LLMs 推荐发展综述-生成型推荐 & 非生成型推荐

原创 方方 方方的算法花园 2024年11月08日 10:02 北京

点击蓝字 关注我们

写在前面

吉林大学、悉尼科技大学、香港理工大学在2024年10月联合发表了一篇论文《 Towards Next-Generation LLM-based Recommender Systems: A Survey and Beyond》，论文里探讨了基于大语言模型（LLMs）的推荐系统的发展，包括其在表示与理解、设计与应用、工业部署等方面的应用，以及面临的挑战和机遇。（论文链接：https://arxiv.org/pdf/2410.19744）

本文主要介绍论文中关于Scheming and Utilizing 部分的研究进展，包含生成型推荐和非生成型推荐。其他部分内容请参考此系列其他文章。

LLMs的出现为推荐系统引入了一种新的范式，引发了关于如何将 LLMs 有效整合到推荐框架中的广泛研究。根据框架是否需要计算每个候选对象的评分以确定推荐，这一领域的研究可以分为非生成型基于 LLMs 的方法和生成型基于 LLMs 的方法。这些方法与传统推荐系统之间的区别如下图所示。

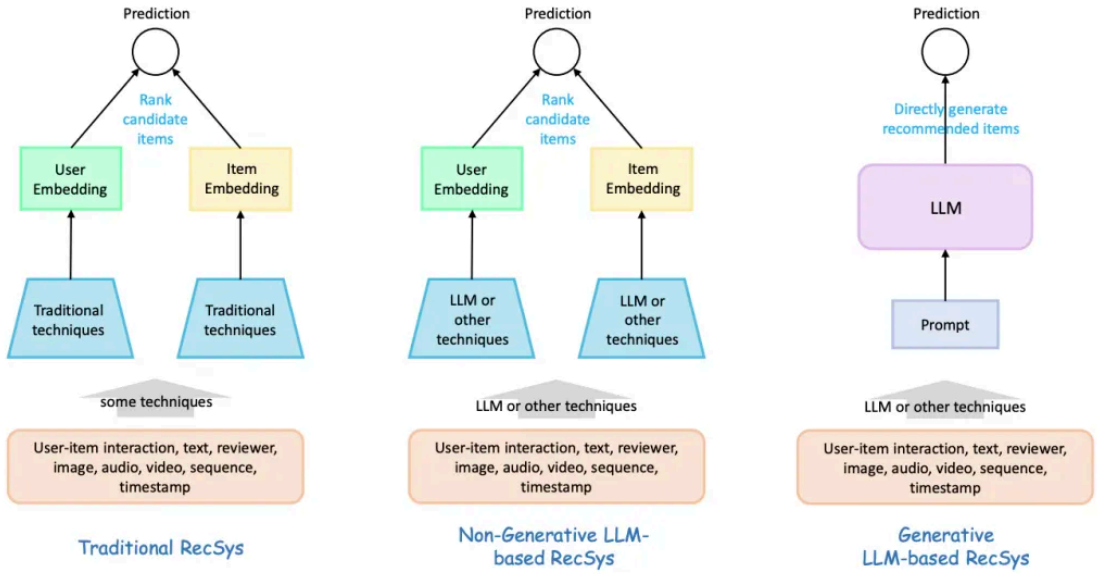
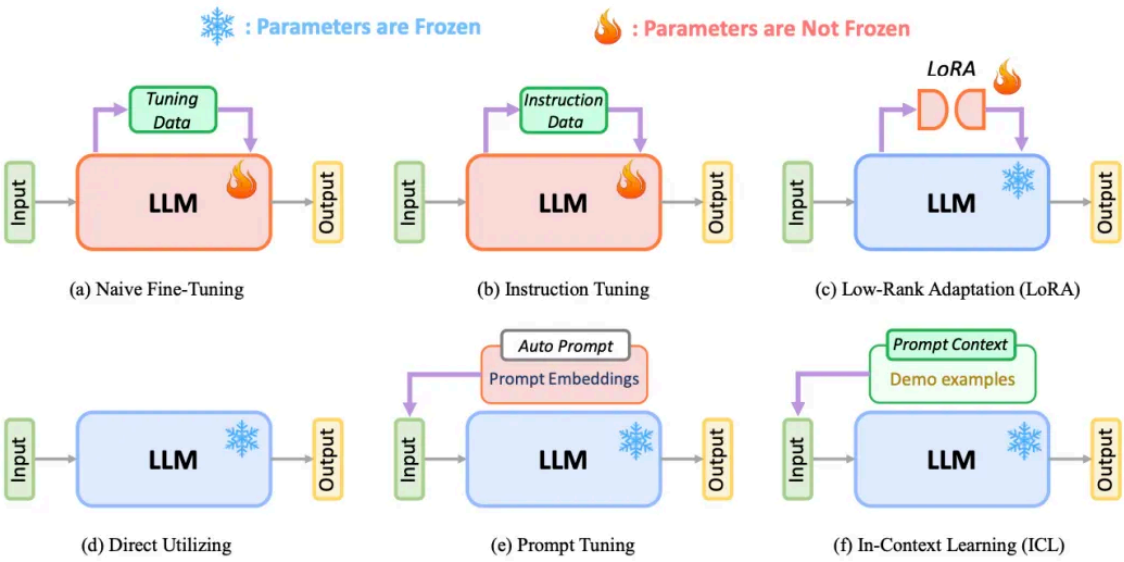


Fig. 4. Different paradigms for recommender systems: (a) Traditional Recommender System; (b) Non-Generative LLM-based Recommender System; and (c) Generative LLM-based Recommender System.

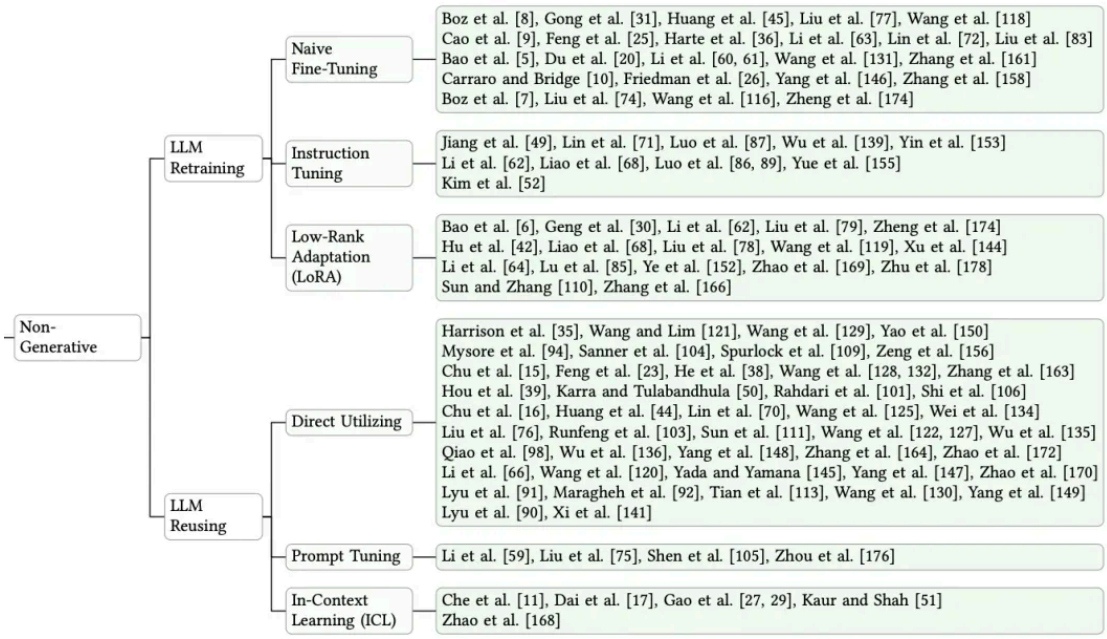
为了更清楚地了解 LLMs 在这些框架中的运用方式，图5 进一步将这些方法分为两种主要策略：LLM 再训练和 LLM 复用。这种分类是基于 LLMs 的参数是否发生改变。



非生成型推荐

非生成型推荐通过将 LLMs 对自然语言的理解纳入推荐过程来增强传统推荐任务。与生成型方法不同，**非生成型的方法不会直接以自然语言输出的形式生成推荐。相反，它们利用 LLMs 来提高推荐模型的准确性和相关性，例如通过利用预训练语言模型中embedding的语义理解来增强排序、评分或特征提取。**这些方法通常涉及多阶段过程，其中 LLMs 在特定阶段发挥作用，例如特征丰富化或排序。

非生成型方法分为LLM retraining和LLM reusing。LLM retraining涉及修改LLM的参数，而LLM reusing 直接使用或者小范围修改LLM参数。



非生成型推荐——LLM retraining

LLM retraining下的三种关键方法：Naive Fine-Tuning、Instruction Tuning 和LoRA。每种方法都代表了一种不同的调整模型参数以提高其在推荐任

务中性能的方法。

① Naive Fine-Tuning

[118] Rethinking large language model architectures for sequential recommendations. arXivpreprint arXiv:2402.09543, 2024.
经过训练后，进一步微调上下文感知嵌入和推荐LLM会带来更好的性能。

[77] Llmrec: Benchmarking large language models on recommendation task. In Proceedingsof the 38th AAAI Conference on Artificial Intelligence, pages 1234–1245, 2024.
研究了监督微调以提高LLM指令遵从能力的有效性。基准测试结果表明，LLM在基于准确性的任务（如顺序推荐和直接推荐）中仅表现出中等水平的熟练程度。

[8] Improving sequential recommendations with llms. arXiv preprint arXiv:2402.01339, 2024.
发现，为推荐任务微调LLM不仅能使它在一定程度上学习任务，还能学习领域的概念。它还表明，微调 OpenAI 的 GPT 比微调 Google 的 PaLM 2表现要好得多。

[31] An unified search andrecommendation foundation model for cold-start scenario. In Proceedings of the 32nd ACM International Conferenceon Information and Knowledge Management, page 4595–4601, 2023.
使用LLM以一种可以帮助解决推荐中的冷启动问题的方式提取领域不变特征。

[45] Recommender ai agent: Integrating largelanguage models for interactive recommendations. arXiv preprint arXiv:2308.16505, 2023. 通过设计从 GPT-4 衍生的模仿数据集来微调一个 70 亿参数的模型，这可以提高交互推荐的能力。

[36]Leveraging large language models for sequential recommendation. In Proceedings of the 17th ACM Conference onRecommender Systems, page 1096–1102, 2023.
以提示-完成对的形式，用特定数据集信息微调LLM，并要求模型为测试提示生成下一个项目推荐。

[25]A large languagemodel enhanced conversational recommender system. arXiv preprint arXiv:2308.06212, 2023.
提出通过从对话推荐系统性能反馈中进行强化学习微调LLM，以提高推荐性能。

[63] Exploring fine-tuning chatgpt for news recommendation. arXivpreprint arXiv:2311.05850, 2023.
探索通过将新闻推荐表述为直接排序和评分任务来微调 ChatGPT。

[9] Aligning large language models with recommendation knowledge. In Proceedings ofthe 25th Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages1051–1066, 2024.

提出通过微调具有推荐知识的数据样本以及编码项目相关性的辅助任务数据样本，使LLM与推荐领域对齐。在电子商务领域，

[83] Conversational recommender system and large language model are made for each other in e-commercepre-sales dialogue. In Proceedings of the 28th Empirical Methods in Natural Language Processing, pages 9587–9605,2023.
研究了将LLM与对话推荐系统相结合以及使用售前对话微调LLM（包括 ChatGLM 和 Chinese-Alpaca-7B）的有效性。

[72] Data-efficient fine-tuningfor llm-based recommendation. arXiv preprint arXiv:2401.17197, 2024.
提出了一种新的数据修剪方法，以高效地识别LLM推荐微调的有影响力样本，这释放了将LLM推荐模型应用于现实世界平台的巨大潜力。

[20] Large language model with graph convolution for recommendation. arXiv preprint arXiv:2402.08859, 2024.
对LLM进行监督微调（SFT），以激活其在任务相关领域的能力。这涉及到用匹配的用户-项目对的描述来训练LLM，使LLM学会对齐用户和项目的描述。

[5] A bi-step grounding paradigm for large language models in recommendation systems. arXiv preprint arXiv:2308.08434,2023.
通过微调LLM来生成项目的有意义标记，并随后识别与生成标记相对应的实际项目，从而在推荐系统中奠定LLM的基础，提高了推荐性能。

[61] Learning structureand knowledge aware representation with large language models for concept recommendation. arXiv preprintarXiv:2405.12442, 2024.
基于概念推荐任务，采用交叉熵损失以端到端的方式微调其结构和知识感知表示学习框架。

[161] Lorec: Large language model for robustsequential recommendation against poisoning attacks. arXiv preprint arXiv:2401.17723, 2024.
通过监督微调利用LLM的开放世界知识，在推荐系统中检测欺诈者，增强了对抗中毒攻击的稳健性。

[131]To recommend or not: Recommendability identification in conversationswith pre-trained language models. arXiv preprint arXiv:2403.18628, 2024
通过在专门为推荐性识别构建的新数据集中训练LLM来微调LLM。这个过程涉及到调整模型的参数，以更好地理解和预测在给定的对话情境中是否需要推荐。

[60] Large language models for nextpoint-of-interest recommendation. arXiv preprint arXiv:2404.17591, 2024.
在基于位置的社交网络数据集上微调LLM，以利用常识知识进行下一个兴趣点推荐任务。

[26] Leveraging large language models in conversational recommender systems.arXiv preprint arXiv:2305.07961, 2023.
专注于在其自身系统内微调LLM，使用大量合成生成的数据。

[146] Fine-tuning large language model based explainable recommendation with explainable quality reward.In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 9250–9259, 2024.
专注于可解释推荐任务，并提出了一种新的基于LLM的可解释推荐模型，称为 LLM2ER，作为骨干，并设计了两个创新的可解释质量奖励模型，用于在强化学习范式中微调这样的骨干。

[10]Enhancing recommendation diversity by re-ranking with large language models.arXiv preprint arXiv:2401.11506, 2024.
展示了LLM如何应用于多样性重排。这项工作采用并比较了两个最先进的LLM家族，即 ChatGPT 和 Llama 2-Chat[115]，它们已经通过 SFT 和 RLHF 从各自的基础模型（GPT 和 LLaMA）进行了微调，以遵循指令。

[158] Notellm: A retrievable largelanguage model for note recommendation. In Proceedings of the 33rd International Conference on World Wide Web,pages 123–132, 2024.
使用笔记压缩提示进行笔记推荐，将笔记压缩成一个特殊标记，并通过对比学习方法进一步学习潜在相关笔记的嵌入。

[174] Harnessing large language models fortext-rich sequential recommendation. In Proceedings of the 33rd International World Wide Web Conference, pages3207––3216, 2024.
构建一个提示文本，包括用户偏好总结、最近的用户交互和候选项目信息，用于基于LLM的推荐。然后使用监督微调技术对该系统进行微调，以生成最终的推荐模型。

[116] Llm4dsr:Leveraing large language model for denoising sequential recommendation. arXiv preprint arXiv:2408.08208, 2024.
提出了 LLM4DSR，这是一种使用LLM进行顺序推荐去噪的专门方法。它引入了一个自我监督微调任务，旨在增强LLM检测序列中噪声项目并建议适当替换的能力。

[74]Beyond inter-item relations: Dynamic adaptive mixture-of-experts forllm-based sequential recommendation. arXiv preprint arXiv:2408.07427, 2024.
提出了 MixRec 来增强基于LLM的顺序推荐。在粗粒度适应的基础上，MixRec 进一步通过上下文屏蔽、协作知识注入和动态专家混合（DAMoE）等技术进行了细化，使其能够有效地管理顺序推荐任务。

[7]Improving sequential recommendations with llms. arXiv preprint arXiv:2402.01339, 2024
设计并探索了在顺序推荐中利用LLM的三种正交方法和两种混合方法。具体而言，它深入研究了每种方法的技术方面，评估了潜在的替代方案，对它们进行了全面微调，并评估了它们的总体影响。

② Instruction Tuning

[71]A multi-facet paradigm to bridgelarge language model and recommendation. arXiv preprint arXiv:2310.06491, 2023.
提出了由数字 ID、项目标题和项目属性组成的多方面标识符。基于这些标识符，它在语言空间中构建用于大型语言模型微调的指令数据，有效地结合了不同方面的排名分数。

[153] Heterogeneous knowledge fusion: Anovel approach for personalized recommendation via llm. In Proceedings of the 17th ACM Conference on RecommenderSystems, page 599–601, 2023.
基于推荐任务和异构知识构建指令数据集，包括个性化推荐的输入、指令和输出，通过整合异构知识和推荐任务，对大型语言模型进行指令微调以实现个性化推荐。

[139] Exploring large language model for graphdata understanding in online job recommendations. In Proceedings of the 38th AAAI Conference on Artificial Intelligence,2024.
构建指令数据集，以弥合在线工作推荐中预训练知识与实际招聘领域之间的差距。

[49]Item-side fairness oflarge language model-based recommendation system. In Proceedings of the ACM Web Conference 2024, 2024.
首先将用户-项目交互数据用自然语言表示，并采用指令微调来微调大型语言模型，增强基于大型语言模型的推荐系统在项目方面的公平性。

[87]Integrating large language models into recommendation via mutual augmentation and adaptive aggregation.arXiv preprint arXiv:2401.13870, 2024.
通过利用指令微调的大型语言模型对传统推荐模型进行数据增强，以缓解数据稀疏和长尾问题。

[86] Recranker: Instruction tuning large language model as ranker for top-k recommendation. arXiv preprintarXiv:2312.16018, 2023.
采用自适应用户抽样来选择高质量用户，为构建指令微调数据集提供便利。然后，该数据集被用于训练指令微调的大型语言模型，以用于前 k 个推荐中的各种排名任务。

[62]E4srec: An elegant effective efficientextensible solution of large language models for sequential recommendation. In Proceedings of the 33rd InternationalWorld Wide Web Conference, pages 1234–1245, 2024.
为LLM设计指令微调过程，以激发其遵循指令的能力，从而提高其在顺序推荐中的能力。

[68]Llara: Largelanguage-recommendation assistant. arXiv preprint arXiv:2312.02445, 2023.
顺序推荐数据被转换为大型语言模型在推荐语料库上的指令微调格式。

[89]Unlocking the potential of large languagemodels for explainable recommendations. arXiv preprint arXiv:2312.15661, 2023.
主张采用名为 LLMXRec 的两阶段框架，将项目推荐与解释生成解耦。该研究利用指令微调来提高大型语言模型生成解释的精度和控制能力。

[155]Llamarec: Two-stagerecommendation using large language models for ranking. arXiv preprint arXiv:2311.02089, 2023.
专注于顺序推荐，并提出了一种用于基于大型语言模型的两阶段推荐的新型 LlamaRec 框架，证明了具有检索和排名的完整解决方案。

[52]Review-drivenpersonalized preference reasoning with large language models for recommendation. arXiv preprint arXiv:2408.06276,2024.
提出了基于LLM的新型推荐系统 Exp3rt，专门设计用于利用用户和项目评论中发现的广泛偏好信息。Exp3rt 有效地利用这一详细反馈来提高推荐的质量和个性化程度。

③ Low-Rank Adaptation (LoRA)

[79]Once: Boosting content-based recommendation withboth open- and closed-source large language models. In Proceedings of the 17th ACM International Conference on WebSearch and Data Mining, page 452–461, 2024.
研究了低秩适应（LoRA）对开源大型语言模型在基于内容推荐方面性能的有效性。

[30]Breakingthe length barrier: Llm-enhanced ctr prediction in long textual user behaviors. In Proceedings of the 47th InternationalACM SIGIR Conference on Research and Development in Information Retrieval, pages 2311–2315, 2024.
引入了行为聚合层次编码（BAHE）来提高基于LLM的点击率（CTR）建模效率。

[6]Tallrec: An effective and efficienttuning framework to align large language model with recommendation. In Proceedings of the 17th ACM Conference onRecommender Systems, pages 1007—-1014, 2023.
在保持原始参数冻结的情况下，通过优化秩分解矩阵有效地整合补充信息。

[174]Harnessing large language models fortext-rich sequential recommendation. In Proceedings of the 33rd International World Wide Web Conference, pages3207—3216, 2024.
将 LoRA 技术应用于参数高效微调（PEFT），解决文本丰富的顺序推荐问题。

[62]E4srec: An elegant effective efficientextensible solution of large language models for sequential recommendation. In Proceedings of the 33rd InternationalWorld Wide Web Conference, pages 1234–1245, 2024.
在其提出的门投影、下投影和上投影模块上额外引入了 LoRA 适配器，以对给定推荐任务进行个性化建模。

[144]Enhancing content-based recommendationvia large language model. arXiv preprint arXiv:2404.00236, 2024.
采用 LoRA 策略为每个领域训练一小部分参数，这些参数可以作为插件，无需进一步重新训练即可无缝添加到目标领域。

[78]Large languagemodel distilling medication recommendation model. arXiv preprint arXiv:2402.02803, 2024.

将 LoRA 微调用于药物推荐任务，更新低秩矩阵集，同时保持大型语言模型的预训练权重冻结。

[68]Llara: Largelanguage-recommendation assistant. arXiv preprint arXiv:2312.02445, 2023.
提出了课程提示微调策略来训练 LoRA，它可以利用顺序推荐器中封装的行为知识来增强大型语言模型。为了在保持高推荐性能的同时实现高效的遗忘。

[42]Exact and efficient unlearning forlarge language model-based recommendation. arXiv preprint arXiv:2404.10327, 2024.
使用 LoRA 以插件方式向大型语言模型的现有权重添加一对对秩分解权重矩阵，仅训练新添加的权重以进行学习任务。

[119]Towards efficientand effective unlearning of large language models for recommendation. arXiv preprint arXiv:2403.03536, 2024.
提出了 E2URec，这是基于大型语言模型的推荐系统的第一个高效且有效的遗忘方法。E2URec 通过仅更新有限的一组额外 LoRA 参数来提高遗忘效率，并通过师徒框架提高遗忘效果。

[169]Llm-based federated recommendation. arXiv preprint arXiv:2402.09959, 2024
纳入了动态平衡策略，涉及为每个客户端设计动态参数聚合和学习速度。

[85]Aligning largelanguage models for controllable recommendations. arXiv preprint arXiv:2403.05063, 2024.
为推荐系统中的大型语言模型引入了一种新颖的对齐方法，显著提高了其遵循用户指令的能力，同时最大限度地减少了格式错误。

[178]Lifelong personalized low-rank adaptation of large language models for recommendation. arXiv preprintarXiv:2408.03533, 2024.
提出了 RecLoRA，该模型包括一个个性化的 LoRA 模块，为不同用户维护独立的 LoRA，以及一个长短模态检索器，以检索不同模态的不同历史长度。

[152]. Harnessing multimodal large language models for multimodal sequential recommendation. arXiv preprintarXiv:2408.09698, 2024.
提出了多模态大型语言模型增强的多模态顺序推荐（MLLM-MSR）模型，旨在利用大型语言模型来改善多模态顺序推荐。

[64]Ganprompt: Enhancing robustness in llm-basedrecommendations with gan-enhanced diversity prompts. arXiv preprint arXiv:2408.09671, 2024.
提出了 GANPrompt，这是一个基于生成对抗网络（GANs）的多维大型语言模型提示多样性框架。该框架通过将 GAN 生成技术与大型语言模型的深层语义理解能力相结合，提高了模型对不同提示的适应性和稳定性。

[110]Delrec: Distilling sequential pattern to enhance llm-based recommendation. arXivpreprint arXiv:2406.11156, 2024.

提出了 DELRec，这是一个旨在从顺序推荐模型中提取知识并使大型语言模型能够轻松理解和利用这些补充信息以进行更有效的顺序推荐的新框架。它使用 AdaLoRA 来微调大型语言模型。

[166]Collm: Integrating collaborativeembeddings into large language models for recommendatio. arXiv preprint arXiv:2310.19488, 2023.
采用了两步调优程序：首先，仅使用语言信息以 LoRA 方式微调大型语言模型，以学习推荐任务，然后专门调优映射模块，使映射的协作信息对于大型语言模型的推荐易于理解和使用，同时在拟合推荐数据时考虑该信息。

非生成型推荐——LLM reusing

> **LLM Reusing的三种主要方法：DirectUtilizing、Prompt Tuning和 In-Context Learning。**

① DirectUtilizing

[150] Knowledge plugins: Enhancing largelanguage models for domain-specific recommendations. arXiv preprint arXiv:2311.10779, 2023.
提出了一种通用范式，即通过特定领域知识增强大型语言模型（DOKE），以提高其在实际应用中的性能。该范式依赖于领域知识提取器，包括为任务准备有效知识、为每个特定样本选择知识以及以大型语言模型可理解的方式表达知识。

[129] Drdt: Dynamic reflection withdivergent thinking for llm-based sequential recommendation. arXiv preprint arXiv:2312.11336, 2023
在检索-重排框架内引入了动态反思与发散思维（DRDT），这是一种新颖的方法，可利用大型语言模型将协作信号和用户偏好的时间演变有效整合到顺序推荐任务中。

[121]Zero-shot next-item recommendation using large pretrained language models. arXivpreprint arXiv:2304.03153, 2023.
提出了一种称为零样本下一项推荐（NIR）提示的提示策略，引导大型语言模型进行下一项推荐。具体而言，基于 NIR 的策略涉及使用外部模块根据用户过滤或物品过滤生成候选物品。[35]提出了一种利用生成式人工智能领域的最新进展进行多模态非平稳内容零样本推荐的方法。

[104]Large language models are competitivenear cold-start recommenders for language-and item-based preferences. In Proceedings of the 17th ACM conference onrecommender systems, pages 890–896, 2023.
针对基于语言的物品推荐任务，为大型语言模型提出了多种提示方法。

[94]Large language model augmented narrative drivenrecommendations. In Proceedings of the 17th ACM Conference on Recommender Systems, pages 777–783, 2023

提出了 Mint，这是一种用于叙事驱动推荐（NDR）任务的数据增强方法。Mint 通过使用 175B 参数大型语言模型在用户喜欢的物品文本上进行条件化，为 NDR 重新利用历史用户-物品交互数据集来撰写长篇叙事查询。

[109] Chatgpt for conversational recommendation: Refiningrecommendations by reprompting with feedback. arXiv preprint arXiv:2401.03605, 2024.
研究了 ChatGPT 作为 top-n 对话推荐系统的有效性。围绕 ChatGPT 构建了一个综合管道，以模拟真实用户交互，在探索模型以获取推荐时进行探测。

[156]Federated recommendation via hybrid retrieval augmentedgeneration. arXiv preprint arXiv:2403.04256, 2024.
提出了 GPTFedRec，这是一个利用 ChatGPT 和一种新颖的混合检索增强生成（RAG）机制的联邦推荐框架。

[132]Re2llm: Reflectivereinforcement large language model for session-based recommendation. arXiv preprint arXiv:2403.16427, 2024.
引入了一种超越上下文学习和大型语言模型微调的学习范式，有效地将通用大型语言模型与特定推荐任务相结合。

[128]RecMind: Large language model powered agent for recommendation. In Kevin Duh,Helena Gomez, and Steven Bethard, editors, Findings of the Association for Computational Linguistics: NAACL2024, pages 4351–4364, Mexico City, Mexico, June 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-naacl.271.
引入了 RecMind，这是一个用于一般推荐目的的大型语言模型驱动的代理。RecMind 无需微调即可适应不同的领域、数据集或任务。RecMind 采用了一种新颖的自我激励（SI）规划技术。

[38] Reindexthen-adapt: Improving large language models for conversational recommendation. arXiv preprint arXiv:2405.12119,2024.
提出了 ReindexThen-Adapt（RTA）框架，该框架将多标记物品标题转换为大型语言模型中的单标记物品标题，并随后调整这些单标记物品标题的概率分布。

[15]Improve temporal awarenessof llms for sequential recommendation. arXiv preprint arXiv:2405.02778, 2024.
提出了三种提示策略，以利用历史交互中的时间信息进行基于大型语言模型的顺序推荐。本研究将显式结构分析作为额外提示纳入输入序列，特别是时间聚类分析。

[163]Tired of plugins? large language models can be end-to-end recommenders. arXiv preprintarXiv:2404.00702, 2024
引入了一个基于大型语言模型的端到端推荐框架 UniLLMRec。UniLLMRec 通过推荐链方法集成了召回、排名和重排等多阶段任务。

[23]Where to move next: Zero-shot generalization ofllms for next poi recommendation. arXiv preprint arXiv:2404.01855, 2024.
专注于利用大型语言模型的能力进行零样本下一个兴趣点（POI）推荐任务。该方法考虑了用户的长期和当前偏好、地理空间距离以及用户移动行为中的顺序转换。

[39]Largelanguage models are zero-shot rankers for recommender systems. In European Conference on Information Retrieval,pages 364–381. Springer, 2024.
将推荐问题形式化为条件排名任务，将顺序交互历史用作条件，将其他模型检索到的物品用作候选。

[50] Interarec: Interactive recommendations using multimodal large languagemodels. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pages 32–43. Springer, 2024.
提出了 InteraRec，这是一种创新的基于截图的用户推荐系统。InteraRec 在用户浏览时捕获实时、高频的网页截图。利用多模态大型语言模型的能力，它分析这些截图以得出对用户行为的有意义见解，并使用相关优化工具提供个性化推荐。

[106]Largelanguage models are learnable planners for long-term recommendation. In Proceedings of the 47th International ACMSIGIR Conference on Research and Development in Information Retrieval, pages 1893–1903, 2024.
提出利用大型语言模型出色的规划能力来处理长期推荐中的稀疏数据。这项工作引入了一个双层可学习大型语言模型规划框架，该框架使用了一组大型语言模型实例。

[101]Logicscaffolding: Personalized aspect-instructed recommendation explanation generation using llms. In Proceedings of the17th ACM International Conference on Web Search and Data Mining, pages 1078–1081, 2024.
提出了一个名为逻辑支架的框架，它结合了基于方面的解释和思维链提示的概念，通过中间推理步骤生成解释。

[70]Rella: Retrieval-enhanced large language models for lifelong sequential behavior comprehension inrecommendation. In Proceedings of the ACM on Web Conference 2024, pages 3497–3508, 2024.
执行语义用户行为检索（SUBR）以提高零样本推荐测试样本的数据质量，显着降低了大型语言模型从用户行为序列中提取关键知识的难度。

[125] Rdrec: Rationale distillation for llm-based recommendation. arXiv preprint arXiv:2405.10587, 2024.
提出了一种紧凑的 RDRec 模型，用于学习大型语言模型生成的交互背后的基本原理。通过从所有相关评论中学习原理，RDRec 通过设计推荐的提示模板有效地指定了用户和物品档案。

[16]Llm-guidedmulti-view hypergraph learning for human-centric explainable recommendation. arXiv preprint arXiv:2401.08217,2024.
促进了基于大型语言模型的细致用户画像，同时仍然考虑了顺序用户行为。通过生成和完善引导超图学习过程的提示。

[44]Large language modelinteraction simulator for cold-start item recommendation. arXiv preprint arXiv:2402.09176, 2024.
提出了一种大型语言模型交互模拟器，用于根据内容方面模拟用户的行为模式。这种基于提示的模拟器允许推荐系统为每个冷物品模拟生动的交互，并直接将它们从冷物品转换为暖物

品。

[134]Llmrec: Large language models with graph augmentation for recommendation. In Proceedings of the 17th ACMInternational Conference on Web Search and Data Mining, pages 806–815, 2024.
使大型语言模型能够生成数据集原本不包含的用户和物品属性，方法是使用从数据集的交互和辅助信息中得出的提示。

[103]Lkpnr: Llm and kg for personalizednews recommendation framework. arXiv preprint arXiv:2308.12028, 2023.
将大型语言模型与知识图谱（KG）相结合。通过结合通用新闻编码器，大型语言模型强大的上下文理解能力能够生成富含语义信息的新闻表示。

[122]Llm4vis: Explainable visualization recommendation using chatgpt. arXiv preprint arXiv:2310.07652, 2023.
引入了 LLM4Vis，这是一种基于 ChatGPT 的创新提示方法，仅使用几个演示示例即可提供可视化推荐并生成类似人类的解释。

[127]Llmrg: Improving recommendations through large language model reasoning graphs. In Proceedingsof the AAAI Conference on Artificial Intelligence, volume 38, pages 19189–19196, 2024.
引入了 LLMRG，它使用大型语言模型构建个性化推理图。这种方法说明了大型语言模型如何在不需要额外信息的情况下增强推荐系统中的逻辑推理和可解释性。

[111] Large language models forintent-driven session recommendations. In Proceedings of the 47th International ACM SIGIR Conference on Research andDevelopment in Information Retrieval, SIGIR '24, page 324–334, New York, NY, USA, 2024. Association for ComputingMachinery. ISBN 9798400704314. doi: 10.1145/3626772.3657688. URL https://doi.org/10.1145/3626772.3657688.
引入了一种简单而强大的范式 PO4ISR，它利用大型语言模型的能力通过提示优化来增强信息检索（ISR）。

[76]Understanding before recommendation:Semantic aspect-aware review exploitation via large language models. arXiv preprint arXiv:2312.16275, 2023.
提出了一种基于链的提示方法，利用大型语言模型的深度语义理解来揭示语义方面感知的交互。这种方法在细粒度语义层面提供了对用户行为更详细的见解。

[135] Leveraginglarge language models (llms) to empower training-free dataset condensation for content-based recommendation.arXiv preprint arXiv:2310.09874, 2023.
提出了 TF-DCon 框架，灵感来自大型语言模型出色的文本理解和生成能力，并利用它们来增强压缩过程中的文本内容生成。

[136]Coral: Collaborative retrieval-augmented large language models improve long-tail recommendation. arXiv preprint arXiv:2403.06447,2024.

提出了 CoRAL，这是一种旨在增强传统基于协同过滤的系统中的长尾推荐的方法。它有效地解决了数据稀疏性和不平衡性带来的挑战，这些挑战往往限制了协同过滤方法的性能。

[148]Common senseenhanced knowledge-based recommendation with large language model. arXiv preprint arXiv:2403.18325, 2024.
提出了一种新颖的框架 CSRec，它开发了一个基于大型语言模型的常识知识图谱，并使用基于互信息最大化（MIM）的知识融合技术将其纳入推荐系统。

[172]Dynllm:When large language models meet dynamic graph recommendation. arXiv preprint arXiv:2405.07580, 2024.
引入了一项新任务，即使用连续时间动态图进行大型语言模型增强的动态推荐，并提出了 DynLLM 模型以有效地将大型语言模型增强的数据与时间图信息相结合。

[164]Finerec: Exploring fine-grainedsequential recommendation. In Proceedings of the 47th International ACM SIGIR Conference on Research and Developmentin Information Retrieval, pages 1599–1608, 2024.
引入了一个新颖的框架 FineRec，旨在通过从评论中挖掘属性意见来探索细粒度顺序推荐。

[98]Llm4sbr: A lightweightand effective framework for integrating large language models in session-based recommendation. arXiv preprintarXiv:2402.13840, 2024.
提出了一种可扩展的两阶段大型语言模型增强框架（LLM4SBR），专门为基于会话的推荐（SBR）而设计。研究了将大型语言模型与 SBR 模型集成的可行性，同时关注有效性和效率。在短序列数据的情况下，大型语言模型可以直接通过其语言理解能力推断偏好，甚至无需微调。

[170]Breaking the barrier: Utilizing large language modelsfor industrial recommendation systems through an inferential knowledge graph. arXiv preprint arXiv:2402.13750,2024.
专注于工业推荐系统，并提出了 LLM-KERec，它使用大型语言模型确定两个实体之间是否存在互补关系，并构建互补图。

[147]Sequential recommendation with latent relations based on large language model. In Proceedings of the 47th International ACM SIGIRConference on Research and Development in Information Retrieval, pages 335–344, 2024
提出了一种新颖的关系感知顺序推荐框架，具有潜在关系发现（LRD）。与依赖预定义规则的先前关系感知模型不同，它提议利用大型语言模型提供项目之间的新型关系和连接。

[145]News recommendation with category description by a large language model. arXivpreprint arXiv:2405.13007, 2024.
提出了一种新颖的方法，即使用大型语言模型自动生成信息丰富的类别描述，而无需人工努力或特定领域的知识，并将其作为补充信息集成到推荐模型中。

[66]Pap-rec: Personalized automatic prompt forrecommendation language model. arXiv preprint arXiv:2402.00284, 2024.

引入了 PAP-REC，这是一个旨在为推荐语言模型生成个性化自动提示的框架，解决了手动制作提示的低效和无效问题。

[120]Large language models as data augmentersfor cold-start item recommendation. In Companion Proceedings of the ACM on Web Conference 2024, pages 726–729,2024.

提出利用大型语言模型作为数据增强器来解决与训练期间冷启动项目相关的知识差距。通过利用大型语言模型，它根据用户历史行为的文本描述和新项目的描述推断冷启动项目的用户偏好。

[113]Mmrec: Llm based multi-modal recommender system.arXiv preprint arXiv:2408.04211, 2024.

研究了大型语言模型在推荐上下文中增强对自然语言数据的理解和利用的潜力。

[149] Darec: A disentangled alignment framework for large language model and recommendersystem. arXiv preprint arXiv:2408.08231, 2024.

提出了 DaRec，这是一种用于将推荐模型与大型语言模型集成的新型即插即用解耦对齐框架。

[91]X-reflect:Cross-reflection prompting for multimodal recommendation. arXiv preprint arXiv:2408.15172, 2024.

引入了 X-REFLECT，这是一个新颖的交叉反射提示框架。该方法提示大型多模态模型（LMMs）同时处理文本和视觉信息，明确识别和协调这些模态之间的任何支持或冲突元素。

[130]Llm4msr:An llm-enhanced paradigm for multi-scenario recommendation. arXiv preprint arXiv:2406.12529, 2024.

提出了 LLM4MSR，这是一种高效且可解释的大型语言模型增强范式。它使用大型语言模型通过定制提示提取场景相关性和跨场景用户兴趣等多层次知识，而无需微调。

[92]Llm-based aspect augmentations for recommendation systems. In Proceedings ofthe 40th International Conference on Machine Learning, Challenges in Deployable Generative AI Workshop, 2023.

旨在衡量使用大型语言模型为用户购买意图生成的项目方面-理由的有效性，以提高排名任务结果。为了实现这一目标，精心设计提示以从电子商务环境中的项目文本数据中得出项目方面。

[141]Towards open-world recommendation with knowledge augmentation from large language models. arXiv preprintarXiv:2306.10933, 2023.

引入了因式分解提示，以引出关于用户偏好的准确推理。生成的推理和事实知识通过混合专家适配器有效地转换和浓缩为增强向量，以与推荐任务兼容。

[90]Llm-rec: Personalized recommendation via prompting large language models. Findings of theAssociation for Computational Linguistics: NAACL 2024, pages 583–612, 2024.

引入了 LLM-Rec，这是一种结合了四种不同提示策略的新颖方法：基本提示、推荐驱动提示、参与引导提示以及推荐驱动和参与引导提示的组合。

② Prompt Tuning

[105] Pmg: Personalized multimodal generation withlarge language models. In Proceedings of the ACM on Web Conference 2024, pages 3833–3843, 2024.
专注于个性化多模态生成问题，并提出了 PMG。它将多模态标记作为可学习参数纳入嵌入表，然后利用线性层将大型语言模型的嵌入空间与生成器的嵌入空间对齐。

[176]Language-based user profiles for recommendation. arXiv preprintarXiv:2402.15623, 2024.
在指令（聊天）模式下对 Llama 2-7B、Llama 2-13B 和 Sakura-SOLAR 10.7B1 进行测试，使用零样本提示调整。使用验证数据集进行硬提示调整，以提高模型的准确性、运行时和可靠性。

[59]Prompt tuning large language models on personalized aspectextraction for recommendations. arXiv preprint arXiv:2306.01475, 2023.
提出了一个端到端框架，将通过提示调整学习基于大型语言模型的个性化方面提取与基于方面的推荐相结合，从而产生更有效的推荐。

[75]Recprompt: Aprompt tuning framework for news recommendation using large language models. arXiv preprint arXiv:2312.10463,2023.
引入了一个使用大型语言模型进行新闻推荐的提示调整框架，它独特地将提示优化器与迭代自举方法相结合，以改进基于大型语言模型的推荐策略。

③ In-Context Learning

[29] Chat-rec: Towards interactiveand explainable llms-augmented recommender system. arXiv preprint arXiv:2303.14524, 2023.
引入了一种名为 Chat-Rec 的新范式，通过将用户档案和历史交互转换为提示，创新地增强了大型语言模型在开发对话推荐系统方面的能力。Chat-Rec 已被证明能够通过上下文学习有效地学习用户偏好，并在用户和产品之间建立联系，从而使推荐过程更具交互性和可解释性。

[27]Dre: Generating recommendationexplanations by aligning large language models at data-level. arXiv preprint arXiv:2404.06311, 2024.
引入了数据级推荐解释（DRE），这是一个为黑盒推荐模型提供解释的非侵入式框架。这项工作提出利用大型语言模型的上下文学习和推理能力，使解释模块与推荐模块对齐。它采用上下文学习方法，并指导大型语言模型生成与推荐系统一致且与用户注意力偏好相对应的逻辑连贯的推荐解释。

[51]Efficient and responsible adaptation of large language models for robust top-krecommendations. arXiv preprint arXiv:2405.00824, 2024.
提出了一种混合任务分配框架，该框架利用大型语言模型和传统推荐系统的能力。通过采用两阶段方法来提高对亚群体的稳健性，该框架促进了任务的战略分配，以实现大型语言模型的高效和负责任的适应。

[17]Uncovering chatgpt's capabilities in recommender systems. In Proceedings of the 17th ACM Conference onRecommender Systems, pages 1126–1132, 2023

的目标是通过与传统信息检索（IR）排名方法（如点式、成对和列表式排名）对齐来增强 ChatGPT 的推荐能力。这项工作采用上下文学习和指令学习，并将不同的能力表示为具有特定领域提示的不同任务。

[11] New community cold-start recommendation: A novel large languagemodel-based method. Available at SSRN 4828316, 2024.

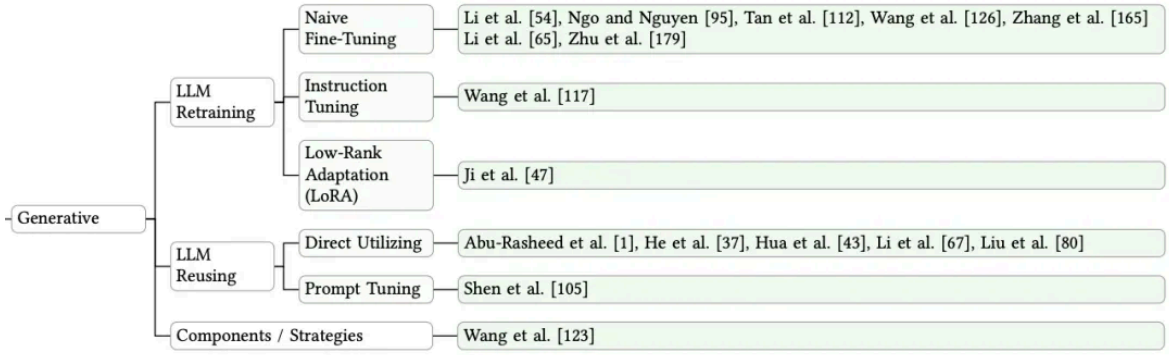
通过提出一种利用大型语言模型的广泛知识和强大推理能力的新推荐方法来解决新社区冷启动（NCCS）问题。它选择上下文学习（ICL）作为提示策略，并设计了一个粗到细的框架，以有效地为创建有效的 ICL 提示选择演示示例。

[168] Lane: Logic alignment of non-tuninglarge language models and online recommendation systems for explainable reason generation. arXiv preprintarXiv:2407.02833, 2024

提出了 LANE，这是一种将大型语言模型与在线推荐系统对齐的有效策略，而无需对大型语言模型进行额外的调整。这种方法降低了成本，同时提高了推荐的可解释性。这项工作利用了大型语言模型出色的上下文学习能力，并精心设计了一个零样本提示模板来提取用户的多种偏好。

生成型推荐

生成型推荐利用LLM通过将推荐任务转换为自然语言任务来执行推荐任务。这种方法允许生成式推荐，即系统直接生成要推荐的项目，而不是像传统推荐模型那样为每个候选项目计算排名得分。



生成型推荐——LLM retraining

① Naive Fine-Tuning

[54] Gpt4rec: A generativeframework for personalized recommendation and user interests interpretation. arXiv preprint arXiv:2304.03879, 2023.

引入了 GPT4Rec，这是一个新的通用生成框架。最初，它根据用户历史中的项目标题生成假设的"搜索查询"，然后通过搜索这些查询来检索推荐的项目。为了有效地捕捉用户在各个方面和细节层次的兴趣，同时提高相关性和多样性，该框架使用了一种多查询生成技术，即波束搜索。它微调了选定的 GPT-2[99]模型，该模型有 1.17 亿个参数，具有复杂的变压器架

构，并在庞大的语言语料库上进行了预训练。这个过程使我们能够有效地捕捉用户兴趣和项目内容信息。

[126]Llm-enhanced user-item interactions: Leveragingedge information for optimized recommendations. arXiv preprint arXiv:2402.09617, 2024.
提出了一种新的提示机制，可以将用户和项目之间的关系以及项目的背景信息转换为自然语言形式，并引入了一种新的融合方法来有效地利用图中的连接信息（即图结构中的边信息）。它使用人群上下文提示对图注意力大型语言模型（图注意力 GPT-2[99]模型）进行预训练，并使用个性化预测提示对图注意力大型语言模型进行微调。

[165] Recgpt: Generative personalizedprompts for sequential recommendation via chatgpt training paradigm. arXiv preprint arXiv:2404.08675, 2024.
专注于顺序推荐，并通过 ChatGPT 训练范式对用户行为序列进行个性化提示建模，提出了一个名为 RecGPT 的新框架。它通过引入用户 ID 模块对个性化自回归生成模型进行预训练，然后通过引入段 ID 对预训练模型进行微调，以生成个性化提示。

[95] Recgpt: Generative pre-training for text-based recommendation. arXiv preprintarXiv:2405.12715, 2024.
专注于基于文本的推荐，引入了名为 RecGPT-7B 的领域适应和完全训练的大型语言模型，以及其指令跟随变体 RecGPT-7B-Instruct。它使用相对较大的 205 亿个标记的特定推荐语料库对 RecGPT-7B 进行预训练，而 RecGPT-7B-Instruct 是通过在 10 万多个指令提示及其响应的数据集上进一步微调 RecGPT-7B 而得到的模型输出。

[112] Idgenrec: Llm-recsys alignmentwith textual id learning. In Proceedings of the 47th International ACM SIGIR Conference on Research and Developmentin Information Retrieval, pages 355–364, 2024.
提出了 IDGenRec，并选择了一个最初为文章标签生成而训练的 T5[100]模型，并使用推荐目标对其进行微调。它专注于标准顺序推荐任务和零样本推荐场景。

[179]Collaborative large language model for recommendersystems. In Proceedings of the ACM on Web Conference 2024, pages 3162–3172, 2024.
引入了 CLLM4Rec，这是一个紧密集成 ID 范式和大型语言模型范式的生成推荐系统。它提出了一种创新的软+硬提示策略，以有效地在表示历史交互和用户/项目特征的异构标记上对 CLLM4Rec 进行预训练。此外，它还提出了一种面向推荐的微调策略来预测保留项目。

[65]Calrec:Contrastive alignment of generative llms for sequential recommendation. arXiv preprint arXiv:2405.02429, 2024.
引入了 CALRec，这是一个为顺序推荐任务量身定制的对比学习辅助的两阶段训练框架，利用 PaLM-2 大型语言模型作为骨干。该框架结合了受少样本学习原则启发的精心制作的模板和独特的准循环 BM25 检索策略。

② Instruction Tuning

[117]Multiple key-value strategy inrecommendation systems incorporating large language model. arXiv preprint arXiv:2310.16409, 2023.
旨在通过将推荐系统与大型语言模型相结合，实现基于多键值数据的顺序推荐。特别是，它指导对一种流行的开源大型语言模型（LLaMA 7B[114]）进行调优，以便将推荐系统的领域

知识注入到预训练的大型语言模型中。由于这项工作使用了多键值策略，因此大型语言模型很难有效地从这些键中学习。因此，它设计了新颖的洗牌和掩码策略，作为一种创新的数据增强方法。

③ LoRA (Low-Rank Adaptation)

[47]Genrec: Largelanguage model for generative recommendation. In European Conference on Information Retrieval, pages 494–502.Springer, 2024.
强调了生成式推荐的有前途的范式，并提出了 GenRec 模型，该模型结合了文本信息，以提高生成式推荐的性能。它选择 LLaMA 语言模型作为骨干。LLaMA 模型在扩展的语言语料库上进行了预训练，为有效捕捉用户兴趣和项目内容信息提供了宝贵的资源。为了节省 GPU 内存，它采用 LLaMA-LoRA 架构进行微调推理任务。

① Direct Utilizing

[43]How to index item ids for recommendation foundationmodels. In Proceedings of the Annual International ACM SIGIR Conference on Research and Development in InformationRetrieval in the Asia Pacific Region, pages 195–204, 2023.
探讨了各种 ID 创建和索引方法，研究了三种基本索引方法：随机索引、标题索引和独立索引，同时强调了它们的局限性。该研究强调了为基础推荐模型选择适当索引方法的重要性，因为它对模型性能有重大影响。此外，还研究了四种简单而有效的索引方法：顺序索引、协作索引、语义索引和混合索引。

[1]Supporting student decisionson learning recommendations: An llm-based chatbot with knowledge graph contextualization for conversationalexplainability and mentoring. arXiv preprint arXiv:2401.08517, 2024.
提出了一种使用聊天机器人作为对话中介和控制有限生成解释的来源的方法。该方法旨在利用大型语言模型的能力，同时减轻其潜在风险。所提出的基于大型语言模型的聊天机器人旨在帮助学生理解学习路径推荐。它利用知识图谱（KG）作为人类策划的信息源，通过定义提示的上下文来调节大型语言模型的输出。此外，还实施了一种群聊方法，将学生与人类导师联系起来，无论是根据请求还是在情况超出聊天机器人的预定义能力时。

[37]Large language models as zero-shot conversational recommenders. In Proceedings of the32nd ACM international conference on information and knowledge management, pages 720–730, 2023.
探索了大型语言模型在零样本对话推荐系统中的应用。它引入了一种简单的提示策略，为大型语言模型定义任务描述、格式要求和对话上下文。然后，这项工作使用处理器将生成结果后处理为排名的项目列表。

[67]Bookgpt: A general framework for book recommendationempowered by large language model. Electronics, 12(22):4654, 2023.
以 ChatGPT 为建模对象，首次将大型语言模型技术融入到理解和推荐图书资源的标准场景中，并将其付诸实践。通过开发一个类似 ChatGPT 的图书推荐系统（BookGPT）框架，它

旨在将 ChatGPT 应用于三个关键任务的推荐建模：图书评级推荐、用户评级推荐和图书总结推荐。此外，它还探讨了大型语言模型技术在图书推荐背景下的可行性。它讨论了三个子任务的构建思路和提示工程方法。已经进行了实证研究来验证两种不同的提示建模方法的可行性：零样本建模和少样本建模。

[80]Once: Boosting content-based recommendation withboth open-and closed-source large language models. In Proceedings of the 17th ACM International Conference on WebSearch and Data Mining, pages 452–461, 2024.

使用提示技术在标记层面丰富了闭源大型语言模型的训练数据。这项工作引入了一种称为 GENRE 的生成推荐方法。通过开发多种提示策略，它增强了可用的训练数据，并获得了更具信息性的文本和用户特征，从而在后续的推荐任务中取得更好的性能。

② Prompt Tuning

[105] Pmg: Personalized multimodal generation withlarge language models. In Proceedings of the ACM on Web Conference 2024, pages 3833–3843, 2024.

关注个性化多模态生成问题，并提出了 PMG。它将多模态标记作为可学习参数并入嵌入表，然后使用线性层将 LLM 的嵌入空间与生成器的嵌入空间对齐。此外，它还使用 P-Tuning V2专门针对生成任务微调 LLM，从而增强其生成能力。在每次推理时，多模态标记都会被附加到用户行为提示中。然后，通过将这些增强的输入传递通过 LLM（增强了 P-Tuning V2）和线性层，生成软偏好嵌入。

生成型推荐——组件或策略

[123] Learnabletokenizer for llm-based generative recommendation. arXiv preprint arXiv:2405.07314, 2024.

专注于物品标记化问题，并全面分析了理想标识符的必要特征，提出了一种名为 LETTER 的新型可学习标记器，以自适应地学习包含层次语义、协作信号和代码分配多样性的标识符。

END

LLM与推荐 15

LLM与推荐 · 目录

上一篇
LLMs 推荐发展综述-Representing篇：单模态推荐 & 多模态推荐

下一篇
亚马逊COSMO：LLM构建高质量电商知识图谱