yipei, LI Yan, et al. LEARN: knowledge adaptation from large language model to recommendation for practical industrial application[EB/OL]. (2024−12−26)[2025−02−24]. https://arxiv.org/abs/2405.03988.

[70] CUI Yu, LIU Feng, WANG Pengbo, et al. Distillation matters: empowering sequential recommenders to match the performance of large language models[C]//18th ACM Conference on Recommender Systems. Bari: ACM, 2024: 507−517.

[71] FAN Xinyan, LIU Zheng, LIAN Jianxun, et al. Lighter and better: low-rank decomposed self-attention networks for next-item recommendation[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. [S.l.]: ACM, 2021: 1733−1737.

[72] WHITE C, SAFARI M, SUKTHANKER R, et al. Neural architecture search: Insights from 1000 papers[EB/OL]. (2023−01−20)[2025−02−24]. https://arxiv.org/abs/2301.08727.

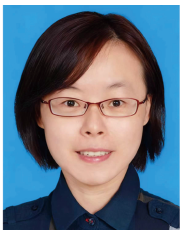
[73] LI Qinbin, WEN Zeyi, WU Zhaomin, et al. A survey on federated learning systems: vision, hype and reality for data privacy and protection[J]. IEEE transactions on knowledge and data engineering, 2023, 35(4): 3347−3366.

[74] ZHAO Haiyan, CHEN Hanjie, YANG Fan, et al. Explainability for large language models: a survey[J]. ACM transactions on intelligent systems and technology, 2024, 15(2): 1−38.

作者简介：

谢广明，教授，博士生导师，中国自动化学会机器人竞赛工作委员会副主任，国际水中机器人联盟创始人，中国仿真学会机器人系统仿真专委会主任委员，主要研究方向为复杂系统动力学与控制、智能仿生机器人多机器人系统与控制。现主持国家自然科学基金重点项目等 8 项，获发明专利授权 20 余项。曾荣获国家自然科学奖二等奖、教育部自然科学奖一等奖、吴文俊人工智能科学技术奖创新奖二等奖，发表学术论文 200 余篇。E-mail：xiegming@pku.edu.cn。

白彦冰，博士研究生，主要研究方向为推荐系统、大模型、智能系统与控制、计算机辅助设计。先后担任新浪、快手、字节跳动等知名公司核心推荐算法团队负责人。E-mail：baiyb@stu.pku.edu.cn。



张艳玲，副教授，中国仿真学会机器人系统仿真专委会委员，主要研究方向为群体智能、演化博弈、博弈学习和推荐系统。主持国家自然科学基金青年基金项目，是国家自然科学基金重点项目的校内负责人。发表学术论文 30 余篇。E-mail：yanlzhang@ustb.edu.cn。