计算机科学
COMPUTER SCIENCE

面向大语言模型的推荐系统综述

卡祖铭, 赵鹏, 张波, 傅晓宁

引用本文

卡祖铭, 赵鹏, 张波, 傅晓宁. 面向大语言模型的推荐系统综述[J]. 计算机科学, 2024, 51(11A): 240800111-11.

KA Zuming, ZHAO Peng, ZHANG Bo, FU Xiaoning. Survey of Recommender Systems for Large Language Models [J]. Computer Science, 2024, 51(11A): 240800111-11.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240800111

面向大语言模型的推荐系统综述

卡祖铭　赵　鹏　张　波　傅晓宁
火箭军工程大学作战保障学院　西安 710025
(553097606@qq.com)

摘　要　大型语言模型已成为自然语言处理领域中的有力使用工具,并且在近期也成功吸引了推荐系统领域的广泛关注。这些模型凭借自监督学习在庞大的数据集上进行了深度训练,从而在通用表征学习上取得了显著成果。通过微调、提示调整等高效的迁移技术,它们有望全面提升推荐系统的各项性能。在利用语言模型的力量来优化推荐质量的过程中,关键在于充分利用其高质量的文本特征表征以及广泛的外部知识库,以此为基础构建物品与用户之间的紧密联系。为了全面而深入地理解当前基于大型语言模型的推荐系统,文中将这些模型细致地划分为两大类:用于推荐的判别式大型语言模型和用于推荐的生成式大型语言模型。同时,对于后者,还进一步将其细分为约束生成和自由生成,并对这两种方法的相关研究进行了详尽的总结。此外,还指出了该领域面临的关键挑战,并分享了一些有价值的发现,希望可以为研究人员和从业者提供宝贵的灵感与启示。

关键词:大语言模型;推荐系统;自然语言处理
中图分类号　TP391

Survey of Recommender Systems for Large Language Models

KA Zuming,ZHAO Peng,ZHANG Bo and FU Xiaoning
College of Operational Support,Rocket Force University of Engineering,Xi'an 710025,China

Abstract　Large language models have emerged as highly effective tools in the field of natural language processing(NLP) and have recently garnered considerable attention in the domain of recommendation systems(RS). These models have undergone extensive training on vast datasets through self-supervised learning,achieving remarkable results in learning universal representations. With efficient transfer techniques such as fine-tuning and prompt tuning,they have the potential to significantly enhance various aspects of recommendation system performance. The crux of leveraging language models to optimize recommendation quality lies in fully utilizing their high-quality text feature representations and extensive external knowledge bases to establish strong connections between items and users. To gain a comprehensive and in-depth understanding of current recommendation systems based on large language models,this paper meticulously categorizes these models into two main types:discriminative large language models for recommendation(DLLM4Rec) and generative large language models for recommendation(GLLM4Rec). Furthermore,the latter is subdivided into constrained generation and free generation,and a detailed summary of relevant research on these two approaches is provide. Additionally,this paper identifies key challenges in this field and shares valuable findings,hoping to provide researchers and practitioners with precious inspiration and insights.

Keywords　Large language model,Recommender system,Natural language processing

1　引言

大型语言模型(Large Language Models,LLM)作为自然语言处理(Natural Language Processing,NLP)领域的代表,近年在推荐系统(Reco-mmendation Systems,RS)领域中赢得了空前的瞩目。这些模型凭借其卓越的自我监督学习能力,在海量数据上进行了深入的训练,不仅在通用表示学习方面取得了显著的成果,更展现出了通过一系列高效的迁移技术(如微调和实时调整等)来强化推荐系统各环节的巨大潜力。在提升推荐质量的关键环节中,LLM 的力量尤为显著。它们凭借对文本文的高精度表示以及对外部知识的广泛覆盖,为构建项目和用户之间复杂而精准的相关性提供了强大的支持。这种强大的关联性建立能力,正是我们在推动推荐系统进步和创新中迫切需要的。因此,深入研究并应用 LLM 在推荐系统中的潜力,不仅具有重要的理论价值,更为实践中的优化与创新指明了方向。为了全面洞悉现有的基于 LLM 的推荐系统,本文深入剖析了这些模型,并将其归类为两大主要范式:判别式 LLM 推荐(Discriminative LLM for Recommendation,DLLM4Rec)和生成式 LLM 推荐(Generative LLM for Recommendation,GLL-M4Rec)。特别地,对生成式 LLM 推荐 GL-LM4Rec 的方法进行了详尽的概述。本文系统地回顾并分析了每个范式中现有的基于大语言模型的推荐系统,深入探讨了这些系统的方法论、技术细节以及性能表现。通过对这些系统的细致研究,本文为研究人员和从业者提供了关于这些推荐系统方法、技术和性能的深刻见解。此外,本文识别并探讨了这些基于 LLM 的推荐系统所面临的关键挑战。这些挑战不仅揭示了当前技术的局限性,也指明了未来研究和发展的方向。同时,还分享了一些有价值的发现,这些发现

通信作者:赵鹏(zpxhh@163.com)

不仅为研究人员提供了新的视角，也为从业者提供了实践中的灵感和指导。通过本研究的全面分析和深入讨论，期望能够为基于 LLM 的推荐系统的研究和发展提供有价值的参考，推动该领域的持续进步和创新。

2 介绍

推荐系统（REcommender Sytems，RecSys）在解决信息过载问题、丰富用户在线体验方面发挥着不可或缺的作用。当用户面对海量信息时，筛选出符合自己兴趣的内容变得尤为困难，这正是推荐系统大显身手的时候。它们通过提供个性化的推荐建议，帮助用户在各种应用场景（如娱乐[1]、电子商务[2]、职位匹配[3]等）中快速找到心仪的项目。以电影推荐为例，像 IMDB 和 Netflix 这样的平台，通过推荐系统能够基于电影的内容特性和用户的观看历史，为用户推荐最新且符合其兴趣的电影，使用户能够轻松发现新的观影选择。推荐系统的核心在于利用用户与项目之间的交互数据以及相关的辅助信息，如用户的偏好、项目的特征以及用户与项目间的交互历史，来预测用户对未知项目的喜好程度，并据此提供个性化的推荐服务[4]。为了实现精准的推荐，推荐系统采用了多种技术和方法。其中，协同过滤是最常用的技术之一，它基于用户或项目的相似性来生成推荐。此外，内容过滤通过分析项目的内容（如电影的类型、演员阵容、导演等）和用户的偏好历史，来推荐与用户兴趣相契合的项目。近年来，随着深度学习技术的蓬勃发展，越来越多的推荐系统开始运用深度学习模型来捕捉用户和项目之间复杂的非线性关系，从而生成更为精准和个性化的推荐结果。特别是文本信息，如项目描述、用户档案和用户评价，对于预测用户与项目之间的匹配度（即用户对该项目的喜好概率）至关重要[5]。具体来说，用户与项目之间的协同行为被广泛应用于设计各种推荐模型，这些模型不仅能够预测匹配度，还能进一步学习用户和项目的深层表示[6-7]。此外，用户和项目的文本辅助信息蕴含着丰富的知识，这些知识在计算匹配度方面起着关键作用，并为深入理解用户偏好、推动推荐系统的持续优化提供了宝贵的见解[8]。

深度学习凭借其卓越的表示学习能力，在多个领域取得了显著成果，因此深度神经网络（Deep Neural Networks，DNNs）被广泛应用于推动推荐系统的发展[9-10]。DNNs 通过多样化的架构，展示了在模拟用户与项目交互过程中的独特优势。例如，循环神经网络（Recurrent Neural Net-works，RNNs）作为处理序列数据的强大工具，被有效应用于捕捉用户交互序列中的高阶依赖关系[11-12]。此外，鉴于用户的在线行为（如点击、购买、社交互动）具有图结构数据的特性，图神经网络（Graph Neural Networks，GNNs）已崭露头角，成为先进的表示学习技术，用于精确学习用户和项目的表征[1,6,13]。与此同时，DNNs 在编码辅助信息方面也展现了其显著优势。例如，已有研究提出了一种基于 BERT 的方法，该方法能够有效提取并利用用户的文本评论，从而进一步提升了推荐系统的性能[14]。

尽管推荐系统在过去取得了显著的成功，但当前大多数先进的推荐系统仍面临一系列固有的限制。1）由于模型规模和数据集大小的局限，先前基于 DNN 的模型（如 CNN 和 LSTM）以及为推荐系统量身定制的预训练语言模型（如 BERT）在捕获关于用户和项目的深层文本知识方面显得力不从心，这暴露了它们在自然语言理解方面的不足，进而影响了在各种推荐场景下的预测性能。2）现有的推荐系统 RecSys 方法普遍针对特定任务进行定制，这导致它们在面对新的推荐任务时缺乏足够的泛化能力。例如，一个经过精心训练的推荐算法可能在用户和项目的评分矩阵上表现出色，能够准确预测电影的评分，但若要它执行带有解释性的前 k 部电影推荐任务，则面临巨大挑战。这是因为这些推荐架构的设计通常高度依赖于特定推荐场景（如前 k 名推荐、评分预测和可解释推荐）的任务特定数据和领域知识。3）尽管大多数基于 DNN 的推荐方法在涉及简单决策的推荐任务（如评分预测和前 k 名推荐）中表现不俗，但它们在处理需要多步骤推理的复杂决策时效果却不理想。以旅行规划推荐为例，多步骤推理在其中扮演着关键角色。推荐系统首先需要基于目的地考虑热门的旅游景点，接着为这些景点安安排合适的行程，最后根据用户的特定偏好（如旅行成本和时间）推荐行程计划。这要求推荐系统具备强大的推理和决策能力，而现有的方法在这方面还有待提升。

近年来，随着自然语言处理技术的飞速发展，拥有数十亿参数的 LLMs 在自然语言处理[15]、计算机视觉[16]以及分子发现[17]等多个前沿领域引发了显著影响。从技术角度来看，这些 LLMs 主要基于 Transformer 架构，在庞大的、来自多元化来源的文本数据上进行了预训练，这些数据涵盖了文章、书籍、网站等公开可取的书面材料。随着训练语料库的持续扩大，LLMs 的参数规模也随之增长，而最新的研究已经证实，这种增长带来了显著的性能提升[18-19]。特别值得注意的是，LLMs 在核心的语言理解和生成任务中展现出了前所未有的强大能力。这些进步使得 LLMs 能够更深入地理解人类意图，并生成更为自然、贴近人类表达的语言回应。

此外，LLMs 还表现出了令人瞩目的泛化和推理能力，这些能力使它们能够轻松地适应各种未曾见过的新任务和领域。具体而言，LLMs 无需针对每个特定任务进行繁琐的微调，只需通过提供简单的指令或几个任务示例，便能将其所学的知识和推理技能灵活地应用于新任务中。更为先进的技术，如上下文学习，更是进一步强化了 LLMs 的泛化性能，使其无需针对特定的下游任务进行微调[19]。

最近，LLMs 作为下一代推荐系统的潜在技术，已经开始受到人们的初步探索。其中，Chat-Rec[3]的提出就是一个典型的例子，它通过 ChatGPT 与用户进行自然的对话交互，进而优化传统电影推荐系统生成的候选集，以此提高推荐的准确性和可解释性。同样，Zhang 等[20]利用 T5 这一 LLM 作为推荐系统的核心，使用户能够用自然语言直接表达他们的明确偏好和意图，这一创新方法相较于仅依赖用户-项目交互的传统方法，展现出了更为卓越的推荐性能。由图 1 可以看到，LLMs 在不同电影推荐任务中的应用实例，这些任务涵盖了前 K 名推荐、评分预测、会话式推荐以及解释生成等多个方面。这些示例不仅展示了 LLMs 在推荐系统中的广泛应用，也体现了其强大的处理能力和灵活性。由于 LLMs 技术的飞速发展，全面回顾并深入分析基于 LLMs 的推荐系统的最新进展与挑战，对于推动该领域的进一步研究和应用具有至关重要的意义。

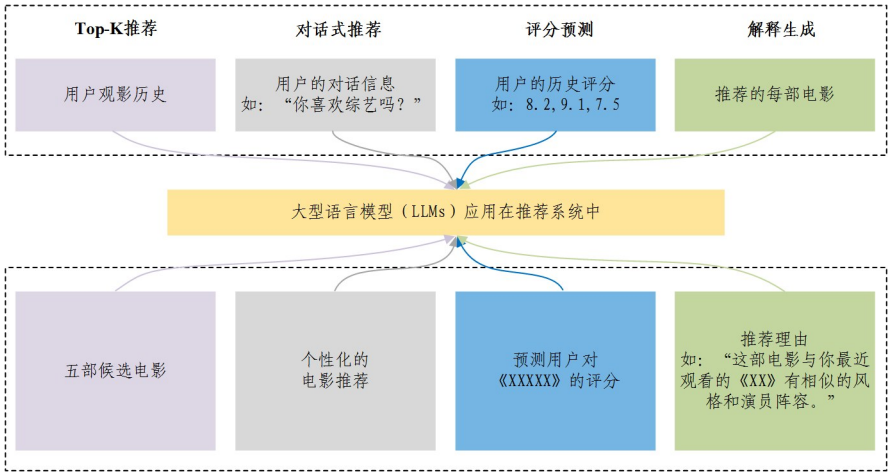

图 1　在电影推荐场景中，LLM 在各种推荐任务中的应用示例

Fig. 1　Application examples of LLM in various recommendation tasks in movie recommendation scenarios

3　相关工作

本章将对推荐系统和大语言模型领域的相关工作进行简要的回顾与梳理。如图 2 所示，该时间表展示了推荐系统和语言模型领域内的里程碑事件，提供了一个跨学科领域发展的清晰脉络。

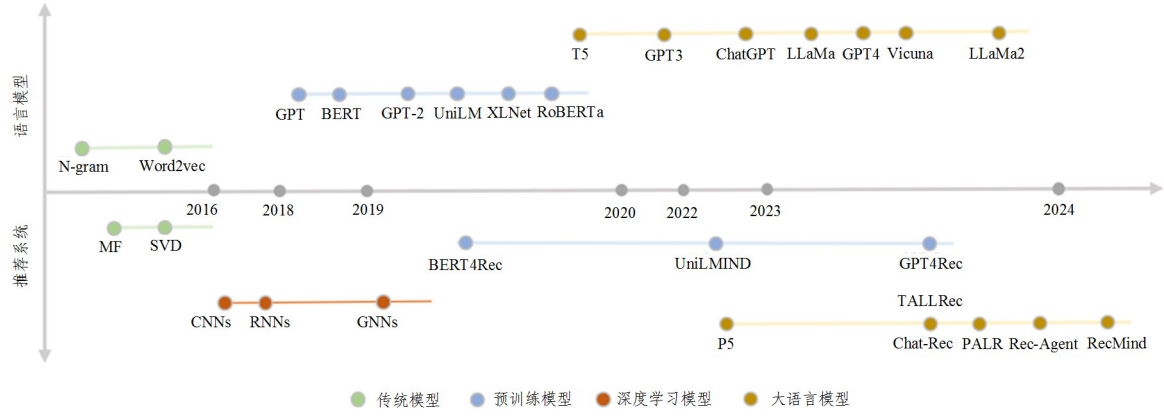

图 2　推荐系统和语言模型邻域的里程碑时间表

Fig. 2　Timeline of milestone in the domains of recommendation systems and language models

3.1　大语言模型

大型语言模型通过在大量的文本数据上进行训练，并具备数十亿个参数，成功捕捉了自然语言的模式和结构。在众多预训练语言模型中，BERT[21]（Bidirectional Encoder Representation from Transformers）、GPT[22]（Chat Generative Pre-trained Tranformer）和 T5[23]（Text-to-Text Transfer Transformer）堪称经典。这些模型常可分为三大类[24]：仅编码器模型、仅解码器模型和编码解码器模型。BERT 作为仅有编码器的模型，采用双向注意力机制处理标记序列，充分考量每个 token 的上下文。通过掩蔽语言建模和下一句预测等预训练任务，BERT 在大量的文本数据中捕捉语言和含义的细微差别，将文本转化为向量空间，实现了高度上下文敏感的分析。GPT 则采用自注意机制进行从左到右的单向词序列处理。它主要用于语言生成任务，通过嵌入向量映射回文本空间，生成与上下文紧密相关的响应。T5 作为编码解码器模型，通过将所有自然语言处理问题转化为文本生成问题，具备处理任何文本到文本任务的能力。随着模型规模的不断扩大，LLMs 展现了前所未有的理解和生成能力[18,25]，彻底改变了自然语言处理领域。这些模型，如 GPT-3[15]、LaMDA[26]、PaLM[27] 和 Vicuna[28] 等，通过大量文本数据训练，能够捕捉复杂的语言模式和细微差别。最近，LLMs 展示了令人瞩目的上下文知识学习（In-Context Learning，ICL）能力，这是它们设计的核心功能。ICL 指模型根据输入上下文理解并提供答案的能力，而非仅仅依赖预训练获得的内部知识。研究如 SG-ICL[29] 和 EPR[30] 等探索了如何在各种任务中利用 ICL，这些研究揭示了 ICL 使 LLMs 能够根据输入上下文调整回答，而非生成通用的响应。此外，思维链（chain-of-thought，CoT）[31] 技术进一步增强了 LLMs 的推理能力。通过在提示中提供多个示例来描述思维链，CoT 引导模型的推理过程。CoT 的扩展概念包括自我一致性，通过答案上的多数表决机制操作，为模型提供了更加稳健的推理机制[32]。当前研究如 StaR[33]，THOR[34] 和 Tab-CoT[35] 等继续深入探索在 LLMs 中应用 CoT 的方法，以提高模型推理的准确性和效率。在化学[17]、教育[36] 和金融[37] 等众多领域，LLMs 展现出了巨大的应用潜力。

3.2　大语言模型推荐系统

在推荐系统中，LLMs 尤其发挥着重要作用。通过分析历史用户交互和偏好，LLMs 能够预测用户对物品的评分，从而提高推荐的准确性。此外，LLMs 还被应用于顺序推荐，通过分析用户交互的序列来预测其下一个偏好。例如，TALLRec[38]，M6-Rec[39]，PALR[40] 和 P5[41] 等模型在这一领域取得

了显著成果。特别值得一提的是，LLMs 如 ChatGPT 在提升 RecSys 的功能和用户体验方面发挥了关键作用。它们不仅用于预测评分和顺序推荐，还用于生成可解释的推荐。例如，Chat-Rec[1] 利用 ChatGPT 提供清晰易懂的推理，增强了用户信任和参与度。此外，LLMs 的互动和对话能力也被用于创建更动态的推荐体验，如 UniCRS[41] 和 UniMIND[42] 等框架，通过预训练语言模型的知识增强提示学习，满足了会话和推荐子任务的需求。

3.3 根据应用场景的分类

3.3.1 自动生成文本

文本生成技术通过对自然语言的理解，将文本生成任务转化为用户的兴趣表达。用户产生的文本可以是文字、语音、图像等形式。文本生成技术与基于规则的推荐方法相比，在推荐系统中的应用范围更广，特别是在推荐信息爆炸和数据稀疏等场景下。随着深度学习技术的发展，基于深度学习的文本生成技术在推荐系统中得到了广泛的应用。文本生成技术主要包括两种：一种是生成式模型，通过对文本进行一定程度上的解码，得到目标词；另一种是生成式与规则结合，通过一定程度上的编码，将规则融入到文本中，实现对目标词更深层次的理解。

3.3.2 检索相关信息

信息检索是一项技术，旨在将用户的查询转化为向量形式，进而利用基于内容的检索策略来搜寻符合用户查询条件的文档[6]。此过程主要包含三大步骤：1)将查询转化为向量，并计算每个文档与查询的相似度；2)从相似文档中筛选出满足用户查询条件的文档，并根据其对查询的关联度进行排序；3)根据用户的个性化偏好，为其呈现最终的搜索结果。该技术融合了传统检索技术和大语言模型技术的优势，有效地解决了传统检索方法中的难题。然而，此方法也面临一些挑战，如基于内容的检索方式在理解用户输入信息时存在局限，且缺乏深入的语义分析。

3.3.3 虚拟助理功能

智能助手是当下人工智能领域的重要研究热点，其应用领域正在不断拓展，涵盖了智能家居、智能医疗[43]、智能客服等多个方面。以 Google Home 为例，这款基于大语言模型的智能助手，能够通过用户输入的关键词迅速检索相关信息。当用户通过语音控制表达需求时，系统能够精准识别关键词并执行相应操作。例如，当用户希望观看电影而电视上未播放相关内容时，系统会准确捕捉用户需求，进一步搜索并呈现相关电影资源。Google Home 以其高效、便捷的特点，为用户解决了寻找和查找电影资源的难题，展现了智能助手在现实生活中的应用价值。

3.3.4 个性化内容推荐

内容推荐旨在将用户需求与现有内容精准匹配，使用户能够迅速定位并获取心仪的信息。具体而言，内容推荐致力于向用户提供个性化且质量上乘的信息，以满足用户多样化的需求。在此过程中，推荐系统的核心目标是精准识别并满足用户的个性化需求。以电商平台为例，系统能够依据用户的浏览历史及购买记录，智能推荐相关商品。在实际应用中，通常借助机器学习模型，通过深度挖掘用户行为数据，分析用户对各类内容的偏好程度，再利用算法模型将符合用户喜好的内容精准呈现[4]。例如，利用大型语言模型，可以为用户提供某一商品在特定方面的推荐结果。

3.4 根据数据特征的分类

数据类型主要指数据的来源，涵盖原始数据、中间特征和最终结果。原始数据指直接从原始文本中提取的信息，这类信息往往包含一些明显的噪声或冗余内容。中间特征则指通过对原始数据进行操作后得到的特征，主要包括两个方面：1)描述个体事物的特征；2)展现个体事物与其他事物间的关联。中间结果指在生成特征后，需要根据其是否为最终结果进行判断。如果中间结果即是最终结果，那么就不存在中间数据。由此，对于大型语言模型而言，数据类型主要划分为两类：1)原始数据；2)中间特征。最终结果的生成则是将原始数据与中间特征进行结合。

4 大语言模型推荐技术的最新研究动态

4.1 推荐技术的整体进展

语言模型推荐技术在电子商务领域的应用已相当普及，尤其体现在商品描述的精准性上，有效助力消费者迅速定位所需商品。用户可借助语言模型深入学习特定商品的描述，并通过语义相似性推荐类似商品。此外，大语言模型同样适用于视频、音乐、新闻等内容的描述，增强用户对这些内容的理解。

在社交媒体领域，大语言模型可助力用户迅速掌握不同群体对同一产品的看法，从而推荐符合其兴趣的信息。例如，通过语言模型获取社交媒体相关内容，以增进理解与应用。而在自然语言处理领域，大语言模型能够识别句子中的词汇，实现句子级别的语义相似性。在推荐系统中，语言模型可被用来学习一句话或一个词，并通过相似度计算为用户推荐其感兴趣的内容。

4.2 关键技术突破研究

为了提升推荐系统的准确性，研究者们通常使用许多技术来提升模型的性能，包括：1)提升模型对用户行为序列的建模能力，通过自回归、循环神经网络等方法将用户行为序列建模为长序列；2)提升模型对用户行为序列的分解能力，通过自适应学习的方式，利用多种不同的深度神经网络对用户行为序列进行分解；3)提升模型对用户行为序列的可解释性，通过将用户行为转化为有意义的文本来实现；4)提升模型对用户行为序列的生成能力，通过自监督学习生成有意义的文本。

4.2.1 深度学习在推荐系统中的应用

在推荐系统领域，深度神经网络因其强大的学习能力为处理用户与物品间复杂的交互模式提供了新的思路[44-45]，如图 3 所示。基于深度学习的用户-物品匹配研究主要集中在两个方向。1)匹配函数学习：这种方法旨在通过深度学习技术构建复杂的用户-物品匹配函数，以捕捉用户与物品之间微妙的关联[46]。例如，神经协同过滤(Neural Collaborative Filtering，NCF)[47] 引入多层感知器(Multi-Layer Perceptrons，MLP)来构建富有表现力和复杂性的匹配函数，以有效地处理含有噪声的隐式反馈数据，从而显著提升推荐性能。2)表征学习：此方向侧重于利用神经网络将用户和物品的特征转化为更易于匹配的潜在空间。Bert4Rec[48] 采用深度双向自注意机制，将用户的历史行为序列转化为潜在空间，为顺序推荐任务提供了有力支持。而 Caser[49] 则提出了一种卷积序列模型，通过水平和垂直卷积滤波器识别用户复杂的历史交互序列，进一步丰富了用户和物品的表征。

此外，鉴于用户-物品交互中固有的图形结构，研究人员开始探索图神经网络在推荐任务中的应用。例如 NGCF[50] 和 LightGCN[51] 等模型利用高阶邻居信息来增强用户和物品的表征，以实现更精准的用户-物品匹配，为推荐系统领域注入了新的活力。

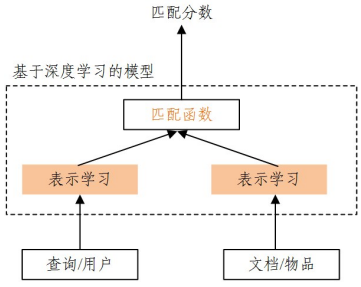

图 3　基于深度学习的推荐系统

Fig. 3　Recommendation system based on deep learning

4.2.2　判别式大语言模型

面向大语言模型的推荐系统可分为判别式和生成式，如图 4 所示。其中，判别式大语言模型的基本思想是利用用户已知的行为特征来预测未知用户的行为特征。这类模型中最常见的是基于历史数据学习得到的模型，其中包括朴素贝叶斯、支持向量机等。一类判别式大语言模型是基于自然语言处理技术，特别是 BERT 系列模型[52]，正逐渐占据重要地位。鉴于判别语言模型在自然语言理解任务中展现出的卓越性能和专业性，它们常被视作下游任务的嵌入核心。同样，在推荐系统中，这些模型也发挥着不可或缺的作用。目前，大多数研究工作都致力于将预训练的判别语言模型（如 BERT）与特定领域的数据进行对齐，通过微调技术来优化模型的表现。这种方法能够充分利用预训练模型在大量文本数据中学到的丰富知识，并将其应用于推荐系统的特定场景中。此外，为了进一步提高模型的训练效率和性能，一些研究还探索了诸如快速调音等先进的训练策略。这些策略旨在加速模型的训练过程，同时保持甚至提升模型的预测能力。表 1 列出了常用数据集。这些数据集为推荐系统领域的研究人员提供了宝贵的参考和实验基础，有助于推动该领域的进一步发展。

表 1　基于 LLM 的推荐方法中使用的常用数据集列表

Table 1　List of common datasets used in existing LLM-based recommendation methods

数据集名称	应用场景	适用任务	链接
Amazon Review	商务	序列推荐；协同过滤推荐	https://jmcauley.ucsd.edu/data/amazon/
Amazon-M2	商务	序列推荐；协同过滤推荐	https://arxiv.org/abs/2307.09688
Steam	游戏	序列推荐；协同过滤推荐	https://cseweb.ucsd.edu/~jmcauley/datasets.html#steam_data
MovieLens	电影	通用	https://grouplens.org/datasets/movielens/
Yelp	商务	通用	https://www.yelp.com/dataset
Douban	电影、音乐、书籍	序列推荐；协同过滤推荐	https://paperswithcode.com/dataset/douban
MIND	新闻	通用	https://msnews.github.io/assets/doc/ACL2020_MIND.pdf
U-NEED	商务	对话推荐	https://github.com/LeeeeoLiu/U-NEED
Beauty	商务	通用	https://tianchi.aliyun.com/dataset/90390/

1）微调：预训练语言模型的微调技术已经成为自然语言处理领域中的一项关键通用技术，尤其在推荐系统等领域中受到了广泛的重视。微调技术的核心思想在于，利用一个预先在大规模文本数据上训练好的语言模型，该模型已经捕获丰富的语言表示能力，随后通过针对特定任务或领域的数据集进行进一步的训练，以使其更好地适应和满足特定任务的需求。

微调的过程包括用学习到的参数初始化预训练的语言模型，然后在特定于推荐的数据集上训练它。该数据集通常包括用户-项目交互、项目的文本描述、用户配置文件和其他相关上下文信息。在微调期间，模型的参数根据特定于任务的数据进行更新，使其能够适应并专门用于推荐任务。预训练阶段和微调阶段的学习目标可能存在差异。预训练阶段的目标是使模型能够从大规模文本数据中学习到普遍适用的语言表示能力，而微调阶段则更侧重于针对特定推荐任务的需求进行优化，以实现更好的推荐效果和性能。这种差异体现了微调技术在适应不同任务需求方面的灵活性和有效性。

2）指令优化：提示调优[24]不是通过设计特定的目标函数使 LLM 适应不同的下游推荐任务，而是试图通过硬/软提示和标签词表达器使推荐的调优对象与预训练的损失保持一致。硬/软提示作为指导模型行为的方式，可以显著影响 LLM 在推荐任务中的表现。硬提示通常指直接插入到模型输入中的固定文本，而软提示则通过修改模型内部参数或权重来影响其行为。标签词表达器则用于将推荐任务中的特定概念或类别映射到模型可以理解的词汇或向量表示中。以 Penha 和 Hauff[53] 的工作为例，他们利用 BERT 模型的掩模语言建模头来探索模型对使用完形提示的项目类型的理解。通过向模型展示部分掩蔽的项目描述，并要求其预测缺失的部分，研究者能够分析 BERT 模型对不同类型项目的识别和表示能力。他们还利用 BERT 的下一个句子预测头部和表征的相似性来比较相关和不相关的搜索和推荐查询文档输入。这种方法允许他们评估模型在区分相关和不相关文档方面的能力，从而为推荐系统提供更准确的候选项目列表。

4.2.3　生成式大语言模型

与判别模型相比，生成模型具有更好的自然语言生成能力。与多数基于判别模型的方法不同，这些方法通常侧重于将大型语言模型（LLM）学习到的表示与推荐领域进行对齐，而基于生成模型的工作则将推荐任务转化为自然语言任务。在这些生成式框架中，研究人员运用了一系列先进的技术，如上下文学习、提示调优（Prompt Tuning）和指令调优（Instruc-

tion Tuning),来调整 LLM,使之能够直接生成个性化的推荐结果。尤其值得注意的是,随着 ChatGPT 等模型所展现出的卓越性能,这一领域的研究近期受到了前所未有的关注。如图 4 所示,根据是否对模型参数进行调优,这些基于 LLM 的生成式推荐方法可以进一步细化为两大范式:非调优范式(Non-tuning Paradigm)和调优范式(Tuning Paradigm)。非调优范式侧重于利用 LLM 的预训练能力,而无需进一步微调其内部参数;而调优范式则通过对模型参数的精心调整,旨在进一步提升模型在特定推荐任务上的性能。这两种范式各有优劣,为推荐系统领域的研究提供了丰富的探索空间。

1)非微调范式:LLM 在许多看不见的任务中表现出强大的零/少射能力[54-55]。因此,最近的一些研究假设 LLMs 已经具有推荐能力,并试图通过引入特定的提示来触发这些能力。他们采用了最近的教学和上下文学习实践[54],采用 LLM 来推荐任务,而不调整模型参数。根据提示是否包含示范实例,该范式的研究主要分为提示和上下文学习两大类。

(1)提示:主要运用于设计合适的指令和提示,帮助 LLMs 更好地理解和解决推荐任务。Liu 等[56]系统地评估了 ChatGPT 在 5 个常见推荐任务上的性能,即评级预测、顺序推荐、直接推荐、解释生成和评论总结。他们提出了一个通用的推荐提示构建框架,该框架包括:任务描述,使推荐任务适应于自然语言处理任务;行为注入,结合用户-物品交互来帮助 LLMs 捕获用户偏好和需求;格式指标,约束输出格式,使推荐结果更具可理解性和可评估性。同样,Dai 等[57]对 ChatGPT 在 3 种常见的信息检索任务上的推荐能力进行了实证分析,包括点式、成对式和列表式排序。他们针对不同类型的任务提出了不同的提示,并在提示的开头引入了角色说明,以增强 ChatGPT 的域适应能力。

(2)上下文学习:上下文学习是 GPT-3 和其他 LLMs 用来快速适应新任务和信息的一种技术。通过一些演示输入的标签,他们可以在没有额外参数更新的情况下预测未知输入的标签[58]。因此,一些作品试图在提示符中添加演示示例,以使 LLMs 更好地理解推荐任务。此外,可以使用合适的演示来控制 LLM 的输出格式和内容[59],这可以改进常规的评价指标。这对于开发一个稳定而强大的推荐系统至关重要。

2)调优范式:随着大型语言模型(LLM)的崛起,研究者们开始探索如何将这些强大的文本生成和理解能力应用于推荐系统中。尽管 LLM 具备强大的零次或少数次学习能力,但其在特定推荐任务上的表现仍受到数据分布和模型训练的局限。因此,为了进一步提升 LLM 在推荐系统中的应用效果,研究者们提出了多种调优范式,包括微调(fine-tuning)、提示调优(prompt tuning)和指令调优(instruction tuning)。

(1)微调范式:这是最直接的方法,通过让 LLM 在特定推荐任务的数据集上进行训练,对模型参数进行微调。在这种范式下,LLM 通常作为编码器来提取用户或项目的特征表示,然后根据推荐任务的特定损失函数对模型参数进行更新。微调范式能够有效地提升模型在特定任务上的性能,但缺点是需要大量的标注数据和计算资源。

(2)提示调优:为了解决微调范式中的标注数据不足问题,研究者们提出了提示调优范式。这种方法通过设计特定的提示(prompt)来引导 LLM 生成符合推荐任务要求的输出。提示可以是文本、模板或其他形式的信息,用于激活 LLM 中与推荐任务相关的知识。通过优化提示的设计,可以在不改变 LLM 主体参数的情况下,提高模型在推荐任务上的性能。

(3)指令调优:指令调优范式是提示调优的一种扩展,它允许 LLM 同时处理多个不同的推荐任务。在这种范式下,LLM 被训练以理解并遵循包含不同类型。

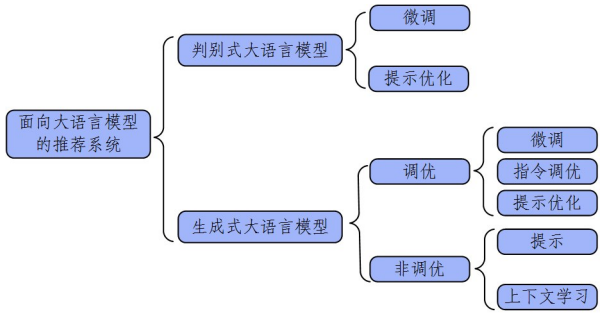

图 4 面向推荐系统的大语言模型研究分类

Fig. 4 Research classification on large language models for recommendation systems

在实践方面,为了实现物品推荐,生成模型在推理阶段执行生成对齐操作[60]。给定自然语言形式的用户表述,生成模型首先通过集束搜索自回归地生成物品标识符。在这里,我们将生成分为两类,即自由生成和约束生成。对于自由生成,在每个生成步骤中,模型会在整个词汇表中搜索,并选择概率最高的前 K 个标记(tokens)作为下一步生成的后续输入。然而,在整个词汇表中搜索可能会导致生成不在语料库中的标识符[61-62],从而使得推荐无效。

为了解决这个问题,早期的研究利用精确匹配进行对齐,即进行自由生成并简单地丢弃无效的标识符。然而,由于无效标识符的存在,它们的准确性仍然较差,尤其是对于基于文本元数据的标识符。为了提高准确性,BIGRec[63]提出通过计算生成的标记序列表示与物品表示之间的 L2 距离,将生成的标识符对齐到有效物品上。这样,每个生成的标识符都能确保对齐到有效的物品标识符上。

在同一时期,对齐生成中的约束生成也被广泛研究[61,64-66]。文献[61]和文献[64]提出利用 Trie 树进行约束生成,确保生成的标识符是有效的。然而,Trie 树严格地从第一个标记开始生成有效的标识符,因此推荐的准确性高度依赖于前几个生成标记的准确性。为了解决这个问题,TransRec[66]利用 FM-索引实现了无位置约束的生成,允许从有效标识符的任何位置生成标记。然后,通过从不同视图聚合生成的有效标记,将其对齐到有效的标识符上。

除了需要有效生成来向用户推荐现有物品的典型推荐外,另一个研究方向是利用模型的生成能力来创造全新的物品[39,67-68]。例如生成个性化的服装搭配[68],可以作为时尚工厂的指导。因此,在这一研究方向中,采用了自由生成,使推荐系统能够充分利用生成潜力。表 2 列出了在推荐领域方面目前的生成式推荐方法。

表2 具有代表性的生成式推荐方法总结

Table 2 Overview of representative generative recommendation methods

方法	架构	生成类型	数据集	领域
RecSysLLM[64]	GLM-10B	Trie	Sports,Beauty,Toys	电子商务
P5[41]	T5-small, T5-Base	Free	Sports,Beauty,Toys,Yelp	电子商务 餐厅
How2index[61]	T5-small	Trie	Sports,Beauty,Yelp	电子商务 餐厅
PAP-REC[69]	T5	Free	Beauty,Sports,Toys	电子商务
VIP5[10]	T5-small	Free	Clothing,Sports,Beauty,Toys	电子商务
UniMAP[67]	Redpajama-3B	Free	Baby,Beauty,Clothing Grocery,Sports,Toys,Office	电子商务
TIGER[65]	Transformer-based model	Free	Sports,Beauty,Toys	电子商务
LC-Rec[71]	LLaMA-7B	Trie	Instruments,Arts,Games	电子商务
TransRec[66]	BART-large,LLaMA-7B	FM-index	Beauty,Toys,Yelp	电子商务 餐厅
M6-Rec[39]	M6	Free	TaoProduct	电子商务
BIGRec[63]	LLaMa-7B	Free	Games,MovieLens25M	电子商务 电影
LMRecSys[72]	BERT-Base,GPT2-Small, GPT2Mediu,GPT2-Large,GPT2-XL	Free	MovieLens1M	电影
NIR[73]	GPT-3	Free	MovieLens100K	电影
RecRanker[74]	LLaMA2-7B	Free	MovieLens100K,MovieLens100M BookCrossing	电影 书籍
InstructRec[75]	Flan-T5-XL	Free	Games,CDs	电子商务
Rec-GPT4V[76]	GPT4-V LLaVA-7B,LLaVA-13B	Free	Sports,Clothing,Beauty,Toys	电子商务
DEALRec[60]	LLaMA-7B	Free	Games,Book,MicroLens-50K	电子商务

5 大语言模型推荐技术面临的挑战

5.1 数据安全与隐私方面的挑战

5.1.1 个人隐私的泄露风险

随着网络与移动技术的广泛渗透,现代社会的生活与网络紧密相连,人们在互联网上留下了丰富的数字足迹。无论是通过浏览器浏览网页,还是在搜索引擎中寻求信息,亦或是在社交平台上分享生活点滴,这些行为都被网站记录并整合为用户个人信息。这些平台利用这些信息,为用户提供个性化服务[77]。例如,用户在搜索引擎中查询书籍时,系统会根据其搜索历史推荐相关书单;用户在社交平台上发布动态时,平台会基于其内容为其推荐相关用户。这些应用程序通过收集用户的浏览历史、行为偏好和社交联系等信息,为用户提供个性化的服务体验[78]。

5.1.2 数据的不当使用

经过深入调研,在实际应用中发现,众多推荐系统存在对用户数据的过度使用现象。用户在使用推荐系统时,并未主动提供个人信息,然而系统通过各种途径依然能够获取到这些数据。这主要是因为用户在享受推荐服务时,往往未意识到自身数据的隐私保护重要性,从而在无意识中泄露了个人信息。这种现象导致用户在使用推荐系统时感到隐私受到侵犯,认为自己在这些系统中毫无隐私可言。另外,存在部分推荐系统通过向用户推送广告来实现盈利,虽然在一定程度上这种做法可能损害用户对推荐系统的信任,但是部分推荐系统为了获得更大的收益,还是没有将其进行删除。想要保证用户能够持续使用推荐系统并同时实现广告收益,就必须高度重视用户隐私保护,确保用户数据得到妥善处理和保护。

5.1.3 法律法规遵循问题

在大语言模型推荐应用中,模型的输出结果往往会受到合规性要求的制约。比如,用户行为数据中可能包含某些特定内容的信息,这些信息在推荐过程中的使用需要慎重考虑,以确保用户能够正确理解并妥善利用。例如,当用户在某一平台上发布广告时,若广告内容被检测出含有敏感信息,平台有权拒绝展示该广告。另外,也存在用户可能无意中将个人敏感信息泄露给拥有相应权限的个体,从而被用于非法活动的风险。因此,在应用大语言模型推荐技术时,必须对相关数据的使用施加必要的限制和规范,以防止用户在使用该技术时陷入合规性困境。

5.2 技术的可解释性问题

在推荐系统中,用户往往不会主动提供过多信息,但他们的负面反馈,如对产品的不满或对某个人的厌恶,可能会对推荐系统产生不利影响。这种负面反馈可能源于其他用户的评论,或是用户在使用产品过程中的不佳体验。当推荐系统接收到这些负面信息时,用户可能会认为系统未能准确理解其需求,进而对系统的可靠性产生怀疑。

此外,在用户模型训练过程中,若训练样本与目标样本的数量存在显著差异,训练出的模型可能无法准确反映目标用户的真实行为特征。这同样可能导致用户在使用推荐系统时对其信任度降低。为了提高推荐系统的可靠性和用户满意度,需要不断优化算法,提升数据处理能力,并尽可能收集更多、更全面的用户信息。

5.2.1 不准确的解释

在实际应用中,模型预测准确性的挑战往往伴随着解释性的迫切需求。特别是在模型输出与实际观测结果产生偏差

时，可解释性的缺乏成为了一个需要解决的问题。以物品类别预测为例，当模型预测的物品类别与用户实际行为所反映的类别不一致时，模型会简单地报告“预测的物品类别与用户行为类别不一致”这一结果。然而，这种直接的输出缺乏深层次的解释，可能导致用户误解为接收到了错误信息，进而引发不必要的疑虑和情绪反应。因此，在构建和部署预测模型时，设计合理的解释机制显得尤为重要。这不仅有助于用户更好地理解模型的工作原理和预测结果，还能在模型出现错误时提供有价值的反馈，促进模型的持续改进和优化。为了提升模型的可解释性，研究者们可以采取多种策略，如采用更透明的模型结构、开发专门的解释工具或算法以及结合领域知识来提供定制化的解释等。通过这些努力，我们可以使预测模型不仅具备高度的准确性，还能在出现偏差时提供清晰、合理的解释，从而增强用户对模型的信任感和满意度。

5.2.2 可解释性与合规性

除了模型的可解释性，推荐系统的合规也是一个很重要的问题。由于推荐系统是服务于大众且面向社会的，在业务上会存在一定的不确定性，会产生一定的合规问题。例如，《网络安全法》和《个人信息保护法》等法律法规，要求公司必须对用户信息进行脱敏处理后才能提供给用户，这种情况下推荐系统需要在用户信息脱敏后才能提供给用户。在没有得到用户授权的情况下就直接给出相关内容是不合适的。此外，在某些特定场景下，如新闻媒体，推荐系统有可能会直接向用户提供敏感信息，由此可见，这种情况下推荐系统需要保证内容的合规性。

5.3 资源消耗问题

由于大语言模型的参数数量大，在训练过程中需要耗费大量的计算资源，然而，大语言模型所消耗的计算资源会随着训练数据量和数据维度的增加而不断上升。对于互联网平台来说，数据量越大，意味着越能吸引用户，从而对系统性能有更高的要求。因此，在互联网平台上建立足够规模的模型是重要问题之一。在现有技术下，可供用户选择的模型较少，这也给大语言模型推荐系统带来了局限性。但是随着深度学习技术和数据规模的不断发展，深度学习在推荐系统中的应用范围将会越来越广。相信在不久的将来，大语言模型推荐系统会成为推荐系统中不可或缺的部分。

6 大语言模型推荐技术的未来发展方向

随着技术的不断发展，用户在使用推荐系统时，会逐步产生对推荐结果的预期。现有的用户体验往往是以用户进行交互后的反馈为准，因此，在推荐系统中将会越来越多地加入一些增强用户体验的技术。例如，可以通过强化学习等技术来改善推荐系统对新用户和新物品的学习效果，提升用户对新推荐物品的喜好程度；也可以在学习过程中，引入一些用户偏好信息、历史购买信息等来丰富模型的数据维度；还可以考虑加入一些新的知识图谱等来增加模型对新物品的理解能力和兴趣程度；还可以引入一些基于大数据、深度学习的方法来丰富模型结构和参数。

6.1 强化数据隐私保护措施

数据隐私在数据安全领域中占据重要地位，其合理保护不仅能够增强用户对于推荐系统的信任度，更是维护用户权益的关键所在。在大语言模型推荐技术中，采取多元化的数据隐私保护策略尤为必要。例如，利用司态加密技术，可以对用户数据进行加密处理，并在随后的计算过程中保持加密状态，从而降低计算成本，实现高效的数据利用与保护。同时差分隐私技术的运用也至关重要。通过将用户的评分向量化，并进行同态加密处理，可以在保护用户隐私的同时确保推荐系统的精准性和有效性[79]。值得注意的是，差分隐私的强大隐私保护能力在与同态加密技术的结合中得到了充分发挥，为用户数据隐私提供了更为坚实的保障。此外，随着技术的不断进步，支持零知识证明的技术也为大语言模型推荐系统带来了新的可能性。将零知识证明应用于推荐系统中，不仅可以进一步提升系统的安全性和可靠性，还能在保护用户隐私的同时，实现更为精准和个性化的推荐服务。

6.2 提高技术的可解释性

推荐系统的用户往往不是理性的，在推荐系统中，如何让用户了解推荐结果背后的原因也是推荐系统在未来研究中需要考虑的一个问题。通过解释推荐结果背后的原因，一方面可以使用户对推荐结果有更加清晰地理解，另一方面也能让用户对推荐结果产生信任，从而提高用户满意度。对于可解释性这一问题，国内外很多学者都进行了研究，并取得了一些成果。例如在 2019 年，美国加州大学伯克利分校的教授 Kim 等提出了一种解释推荐结果的方法，即基于解释模型，通过给用户呈现预测模型和真实预测模型之间的差异来提高用户对推荐结果的信任。Kim 等在预测结果中加入了解释因子，并构建了一种解释模型，对用户的行为进行解释，从而提高用户对预测结果的信任。为了更好地解释推荐结果，许多学者还提出了一些方法，例如 Chan 等提出了一种基于用户和物品的关系预测模型的方法，并将其应用于推荐系统中。但也有一些学者指出，目前的可解释性方法还存在一些问题。例如，可解释性方法会对模型造成一定程度的限制；再如，如何在不影响模型预测结果准确性的前提下提供更多的信息来帮助用户理解推荐结果等。未来可以针对这些问题进行进一步研究。

6.3 发展自动化内容创作能力

随着用户对推荐系统需求的不断升级，内容生成功能在推荐系统中的地位愈发重要。展望未来，自动化内容生成能力将持续受到重视，成为推荐系统技术应用的关键一环。具体而言，大语言模型推荐技术将在内容生成领域发挥重要作用，为用户提供更优质的使用体验，并助力提升生产力。自动化内容生成作为大语言模型推荐技术的核心应用领域之一，将持续拓展其影响力[80]。大语言模型推荐技术将成为写作助手的重要组成部分，协助作者根据用户喜好生成高质量、个性化的作品，以满足不同用户的需求。媒体制作和广告领域也将广泛应用大语言模型推荐技术，实现文本、图像和视频内容的自动生成，以提高生产效率、降低成本，并实现精准推送，满足用户的个性化需求。新闻机构将借助大语言模型推荐技术自动生成新闻报道，提升新闻发布的速度和覆盖范围，同时增加新闻的多样性和个性化，为用户提供更加丰富的信息获取体验。总之，大语言模型推荐技术在内容生成领域的应用将持续深化，助力提升用户体验和生产效率，满足用户不断升级的需求。

结束语 本文综述了大型语言模型在推荐系统中的研究进展,将已有的工作分为判别式推荐和生成式推荐,并对常用数据库和生成式推荐方法进行了总结,进一步地,通过领域自适应的视角,详细阐述了这些方法的实际应用和适应性调整。在判别式推荐和生成式推荐中提供了调优、提示、提示调优和指令调优的定义和区别,总结了许多相关研究的共同发现和挑战。为研究人员全面了解大型语言模型在推荐系统中的应用,探索潜在的研究方向提供了宝贵的资源。展望未来,随着计算能力的不断提高和人工智能领域的扩展,大语言模型推荐系统在现代社会发展中有着巨大的发展未来,在多个领域中都能够发现其身影,为此,想要保证推荐系统能够顺应时代发展潮流,便要结合现代社会发展趋势,对其进行深度研究和创新突破,这样才能保证在面对多种问题和挑战时得到及时的解决。此外,随着道德考虑的日益突出,未来面向大语言模型的推荐系统也可能从本质上整合公平、问责制和透明度。总之,虽然我们在理解和实施 LLMs 方面取得了实质性进展,但未来的旅程充满了创新和改进的机会。

参 考 文 献

[1] GAO Y,SHENG T,XIANG Y Y,et al. Chat-rec:Towards interactive and explainable llms augmented recommender system[J]. 2003.

[2] CHEN J,MA L,LI X,et al. Knowledge grap hcompletion models are few-shot learners:An empirical study of relation labeling in e-commerce with llms[J]. arXiv:2305.09858,2023.

[3] CHEN X,FAN W,CHEN J,et al. Fairly adaptive negative sampling for recommendations[C]// Proceedings of the ACM Web Conference. 2023:3723-3733.

[4] QIN L,WU W S,LIU D,et al. Autonomous planning and processing framework for complex tasks based on large language models[J]. Acta Automatica Sinica,2024,50(4):862-872.

[5] FAN W,ZHAO X,CHEN X,et al. A comprehensive survey on trustworthy recommender systems[J]. arXiv:2209.10117,2022.

[6] HE X,DENG K,WANG X,et al. Lightgcn:Simplifying and powering graph convolution network for recommendation[C]// Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020:639-648.

[7] FAN W,DERR T,MA Y,et al. Deep adversarial social recommendation[C]// 28th International Joint Conference on Artificial Intelligence(IJCAI-19). International Joint Conferences on Artificial Intelligence,2019:1351-1357.

[8] ZHENG L,NOROOZI V,YU P S. Joint deep modeling of users and items using reviews for recommendation[C]// Proceedings of the tenth ACM International Conference on Web Search and Data Mining. 2017:425-434.

[9] ZHANG S,YAO L,SUN A,et al. Deep learning based recommender system:A survey and new perspectives[J]. ACM Computing Surveys(CSUR),2019,52(1):1-38.

[10] FAN W,LIU C,LIU Y,et al. Generative diffusion models on graphs:Methods and applications[J]. arXiv:2302.02591,2023.

[11] HIDASI B,KARATZOGLOU A,BALTRUNAS L,et al. Session-based recommendations with recurrent neural networks[J]. arXiv:1511.06939,2015.

[12] FAN W,MA Y,YIN D,et al. Deep social collaborative filtering[C]// Proceedings of the 13th ACM Conference on Recommender Systems. 2019:305-313.

[13] FAN W,MA Y,LI Q,et al. Graph neural networks for social recommendation[C]// The World Wide Web Conference. 2019:417-426.

[14] QIU Z,WU X,GAO J,et al. U-bert:Pre-training user representations for improved recommendation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021:4320-4327.

[15] BROWN T,MANN B,RYDER N,et al. Language models are few-shot learners[J]. NeurIPS,2020.

[16] ZHOU L,PALANGI H,ZHANG L,et al. Unified vision-language pre-training for image captioning and vqa[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:13041-13049.

[17] LI J,LIU Y,FAN W,et al. Empowering molecule discovery for mole-cule-caption translationwith large language models: A chatgpt perspective[J]. a-Xiv:2306.06615,2023.

[18] CHEN Z,MAO H,LI H,et al. Exploring the potential of large language models(llms) in learning on graphs[J]. arXiv:2307.03393,2023.

[19] ZHAO W X,ZHOU K,LI J,et al. A survey of large language models[J]. arXiv:2303.18223,2023.

[20] ZHANG J,XIE R,HOU Y,et al. Recommendation as instruction following:A large language model empowered recommendation approach[J]. arXiv:2305.07001,2023.

[21] DEVLIN J,CHANG M W,LEE K,et al. Bert:Pre-training of deep b-idirectional transformers for language understanding[J]. arXiv:1810.04805,2018.

[22] RADFORD A,NARASIMHAN K,SALIMANS T,et al. Improving Language Understanding by Generative Pre-Training[OL]. https://openai.com/index/language-unsupervised/.

[23] RAFFEL C,SHAZEER N,ROBERTS A,et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. The Journal of Machine Learning Research,2020,21(1):5485-5551.

[24] VASWANI A,SHAZEER N,PARMAR N,et al. At-tention is all you need[J]. Advances in Neural Information Processing Systems,2017,30:5998-6008.

[25] ZHANG Z,ZHANG G,HOU B,et al. Certified robustness for large language-models with self-denoising[J]. arXiv:2307.07171,2023

[26] THOPPILAN R,DE FREITAS D,HALL J,et al. Lamda:Language models for dialog applications[J]. arXiv:2201.08239,2022.

[27] CHOWDHERY A,NARANG S,DEVLIN J,et al. Palm:Scaling language modeling with pathways[J]. arXiv:2204.02311,2022.

[28] CHANG W L,LI Z,LIN Z,et al. Vicuna:An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality[C/OL]// https://vicuna.lmsys.org(accessed 14 April 2023). 2023.

[29] KIM H J,CHO H,KIM J,et al. Self-generated in-context learning:Leveraging auto-regressive language models as a demons-

tration generator[J]. arXiv:2206.08082,2022.
[30] RUBIN O,HERZIG J,BERANT J. Learning to retrieve prompts for incontext learning[J]. arXiv:2112.08633,2021.
[31] WEI J,WANG X,SCHUURMANS D,et al. Ch-ain of thought prompting elicits reasoning in large language models[J]. arXiv:2201.11903,2022.
[32] WANG X,WEI J,SCHUURMANS D,et al. Self-consistency improves c-hain of thought reasoning in language-models[J]. arXiv:2203.11171,2022.
[33] ZELIKMAN E,WU Y,MU J,et al. Star:Bootstrapping reasoning with reasoning[J]. Advances in Neural Information Processing Systems,2022,35:15476-15488.
[34] FEI H,LI B,LIU Q,et al. Reasoning implicit sentiment with chain-of-thought prompting[J]. arXiv:2305.11255,2023.
[35] JIN Z,LU W. Tab-cot:Zero-shot tabular chain of thought[J]. arXiv:2305.17812,2023.
[36] KASNECI E,SEßLER K,KÜCHEMANN S,et al. Chatgpt for good? on opportunities and challenges of large language models for education[J]. Learning and Individual Differences,2023,103:102274.
[37] WU S,IRSOY O,LU S,et al. Bloomberggpt:A large language model for finance[J]. arXiv:2303.17564,2023.
[38] CUI Z,MA J,ZHOU C,et al. M6-rec:Generative pretrained language models are open-ended recommender systems[J]. arXiv:2205.08084,2022.
[39] CUI Z Y,MA J X,ZHOU C,et al. M6-rec:Generative pretrained language models are open-ended recommender systems[J]. arXiv:2205.08084,2022.
[40] GENG S,LIU S,FU Z,et al. Rec-ommendation as language processing(rlp):A unified pretrain,personalized prompt & predict paradigm(p5)[C]// Proceedings of the 16th ACM Conference on Recommender Systems. 2022:299-315.
[41] WANG X,ZHOU K,WEN J R,et al. To-wards unified conversational recommender systems via knowledge-enhanced prompt learning[C]// Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022:1929-1937.
[42] DENG Y,ZHANG W,XU W,et al. A unified multi-task learning framework for multi-goal conversational recommender systems[J]. ACM T ransactions on Information Systems,2023,41(3):1-25.
[43] SHENG B,GUAN Z Y,LEE L L,et al. diabetes management based on the big language model:potential and prospect [J]. Science Bulletin,2024,69(5):583-588.
[44] CHEN X,ZHANG Y F,QIN Z. Dy-namic explainable recommendation based on neural at-tentive models[C]// AAAI. 2019:53-60.
[45] KANG W C,MCAULEY J L. Self-attentive sequential recommendation[C]// ICDM. IEEE,2018:197-206.
[46] ZHANG Y F,AI Q Y,CHEN X,et al. Joint representation learning for top-n recommendation with heterogeneous information sources[C]// CIKM. 2017:1449-1458.
[47] HE X N,LIAO L Z,ZHANG H W,et al. Neural Collaborative Filtering[C]// WWW. ACM,2017:173-182.
[48] SUN F,LIU J,WU J,et al. BERT4Rec:Sequen-tial recommendation with bidirectional encoder repre-sentations from transformer[C]// CIKM. ACM,2019:1441-1450.
[49] TANG J X,WANG K. Personalized top-n sequential recommendationvia convolutional sequence embedding [C] // WSDM. ACM,2018:565-573
[50] WANG X,HE X N,WANG M,et al. Neural Graph Collaborative Filtering[C]// SIGIR. ACM,2019:165-174.
[51] HE X N,DENG K,WANG X,et al. Lightgcn:Simpli-fying and powering graph convolution network for recommendation[C]// SIGIR. 2020:639-648.
[52] DEVLIN J,CHANG M W,LEE K,et al. BERT:pre-training of deep bidirectional transformers for language understanding [C]// Association for Computational Linguistics. NAACL-HLT (1),2019:4171-4186.
[53] PENHA G,HAUFF C. What does BERT know about books,movies and mu-sic? probing BERT for conversational recommendation[C]// RecSys. ACM,2020:388-397.
[54] BROWN T B,MANN B,NICKRYDER,et al. Language models are fewshot learners[C]// NeurIPS. 2020.
[55] OUYANG L,WU J,JIANG X,et al. Training lan-guage models to follow instructions with human feed-back[C] // NeurIPS. 2022.
[56] LIU J L,LIU C,LV R J,et al. Is chatgpt a good recommender? A preliminary study[J]. CoRR,abs/2304.10149,2023.
[57] DAI S B,SHAO N L,ZHAO H Y,et al. Uncovering chatgpt's capabilities inrecom-mender systems [J]. CoRR, abs/2305.02182,2023.
[58] DAI D M,SUN Y T,DONG L,et al. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers[J]. CoRR,abs/2212.10559,2022.
[59] WANG W J,LIN X Y,FENG F L,et al. Generative recommendation:Towards next-generation recommender paradigm. CoRR [J]. abs/2304.03516,2023.
[60] LIU Y Q,WANG Y,SUN L C,et al. Rec-GPT4V:Multimodal Recommendation with Large Vision-Language Models[J]. arXiv:2402.08670.2024.
[61] HUA W Y,XU S Y,GE Y Q,et al. How to I-ndex Item IDs for Recommendation Foundation Models[J]. arXiv:2305.06569,2023.
[62] LI L,ZHANG Y F,LIU D G,et al. Large language models for generative recommendation:A survey and visionary discussions [J]. arXiv:2309.01157,2023.
[63] BAO K Q,ZHANG J Z,WANG W J,et al. A bi-step grounding paradigm for large language models in recommendation systems [J]. arXiv:2308.08434,2023.
[64] CHU Z X,HAO H Y,OUYANG X,et al. Leveraging large language models for pre-trained recommender systems[J]. arXiv:2308.10837,2023.
[65] RAJPUT S,MEHTA N,SINGH A,et al. Recommender Systems with Generative Retrieval[C] // NeurIPS. Curran Associates,Inc,2023.
[66] LIN X Y,WANG W J,LI Y Q,et al. A multi-facet paradigm to bridge large language model and recom-mendation[J]. arXiv:2310.06491,2023.

[67] WEI T X,JIN B W,LI R R,et al. Towards Universal Multi-Modal Personalization:A Language Model Empowered Generative Paradigm[C]// ICLR. 2024.

[68] XU Y Y,WANG W J,FENG F X,et al.. DiFashion:Towards Personalized Outfit Generation[C]// SIGIR. 2024

[69] LI Z L,JI J C,GE Y Q,et al. PAP-REC:Personalized Auto-matic Prompt for Recommendation Language Model[J]. arXiv:2402.00284,2024.

[70] GENG S J,TAN J T,LIU S C,et al. VIP5:Towards Multimodal Foundation Models for Recommendation[C]// EMNLP. 2023:9606-9620.

[71] ZHENG B W,HOU Y P,LU H Y,et al. Adapting large language models by integrating collabora-tivesemantics for recommendation[J]. arXiv:2311.09049,2023.

[72] ZHANG Y H,DING H,SHUI Z R,et al. Language models as recommender systems:Evaluations and limi-tations[C]// 2021.

[73] WANG L, LIM E P. Zero-Shot Next-Item Recommendation using Large Pretrained Language Models[J]. arXiv: 2304.03153,2023.

[74] LUO S C,HE B W,ZHAO H H,et al. RecRanker:Instruction Tuning Large Language Model as Ranker for Top-k Recommendation[J]. arXiv:2312.16018,2023.

[75] ZHANG J J,XIE R B,HOU Y P,et al. Recommendation as instruction following:A large language model empowered recommendation approach[J]. arXiv:2305.07001,2023.

[76] LIU Y Q,WANG Y,SUN L C,et al. Rec-GPT4V:Multimodal Recommendation with Large Vision-Language Models[J]. arXiv:2402.08670,2024.

[77] GU Z,HE X,YU P,et al. Automatic quantitative stroke severity assessment based on Chinese clinical named entity recognition with domain adaptive pretrained large language model[J]. Artificial Intelligence In Medicine,2024,150:102822.

[78] LI N. Exploration of the Application of Big Language Model in Financial Shared Center [J]. China Agricultural Accounting,2024,34(6):88-90.

[79] LU M F. Research on the Application Principles,Challenges, and Implementation Paths of Large Language Models in the Financial Sector [J/OL]. Journal of Chongqing Technology and Business University(Social Sciences Edition),1-13. [2400-04-08].

[80] LEE G G,LATIF E,WU X,et al. Applying large language models and chain-of-thought for automatic scoring[J]. Computers and Education:Artificial Intelligence,2024,6:100213.

KA Zuming, born in 2000, postgraduate. Her main research interests include recommendation systems and large language models.

ZHAO Peng, born in 1979, Ph.D, associate professor. His main research interests include intelligent information processing, recommendation systems, and distributed computing.