计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.010058] [CSTR: 32024.14.csa.010058]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

大语言模型用于推荐系统可解释性综述①

王晶鑫, 徐占洋, 于佳一, 卞刘尉, 李梦婷

(南京信息工程大学 软件学院, 南京 210044)
通信作者: 徐占洋, E-mail: 001280@nuist.edu.cn

摘　要: 随着大语言模型 (large language model, LLM) 的快速发展, 其在推荐系统可解释性方面的应用成为研究热点. 本文系统综述了 LLM 在推荐系统可解释性研究中的进展, 从领域研究现状、评价指标、数据集到应用场景进行了全面梳理. 从技术角度将现有研究分为基于 LLM 的推荐系统和 LLM 辅助型推荐系统, 并依据是否需要微调对此分类进一步细分. 在评价指标方面, 总结了人工评价与自动评价指标, 其中自动评价指标又包括传统指标、结合 LLM 指标以及拓展指标. 此外, 本文还整理了公开和私有数据集的使用情况, 强调了评论数据在可解释推荐中的重要性. 最后, 探讨了 LLM 在多个领域推荐系统可解释性方面的实际应用, 并分析了当前研究面临的挑战及未来可行的研究方向.

关键词: 大语言模型; 推荐系统; 可解释性; 评价指标; 数据集

引用格式: 王晶鑫,徐占洋,于佳一,卞刘尉,李梦婷.大语言模型用于推荐系统可解释性综述.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/10058.html

Survey on Large Language Model for Explainable Recommendation

WANG Jing-Xin, XU Zhan-Yang, YU Jia-Yi, BIAN Liu-Wei, LI Meng-Ting

(School of Software, Nanjing University of Information Science & Technology, Nanjing 210044, China)

Abstract: With the rapid development of large language models (LLMs), their application in the explainability of recommender systems has become a research hotspot. This study systematically reviews the research progress of LLMs in the explainability of recommender systems, providing a comprehensive overview covering current research status, evaluation metrics, datasets, and application scenarios. From a technical perspective, existing research is categorized into LLM-based recommender systems and LLM-aid recommender systems, further subdivided according to whether fine-tuning is required. In terms of evaluation metrics, manual evaluation and automated evaluation metrics are summarized, with automated evaluation metrics including traditional metrics, LLM-integrated metrics, and extended metrics. Moreover, the usage of public and private datasets is reviewed, with emphasis on the importance of review data in explainable recommendations. Finally, practical applications of LLMs in the explainability of recommender systems across various domains are explored, and the challenges faced by current research as well as potential future research directions are analyzed.

Key words: large language model (LLM); recommender system; explainability; evaluation metrics; dataset

随着深度学习模型的广泛应用, 可解释性已成为如今的研究热点[1–3]. 目前, 可解释性研究主要集中在医疗[4]、金融[5]等高风险的决策领域, 并广泛采用 LIME[6]、SHAP[7]等事后解释方法. 与此同时, 推荐系统也受到深度学习的影响, 神经网络模型的应用使得推荐精度在不断提升的同时, 其“黑箱”特性却严重限

① 收稿时间: 2025-07-02; 修改时间: 2025-08-01; 采用时间: 2025-08-19; csa 在线出版时间: 2025-12-19

制用户对推荐结果的信任与采纳. 因此, 如何平衡推荐精度与可解释性已成为当前推荐系统领域亟待解决的关键问题[8].

在推荐系统中, 解释形式众多. 例如 Zhang 等人[9]就根据解释的信息来源及展示形式将解释分为协同过滤驱动的解释、基于物品特征的解释、视觉化解释以及社交关系衍生的解释等. 其中, 图像相关的可视化解释[10] (如节点链接图、条形图等) 已有较多研究, 但这种解释形式计算复杂度高, 难以满足实时性要求严格的推荐场景. 与之相比, 文本解释因其灵活性和普适性, 仍然是当前研究的主流.

高质量文本解释不仅能帮助用户理解推荐逻辑, 还能显著提升用户对推荐结果的满意度和信任度. 例如, Feng 等人[11]发现, 基于用户历史行为的上下文解释比通用解释更能提高电影推荐的接受度. Lu 等人[12]的研究则进一步证明人类撰写的解释在用户意图理解和偏好构建方面都优于系统生成的模板解释, 因此人类解释一直是推荐系统的理想选择, 但由于其成本高、获取周期长, 在实际应用中并未得到广泛采用.

近年来, 大语言模型 (large language model, LLM) 的崛起为推荐解释带来了新的机遇. LLM 凭借其强大的语言生成能力, 有望达到与人类解释相媲美的水平. 在这一背景下, Silva 等人[13]通过设计推荐和解释提示词, 利用 LLM 生成了个性化的推荐解释. Petruzzelli 等人[14]则首次将 LLM 应用于跨域推荐, 验证了其在知识迁移与解释生成方面的潜力. 此外, LLM 还具备生成多种类型解释的能力. 例如 Lubos 等人[15]利用 LLM 生成了基于后果的通用解释, 这类解释从物品依赖性出发, 拓展了解释的适用场景. Okoso 等人[16]的研究则表明, 解释的语气、领域适配性以及用户个性化需求均会显著影响用户对推荐结果的感知和接受程度. 基于这一发现, 他们设计了针对语气等因素的提示词, 通过 LLM 生成的推荐解释显著提高了用户满意度. 这些研究充分展现了 LLM 在生成高质量推荐解释方面的强大能力.

然而, 现有的综述主要聚焦于 LLM 用于推荐性能的优化, 对 LLM 如何增强推荐可解释性这一关键问题的系统性总结尚显不足. 尽管部分研究在探讨模型架构或性能评估时也会涉及可解释性议题, 但该领域至今缺乏专门的全面综述. 据我们所知, Said 等人[17]在 2025 年发表的综述是首个探讨 LLM 用于推荐可解释性方面的综述, 但其仅基于当时筛选的 6 篇文献展开讨论, 研究深度和系统性总结仍有不足. 经过近一段时间的快速发展, 本领域已涌现出诸多优秀框架和模型, 亟需新的综述工作对这些进展进行系统梳理. 我们以此为动机, 希望通过全面梳理最新研究成果, 为该领域研究者提供清晰的 LLM 应用于推荐可解释的实现路径, 以促进本领域的发展.

本综述第 1 节首先提出针对现有研究的分类体系, 进而对该领域的文献进行系统梳理和归类, 并对比不同技术路线的特点, 以帮助研究人员选择合适的研究路线. 第 2 节按类别归纳可解释推荐的评价指标, 并分析各指标存在的局限性. 第 3 节介绍可解释推荐系统常见的数据集, 重点分析评论数据的作用以及使用方式. 第 4 节分领域讨论 LLM 在可解释推荐中的实际应用场景, 突出领域差异对解释需求的影响, 以及由此对解释方法产生的约束. 第 5 节指出该领域当前的挑战和未来方向. 第 6 节总结全文.

1 研究现状

1.1 分类体系

现如今, LLM 在推荐系统中的应用日益广泛, 但现有研究不管在技术路线, 还是在应用方式上都呈现出显著的多样性. 为梳理这一领域的研究进展, 本综述提出一种二维分类体系, 该分类体系从以下两个维度对 LLM 在推荐系统中的应用方法进行分类.

第 1 个分类维度为 LLM 在推荐系统中所担任的功能角色, 根据此维度可以将相关研究分为两大类, 即基于 LLM 的推荐系统 (LLM-based) 和 LLM 辅助型推荐系统 (LLM-aid). 其中, LLM-based 完全依赖 LLM 来完成推荐任务, 并不使用传统的推荐模型, 整个系统由 LLM 端到端驱动. 而 LLM-aid 则将 LLM 作为辅助工具, 用于增强推荐系统的可解释性或其他特定功能, 在整个推荐系统中, 仍然保留传统推荐模型作为核心推荐引擎. 其中, LLM-aid 又可根据推荐模型与解释模型 (这里指的是 LLM, 下同) 的耦合程度进一步细分为 3 种类型: 推荐模型无关 (model-agnostic)、推荐模型相关 (model-specific) 和推荐模型半相关 (semi-specific). model-agnostic 类型将推荐模型和解释模型完全解耦, 使得推荐模型可以独立替换而不影响解释模型; 而在 model-specific 类型中, 推荐与解释模型紧密结合, 两者相互依赖, 无法单独修改; semi-specific 方法则介

于前两者之间, 虽然推荐模型仍可替换, 但需部分调整推荐模型结构以适配解释模型.

第 2 个分类维度则是 LLM 是否需要微调, 据此维度可将相关研究再分为需要微调 (finetune) 和无需微调 (no finetune) 两种类别. 前者需要为解释任务对 LLM 进行参数微调优化, 而后者直接利用预训练 LLM 的世界知识和推理能力, 通过提示工程等方法实现推荐的可解释. 研究者可根据自身任务需求和计算资源, 基于此分类合理选择 LLM 的应用策略.

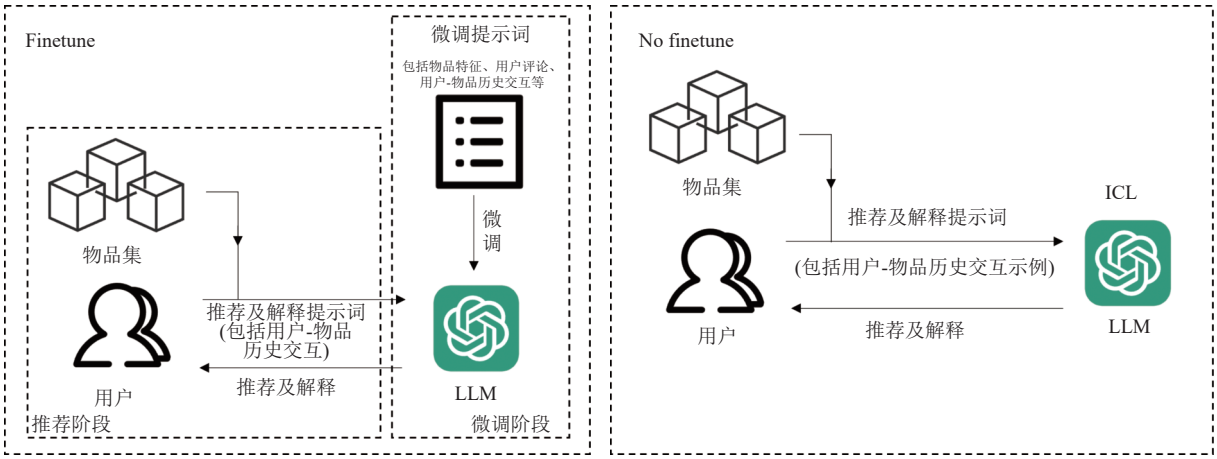
在上述分类体系的基础上, 我们收集了该领域近 3 年的文献, 并对文献进行了整理与分类, 分类结果如表 1 所示. 接下来我们将通过列举相关研究工作的方式介绍各类技术方法的具体实现细节.

1.2 基于 LLM 的推荐系统

基于 LLM 的推荐系统结构如图 1 所示, 图 1(a) 为参数微调场景, 图 1(b) 为非参数微调 (即样本提示) 场景.

表 1 LLM 应用于推荐系统可解释方式

LLM使用方式	是否需要微调	推荐模型相关性	模型/现有工作
LLM-based	finetune	—	LLM4CDR[14]、ECCR-LLM[18]、CIER[19]、Pleaser[20]、Exp3rt[21]、HDRec[22]
	no finetune	—	Zeng等人[23]、Chat-REC[24]、ELMAR[25]
LLM-aid	finetune	model-specific	LLM2ER-ERQ[26]、InteRecAgent[27]、MuseChat[28]
		model-agnostic	XRec[29]、RecExplainer[30]、G-Refer[31]
	no finetune	model-specific	LR-Recsys[32]、CrossDR-Gen[33]
		model-agnostic	LLME4RS[34]、DRE[35]、Lin等人[36]、Logic-Scaffolding[37]、Peng等人[38]、Abu-Rasheed等人[39]、PRAG[40]、Kovacs等人[41]、Ashaduzzaman等人[42]、Li等人[43]、Chun等人[44]
		semi-specific	LANE[45]



(a) 参数微调的解释生成 (b) 样本提示的解释生成

图 1 基于 LLM 的推荐系统

1.2.1 参数微调的解释生成

如图 1(a) 所示, 基于 LLM 微调方法在推荐系统中的研究主要聚焦于如何通过指令微调、参数高效微调等技术提升推荐性能和可解释性. 这些工作通常包含微调和推荐两个阶段: 在微调阶段利用物品特征、用户评论、用户-物品历史交互等数据对 LLM 进行微调训练; 而在推荐阶段则通过设计推荐及解释提示词让 LLM 生成推荐结果及其解释, 以下所列举的相关研究就采用了该方法来生成推荐及解释.

LLM4CDR[14]使用 LLM 处理跨域推荐问题, 其方法是将数据集划分为指令调优集和提示测试集, 构建用户源域喜好项、厌恶项和目标域待排序项, 并提取物品特征. 通过最小化交叉熵损失进行指令调优, 使 LLM 适应跨域推荐 (CDR) 任务. 在进行推荐时, 过滤幻觉物品, 确保推荐项在目标域候选集中, 最终在得到推荐结果的同时, 形成自然语言推荐解释.

ECCR-LLM[18]在使用 LLM 时, 结合 QLoRA 微调方法解决互补推荐中的连贯性问题. 该框架引入多任务学习, 除基本的推荐和解释生成任务外, 还包含互补分类和替代分类任务, 以此提升生成解释的质量.

CIER[19]基于 LLaMA 构建推荐系统, 采用 LoRA 微调解决推荐评分与解释文本不一致的问题. 其核心在于预测用户评分时, 将离散的硬评分转化为连续的软评分向量, 从而生成与预测评分语义高度一致的解释文本.

Pleaser[20]提出基于 T5 架构的编码器-解码器方案进行推荐及解释. 编码器捕获序列依赖关系进行推荐, 解码器则结合物品描述和用户偏好生成解释. 该工作采用参数高效微调 (PEFT) 降低成本, 并设计个性化偏好提取模块, 使解释同时反映物品特征和用户偏好.

Exp3rt[21]通过利用用户和物品评论中的偏好信息进行个性化推理, 以提高评分预测的准确性和推荐的可解释性. 具体地, 该模型首先通过 LLM 预处理评论数据, 以此来提取用户偏好和物品资料, 随后采用监督微调 (SFT) 进行微调训练, 最终通过分步文本推理预测用户评分, 并生成可解释的推理过程.

HDRec[22]提出了一种基于分层蒸馏的推荐方法, 旨在从用户评论中提取交互逻辑 (包括用户偏好、个性特征和商品属性等), 并生成具有解释性的摘要. 该方法采用 LLM 分阶段蒸馏关键信息, 并通过微调 T5-small 来优化推荐任务. 系统通过生成详细的真实解释和评论摘要来增强推荐结果的可解释性.

1.2.2 样本提示的解释生成

在现有研究中, 由于微调成本较高等因素, 基于 LLM 未微调的工作占据了主要部分. 这类范式的基本方法是通过设计提示词来完成推荐及解释任务, 其中提示词通常包含用户-物品历史交互数据和候选物品集合. 相较于微调方法依赖大量领域训练语料使 LLM 掌握相关知识, 无微调方法下的 LLM 会缺乏特定领域的推荐知识. 为此, 该范式下的研究人员通常会采用上下文学习 (ICL) 方法, 通过在提示词中嵌入相关示例来使 LLM 掌握推荐领域的相关知识, 以此让 LLM 能更好地理解任务需求, 从而输出令用户满意的推荐及解释. 这种方法生成推荐及解释的逻辑如图 1(b) 所示. 当前, 多项研究实践了该方法并取得良好进展.

例如, Zeng 等人[23]提出了一种利用 LLM 的零样本医生推荐框架, 通过整合医生的多维专业信息 (如专长、研究、奖项等) 和疾病-治疗配对数据, 实现了高准确性和可解释性的医生排名. 特别地, 此框架在生成针对特定疾病—治疗对的排名标准时, 采用了一次提示法, 即提供一个最佳标准作为示例, 例如胃癌手术治疗的评估标准示例. Chat-REC[24]则采用更直接的思路, 它将用户画像和历史交互数据转换为提示词, 以此构建完整的对话式推荐系统. ELMAR[25]专注于金融场景下的养老基金推荐, 该模型借助 LLM 分析人口统计数据以寻找相似用户, 进而生成推荐. 此方法可支持冷启动场景, 并能在推荐时给出解释.

第 1.2 节中的研究均直接将 LLM 作为推荐器使用, 在生成推荐的同时提供解释. 虽然这种方法具有便捷性, 但通常在用户或物品资料缺失的推荐中存在推荐精度不足的问题. 后续关于 LLM 辅助型推荐系统的研究突破了这一范式, 研究人员开始探索将 LLM 与传统推荐模型相结合来增强系统可解释性的新思路.

1.3 带微调的 LLM 辅助型推荐系统

带微调的 LLM 辅助型推荐系统如图 2 所示. 图 2(a) 对应推荐模型相关范式, 图 2(b) 则是推荐模型无关范式.

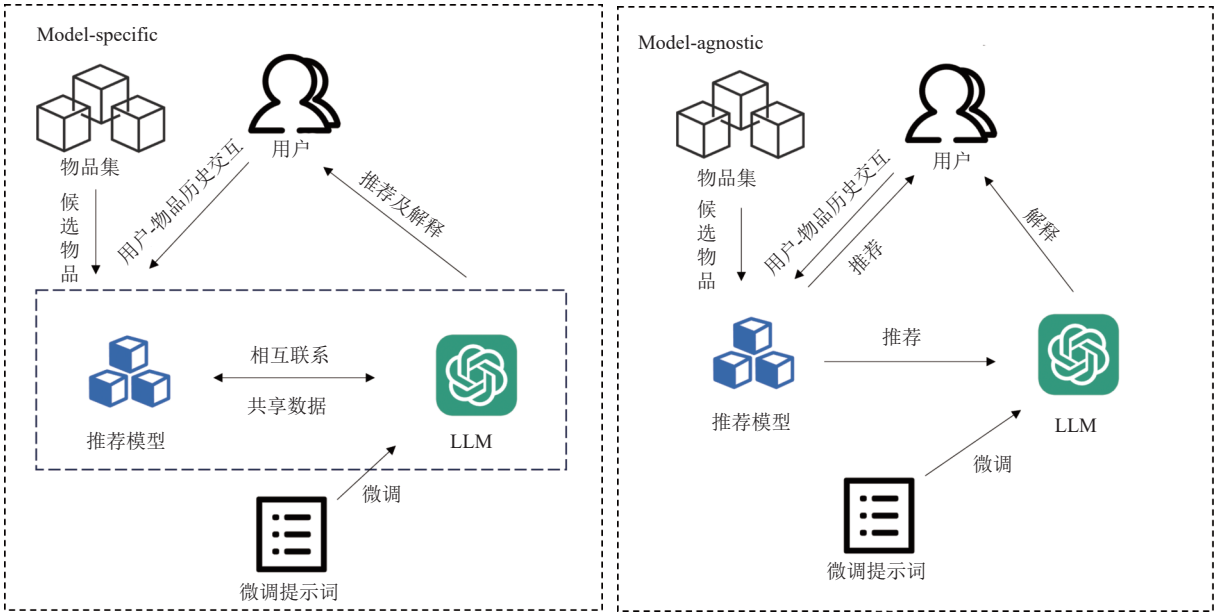
1.3.1 推荐模型相关的解释生成

如图 2(a) 所示, 在推荐模型相关范式中, 推荐模型和 LLM 相互联系不可分割, 属于一个整体, 他们共享数据. 其中, 推荐模型接收用户-物品历史交互和候选物品输入后, 它会根据具体的推荐算法形成用户对于物品的预测分数等数据, 然后 LLM 再利用这些数据形成推荐结果并附带解释. LLM2ER-EQR[26]是采用这种方法的典型框架, 我们将详细介绍这个框架, 除此之外, 还会简单介绍这类范式下的其他工作.

LLM2ER 框架基于概念图, 包含 3 个核心模块: 评分预测模块、个性化提示学习模块以及解释生成模块[26]. 其中, 评分预测模块利用异构图神经网络 (HGT) 学习用户和物品嵌入, 并通过多层感知机 (MLP) 预测评分; 个性化提示学习模块接收到评分预测模块的用户和物品嵌入后, 结合任务描述、用户情感和候选概念, 构建个性化提示; 解释生成模块则将该提示输入因果语言模型 (causal LM), 最终生成自然语言形式的推荐解释.

为了提高解释文本的质量, 文献[26]对 LLM2ER 进行强化范式微调, 形成 LLM2ER 的升级版——LLM2EQ-EQR. 这种微调包括两种奖励模型, 第 1 种模型是概念一致性奖励模型 (CCR), 在使用该模型之前先需要使用对比训练来提升解释与用户偏好、物品特征的一致性, 然后通过 BERT 计算生成解释与候选概念的余弦相似度作为奖励, 以此来微调 LLM.

第 2 种模型是高质量对齐奖励模型 (HQAR), 它采用了生成对抗网络 (GAN) 结构: 以 LLM2ER 作为生成器, 并用与 LLM2ER 同结构的 HQAR 作为判别器, 通过对抗训练使得模型能够生成更加高质量的解释. 最终 EQR 通过叠加上述两种奖励模型, 显著提高了推荐解释的质量.



(a) 推荐模型相关的解释生成 (b) 推荐模型无关的解释生成

图 2 带微调的 LLM 辅助型推荐系统

InteRecAgent[27]将 LLM 作为决策中枢, 与传统推荐模型协同工作. 通过指令微调使 LLM 能够处理推荐模型提供的查询、检索和排序结果, 最终生成推荐并给出解释. MuseChat[28]是一个基于对话的视频音乐推荐系统, 它首先利用对比学习让推荐模型学习一个多模态共享向量空间, 然后从音乐库中检索出与当前视频、用户偏好相匹配的目标音乐作为推荐候选集, 再然后利用 LLM 综合考虑用户-物品历史交互记录、匹配分数等数据形成推荐结果. 在此过程中, LLM 除了产生推荐, 它还会通过 LoRA 微调来生成高质量自然语言形式的推荐解释.

1.3.2 推荐模型无关的解释生成

大部分研究人员认为, 为了实现推荐模型的动态替换, 推荐模型应与解释模型分离. 因此, LLM 辅助微调方法中有较多与推荐模型无关的研究. 这种范式不再将推荐模型和 LLM 视为一个整体, 而是把它们分开后再单独研究. 如图 2(b) 所示, 推荐模型接收用户-物品历史交互数据和候选物品输入后, 形成的推荐结果会直接返回给用户. 与此同时, 推荐模型也会将推荐结果送入解释模型, 解释模型再根据微调阶段学到的推荐领域知识为用户生成推荐解释. 为深入理解这一范式下的技术方法, 我们整理了以下几项具有代表性的工作.

XRec[29]设计了一个混合专家系统 (MoE) 适配器, 用于连接推荐模型与 LLM. 该适配器将推荐模型中的协同信号与 LLM 的文本语义进行了对齐. 对齐后, 用户和物品的协同嵌入作为特殊词元 (token) 输入 LLM. 通过此步骤, LLM 就能有效理解非语义内容. 为了实现高效微调并生成高质量的推荐解释, 该框架在微调训练过程中冻结 LLM 主体部分, 仅更新少量新增参数. 同时, XRec 通过利用外部 LLM 处理用户资料、物品信息以及评分数据, 将原始主观评论转化为真实解释文本, 然后将这些数据作为监督信号, 指导模型在训练过程中优化新增参数, 从而提升生成解释的准确性.

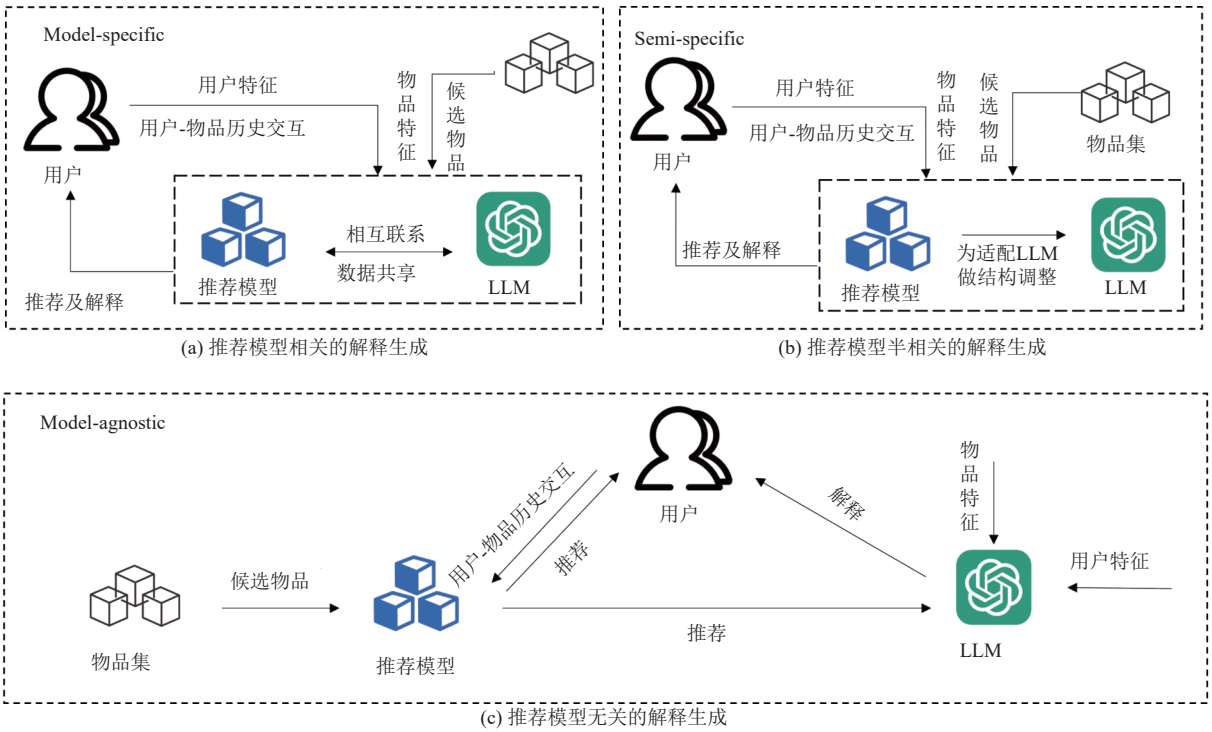
RecExplainer[30]则将 LLM 作为代理模型, 通过模仿推荐模型的行为来学习并复现其推荐能力, 然后再利用 LLM 强大的推理能力生成推荐解释. 该框架提供 3 种方法对齐推荐模型行为, 即行为对齐、意图对齐和混合对齐. 其中, 行为对齐通过将用户和物品资料转化为文本, 然后在语义空间中对 LLM 进行指令微

调. 意图对齐则是将用户和物品嵌入视为一种新的模态, 通过将其加入微调数据来在潜在空间训练 LLM. 作者针对这种对齐方法设计了历史重建任务, 以增强嵌入信息保真度. 混合对齐综合以上两种对齐方法在语义空间和潜在空间的优势, 通过在指令微调的提示词中同时包含文本及嵌入信息, 提高 LLM 对推荐模型行为的理解. 作者发现, 这 3 种方式中通过混合对齐方法微调的 LLM 生成的推荐解释质量最好, 这证明其方法的有效性.

G-Refer[31]利用路径级检索器和节点级检索器显式捕获基于图神经网络 (GNN) 的推荐模型的协同过滤 (CF) 信号. 具体地, 在使用路径级检索器获取结构 CF 信号时训练 GNN 模型, 通过掩码学习和 Dijkstra 算法检索解释路径. 节点级检索器则通过基于双编码器的密集检索架构计算用户和物品资料语义相似度来检索最相关节点以获取语义 CF 信号. 后续将检索到的结构和语义 CF 信号通过图翻译转换成自然语言文本以便 LLM 处理. 值得注意的是, 在使用 LoRA 对 LLM 进行微调之前, 作者首先采用基于知识剪裁的方法筛选训练样本. 具体地, 通过计算用户和物品资料与真实解释之间的语义相似度, 过滤掉那些对 CF 信号依赖程度较低的样本, 以降低噪声并提升训练效率. 随后, 作者将用户和物品资料以及通过检索增强获得的 CF 信号输入 LLM 进行微调, 最终生成高质量的解释文本.

1.4 无需微调的 LLM 辅助型推荐系统

与基于 LLM 的推荐系统类似, LLM 辅助型推荐系统中也有较多无需微调的工作. 此部分工作为了弥补不微调带来的解释偏差问题, 常使用的方法是引入用户和物品特征数据. 不带微调的 LLM 辅助型推荐系统结构如图 3 所示. 图 3(a)、图 3(b) 分别展示了推荐模型相关和推荐模型半相关范式, 图 3(c) 则展示了推荐模型无关范式.



(a) 推荐模型相关的解释生成

(b) 推荐模型半相关的解释生成

(c) 推荐模型无关的解释生成

图 3 不带微调的 LLM 辅助型推荐系统

1.4.1 推荐模型相关的解释生成

与需要微调的推荐模型相关 LLM 辅助方法类似, 其无需微调的版本省略了微调步骤, 并加入了用户和物品特征数据, 其结构如图 3(a) 所示.

LR-Recsys[32]实践了这种范式方法, 该框架将 LLM 与深度神经网络 (DNN) 相结合, 首先将用户和物品相关信息经 LLM 推理生成两种自然语言解释 (正向解释和负向解释), 然后将解释通过文本编码器转换成嵌入后, 最后将该嵌入进行特征拼接输入用于推荐的 DNN 进行训练, 从而提升推荐精度. 在目前收集到的该领域

工作中, 只有这项研究将 LLM 生成的解释应用于提升推荐模型的精度上, 因此它有较大的研究价值. 在 CrossDR-Gen[33]模型中 LLM 也起到关键的作用, 一方面, 它从用户访问的兴趣点 (POI) 类别挖掘信息, 生成伪用户画像, 帮助序列推荐模型更好地理解用户偏好, 以此提升推荐性能. 另一方面, 它可基于推荐结果与 POI 相关信息, 生成推荐解释.

1.4.2 推荐模型无关的解释生成

无需微调的 LLM 辅助方法中推荐模型无关范式, 因为其简单性, 所以在我们收集到的文献中占绝大部分, 其工作流程如图 3(c) 所示. 这部分的模型或工作众多, 我们将介绍部分, 如下.

LLME4RS[34]探讨利用 LLM 为推荐系统生成高质量解释方法. 该工作研究了 3 种推荐方式对应的解释生成, 分别为基于特征的推荐、基于物品协同过滤的推荐以及基于知识的推荐. 其解释生成方法首先构建包含用户偏好和物品特征的提示模板, 然后由 LLM 生成自然语言解释.

DRE[35]将用户-物品历史交互数据和推荐模型预测的推荐物品作为输入, 让 LLM 总结用户偏好和推荐物品属性之间的关系. 然后利用用户交互过的物品的评论数据和推荐物品的评论进行目标感知用户偏好蒸馏. 具体而言, LLM 先从推荐物品描述及评论中提取关键特征, 形成目标物品概要. 然后再结合目标物品概要, 利用 LLM 从用户交互过的物品评论中筛选与目标物品相似特征, 从而生成具有目标感知的历史物品概要. 最后, 将目标物品概要和用户历史物品概要输入到 LLM, LLM 凭借其强大的上下文学习能力生成逻辑连贯的推荐解释.

Lin 等人[36]训练联合框架, 通过评分预测模块学习用户物品表征, 并输出最终推荐分数, 然后将评分预测模块中得到的用户或物品嵌入共同作为预测提示输入 GPT-2, 引导其生成解释文本. Logic-Scaffolding[37]则是通过计算推荐物品嵌入与用户历史中每个物品嵌入点积, 选择得分最高的前 k 个物品, 然后利用少样本学习通过 LLM 进行方面 (aspect) 提取. 最后, 采用思维链提示技术基于物品属性和提取的方面特征生成推荐解释. 类似地, Peng 等人[38]将用户和物品 ID 向量作为连续提示, 并采用多任务学习框架联合优化推荐任务和解释生成任务.

属于此类工作的还有 Abu-Rasheed 等人[39]、Xie 等人[40]、Kovacs 等人[41]、Ashaduzzaman 等人[42]、Li 等人[43]、Chun 等人[44]的几篇论文.

1.4.3 推荐模型半相关的解释生成

特别地, 在无需微调的 LLM 辅助方法中有一类范式属于推荐模型半相关结构. 这类范式的推荐模型原则上可以替换, 但是替换相对来说比较麻烦, 需要修改推荐模型部分结构才能实现, 其工作流程如图 3(b) 所示.

LANE[45]框架属于该范式, 它首先利用文本编码器 (TextEncoder) 对物品集中的标题进行编码, 生成语义嵌入矩阵. 同时, 将用户的历史交互序列转换成固定长度的向量序列. 这两部分内容随后被输入到一个改进的 SASRec 模型 (修改了其嵌入层和预测层) 中. SASRec 经过处理, 输出用于计算排名的分数特征向量, 从而形成查询向量 (Query). 接着, 利用 LLM 预设定好的用户偏好提示模板从用户历史交互序列中提取用户编码再通过 TextEncoder 转换为嵌入向量, 然后使用 Transformer (包含多头注意力机制和前馈神经网络, 并结合残差连接和层归一化) 将用户序列特征向量与多个用户偏好嵌入向量进行语义对齐. 对齐后的特征经过预测模块处理, 生成最终的推荐排序结果. 以上步骤是此框架生成推荐的逻辑, 除此之外, 它还包含可解释推荐文本生成模块, 其基于零样本提示模板, 融合用户交互序列、多偏好信息及注意力权重等数据, 最终生成个性化推荐解释.

综上, 第 1.3 和 1.4 节分别总结了 LLM 辅助推荐系统带微调和不带微调两类工作. 带微调方法分为推荐模型相关和无关范式, 其中模型相关深度融合推荐模型与 LLM 以协同生成解释, 模型无关方式则用适配器等技术生成高质量解释. 不带微调依赖用户和物品特征等数据生成解释, 分为推荐模型相关、无关及半相关范式, 这些范式效率高但推荐精度依赖推荐模型. 第 1.5 节我们将通过回答一些问题来对比第 1.2–1.4 节中各方法优劣, 以供领域研究人员选择适合自己的研究路线.

1.5 技术路线对比

为方便领域研究人员根据自身条件选择合适的技术路线以解决遇到的推荐系统可解释性问题. 我们在此节整理了以上分类体系中主要技术方法的量化对比. 具体地, 从 3 个方面进行对比: 推荐性能、解释质量以及计算开销与生成效率. 在推荐性能方面, 对于 Top-K 推荐任务, 使用精准度 (Precision)、召回率 (Recall)、

归一化折损累计增益 (NDCG)、命中率 (HR) 以及准确率 (ACC) 作为评估指标, 这 5 个指标均是越高越好; 而对于评分预测任务, 则使用平均绝对误差 (MAE) 和均方根误差 (RMSE) 进行衡量, 这 2 个指标是越低越好. 在解释质量方面, 采用传统指标 (BLEU、ROUGE、Distinct)、结合 LLM 指标 (Rat) 以及拓展指标 (ASP) 进行对比分析. 最后, 在计算开销与生成效率方面, 整理了微调方式和非微调方式的计算开销, 包括硬件配置和使用的 LLM 基准模型 (包括类型, 参数量), 另外若相关论文提及解释生成时间、延迟等信息, 我们也会一同整理. 需要说明的是, 除非特别注明, 本节所引用的数据均来自各技术方法论文的原始实验数据.

1.5.1 推荐性能对比

在整理文献时, 我们发现 LLM-based 范式中所涉及的所有研究均对 LLM 的推荐性能进行了讨论, 而 LLM-aid 范式对于推荐能力的讨论却较少, 特别是推荐模型无关方法. 从这可以看出用 LLM 做推荐可解释研究, 推荐能力并不是我们研究的重点, 但推荐系统的核心功能仍然是推荐, 所以我们在此节对分类中涉及的推荐性能进行对比. 在对比的过程中, 我们将回答以下两个推荐性能方面的关键问题.

问题 1: LLM-based 和 LLM-aid 范式哪个推荐性能更优?

问题 2: 推荐模型无关方法是否在推荐精度上低于推荐模型相关方法?

LLM-based 范式依赖 LLM 做推荐, LLM-aid 则依靠推荐模型做推荐, 两者做推荐的方式有明显的区别. LLM4CDR[14]是 LLM-based 范式的典型代表, 其作者在论文中对比了基于 LLM 的跨域推荐模型和传统的跨域推荐模型 (例如 SSCDR、BiTGCF、CLFM) 在推荐性能的表现. 实验结果表明, 在多数情况下, 基于 LLM 的模型在性能上更具优势. 具体而言, 该工作在 Amazon 数据集上开展实验, 我们以 Books 为目标域, Movies 为源域的设置条件为例, 可以观察到 LLM4CDR 在 Top-K 推荐任务 (K=5) 中 Precision、Recall 和 NDCG 等指标均取得最佳结果 (如表 2 所示, 加粗为最优), 这显示出 LLM 在跨域推荐任务中良好的适应性和有效性.

同样属于 LLM-based 范式的 Exp3rt[21]也在 Amazon 数据集上进行了实验. 其对比了传统协同过滤模型 MF、基于 LLM 的模型 LLMRec 以及 Exp3rt 在评分预测任务中的性能表现, 实验结果如表 3 所示, 加粗为最优, 下划线为次优. 通过实验数据, 我们除了可以观察到 Exp3rt 在各项指标上取得最佳, 还可以看到 LLMRec 在冷启动推荐场景下的 MAE 指标低于 MF. 更重要的是, 其能够在未见用户/物品情况下完成推荐任务, 这表明 LLM-based 方法在冷启动或新用户/物品场景下推荐具有优势.

表 2 LLM 应用于跨域推荐实验结果

场景	模型	Precision@5↑	Recall@5↑	NDCG@5↑
Movies ↓ Books	SSCDR	0.5262	0.3380	0.8268
	BiTGCF	0.5246	0.3448	0.8286
	CLFM	0.5358	0.3492	0.8308
	LLM4CDR	**0.5383**	**0.3536**	**0.8980**

表 3 LLM 应用于冷启动推荐实验结果

模型	热启动场景		冷启动场景		未见用户/物品场景	
	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓
MF	0.6663	0.4588	0.6769	0.4501	—	—
LLMRec	0.7888	0.4623	0.7688	0.4286	1.4993	**1.0300**
Exp3rt	**0.6572**	**0.4369**	**0.6072**	**0.3858**	**1.4398**	1.1192

以上实验表明, LLM-based 范式在跨域、冷启动及未见用户/物品等场景下有着出色的表现. 然而, LLM 并非适用于所有场景. 例如, RecExplainer[30]对比了直接使用 LLM、LLM 结合 ICL 以及 SASRec 这 3 种方式在推荐任务上的性能表现. 实验结果如表 4 所示, 加粗为最优, Task1、Task2、Task3 分别对应下一个物品预测、物品排序和用户兴趣分类任务. 可以观察到, 使用 LLM 进行推荐的效果较差, 即便采用当时性能最强的 GPT-4 结合 ICL, 推荐效果仍明显不如 SASRec.

表 4 基于 LLM 的推荐与传统推荐对比实验结果

模型	Task1		Task2	Task3
	HR@5↑	NDCG@5↑	NDCG@5↑	ACC↑
Vicunna-7B	0.0026	0.0014	0.2391	0.5026
Vicunna-7B-ICL	0.0379	0.0304	0.2661	0.5070
GPT4-ICL	0.1105	0.064	0.6492	0.6338
SASRec	**0.6736**	**0.5234**	**0.8759**	**0.7768**

基于以上的实验及分析, 我们可以给出问题 1 答案: LLM-based 和 LLM-aid 范式在不同的推荐场景中推荐性能各有不同, LLM-based 因其丰富的世界知识以及强大的推理能力, 所以适合跨领域及冷启动推荐场景. 而 LLM-aid 则适合传统的如具有协同信号的协同过滤推荐、时间感知的序列推荐以及具有丰富用户-物品历史交互的热推荐场景.

在推荐模型无关范式中, 我们使用 MF、SASRec

这两种不同的推荐模型, 利用 RecExplainer 论文中提供的源代码来微调 LLaMA3. 在微调过程中, 仅设置推荐模型不同, 其余的如数据集、微调方法、超参数都相同, 实验结果如表 5 所示, 加粗为最优. 可以看到, 两种推荐模型微调出的代理解释模型在推荐性能上有很大差异 (此部分数据非原论文数据).

表 5 推荐模型无关推荐性能对于模型依赖性实验结果

模型	Task1		Task2	Task3
	HR@5↑	NDCG@5↑	NDCG@5↑	ACC↑
RecExplainer (MF)	0.1130	0.0731	0.6935	0.7292
RecExplainer (SASRec)	**0.7711**	**0.6681**	**0.8813**	**0.9657**

通过此实验, 我们可以给出问题 2 答案: 推荐模型无关不能笼统地认为推荐精度低于模型相关方法, 这取决于推荐模型无关方法中使用的具体推荐模型. 当采用推荐精度低的传统推荐模型时, 推荐模型无关推荐能力自然比推荐模型相关的差, 但当使用推荐精度高的推荐模型时, 推荐精度就会比模型相关的好. 所以我们也推荐该领域的研究人员在使用推荐模型无关方法时使用领域的 SOTA 推荐模型, 这样不光可以保证推荐性能, 也能在一定程度上减少 LLM 强行解释推荐原因而造成的“幻觉”问题.

1.5.2 解释质量对比

在我们的分类体系中, 推荐解释的质量是各框架研究的重点. 我们将对比微调方法和非微调方法 (即提示工程) 在解释方面的差异性, 并回答以下两个解释质量方面的关键问题.

问题 3: 微调方法是否在解释一致性上显著高于非微调方法?

问题 4: 微调方法适合哪些场景?

在用 LLM 进行推荐解释生成时, 微调方法往往会带来较好的解释效果, 但是这种解释效果并不适合该领域所有的研究人员, 因为微调对于算力要求极高. 例如 RecExplainer 在 Amazon 数据集训练 10 个 epoch 所使用的硬件条件就为 8 张 V100 32 GB 显卡, 我们复现其实验时使用 2 张 L40S 48 GB 显卡, 仅训练一轮就需要 4 h. 这对于大多数的研究人员来说是很难达到的研究条件. 而且微调方法是通过更新 LLM 参数来达到增强解释效果的目的的, 其既有优势又有劣势, 优势是微调使得 LLM 掌握推荐领域知识, 增强了推荐解释的效果. 劣势则是 LLM 容易出现灾难性遗忘. DRE[35]就指出 RecExplainer 这方面的问题, 在指出问题后, DRE 在充分利用评论数据的情况下, 通过细致地设计各种提示词, 并通过实验证实了其提出的方法在 Amazon 的 3 个子数据集上达到了比 RecExplainer 更好的性能表现, 实验结果如表 6 所示.

表 6 LLM 微调和非微调实验结果

方法	家居&厨房		服装鞋帽&珠宝		手机&配件	
	ASP↑	Rat↑	ASP↑	Rat↑	ASP↑	Rat↑
RecExplainer	0.6057	2.64	0.5628	2.68	0.6028	2.64
Mistral	0.7028	2.65	0.5757	2.79	0.6571	2.00
ChatGPT	0.6971	2.51	0.6362	2.86	0.6229	2.67
DRE-M	0.7142	2.68	0.6485	2.89	0.6857	2.57
DRE-C	**0.7714**	**2.88**	**0.6728**	**2.94**	**0.7400**	**2.90**

所以, 显而易见问题 3 的答案为“不是”. 微调方法的优势是简单粗暴, 让解释模型充分学习该领域的知识, 从而达到较好的解释效果, 但是这种方式会容易出现“幻觉”问题, 该问题会大大降低解释的质量.

LLM2ER-EQR[26]做的相关研究则可以回答问题 4. 该框架的核心模块为评分预测、个性化提示学习以及解释生成模块. 具有了这 3 个模块即可进行可解释推荐. 但是作者在文中指出这样做解释质量并不高, 所以他们又添加了 2 个奖励模型进行强化式微调, 以此来提升模型推荐解释的质量. 其在 Amazon Movies & TV 所作的实验结果如表 7 所示, 可以发现微调后的模型解释质量更高.

表 7 微调增强解释质量实验结果

方法	BLEU		ROUGE		Distinct	
	N=1	N=2	N=1	N=2	N=1	N=2
LLM2ER	16.37	1.193	16.874	14.176	18.400	64.982
LLM2ER w/HQAR	16.973	1.237	17.669	14.614	19.089	66.228
LLM2ER w/CCR	17.319	1.564	18.029	14.747	18.622	65.631
LLM2ER w/ERQ	**17.571**	**1.572**	**18.291**	**15.157**	**19.370**	**67.004**

所以我们可以给出问题 4 的答案: LLM 微调适合的场景为要求解释详细, 质量高, 且高度有领域特色的解释场景. 一般要求不是很高的场景可以通过预先处理例如评论等数据, 提取用户偏好、物品特征等因素构建提示词来实现简单的推荐解释.

1.5.3 计算开销与生成效率对比

在计算开销与解释生成效率方面, 我们整理了微调与非微调方法使用的预训练模型, 包括类型, 参数量等. 最终整理的结果如表 8 所示.

表 8 各模型/工作使用预训练 LLM 及生成效率

微调			不微调		
模型/工作	预训练LLM	备注	模型/工作	预训练LLM	备注
LLM4CDR[14]	(1) GPT3.5-Turbo (API) (2) LLaMA2-7B-chat (3) Mistral-7B Instruct v0.2	GPU: 4 NVIDIA A100 40 GB epoch: 15, batch: 64/GPU 训练方式: DeepSpeed (ZeRO) 训练速度: 3 h/模型	Zeng等人[23]	(1) Qwen2.5 (0.5B、1.5B、3B、7B、14B、32B、72B) (2) MING-MOE-7B (医学领域专用大模型)	(1) 评分时间 (s/batch): 1.23 (Qwen2.5-7B-Instruct) 8.67 (Qwen2.5-72B-Instruct) 2.45 (MING-MOE-7B) (2) 解释生成时间 (s/batch): 3.56 (Qwen(2)5-7B) (3) GPU: 8 NVIDAI A800 80 GB, 本地部署
ECCR-LLM[18]	(1) T5-Small (60M) (2) T5-Base (220M) (3) OpenLLaMA-3B (QLora, 仅微调200M)	GPU: 8 NVIDIA A10G 24 GB epoch: 30, batch: {8, 16, 32, 64} 训练速度(min/epoch): 依次为19.0、69.0、693.5	Chat-REC[24]	(1) GPT3.5-Turbo (API) (2) text-davinci-003 (API) (3) text-davinci-002 (API)	—
			ELMAR[25]	GPT-4 (API)	—
CIER[19]	LLaMA2-7B	GPU: NVIDIA H800 epoch: 3	LR-Recsys[32]	(1) GPT-3.5 (API) (2) GPT-4 (API) (3) LLaMA3.1 (参数量未知)	调用时间 (ms/50 words): 1 700 (GPT-3.5) 9 800 (GPT-4) 经济成本: 调用一次0.01美元, 2.96 ms/token
Pleaser[20]	(1) T5-small (2) T5-base (3) T5-large	GPU: 4 NVIDIA A100 40 GB CPU: AMD EPYC 7543 (2.8 GHz) 参数(FPFT): 依次为 (1) 推荐35.9M, 解释78.5M (2) 推荐111M, 解释251M (3) 推荐339M, 解释780M	CrossDR-Gen[33]	GPT3.5-Turbo (API)	—
			LLME4RS[34]	LLaMA2-13B (API)	—
			DRE[35]	(1) Mistral (8×70B) (2) GPT3.5-Turbo-0613 (API)	—
Exp3rt[21]	(1) GPT-3.5 (API, 教师LLM) (2) LLaMA3-8B (学生LLM)	GPU: NVIDIA RTX3090 32 GB epoch: 10	Lin等人[36]	GPT-2 (参数量未知)	—
			Logic-Scaffolding[37]	Falcon-40B	—
HDRec[22]	(1) LLaMA3-8B-instruct (2) T5-small	GPU: NVIDAI V100 32 GB	Peng等人[38]	GPT-2 (参数量未知)	—
			Abu-Rasheed等人[39]	GPT4-1106-preview (API)	—
LLM2ER-ERQ[26]	(1) GPT-2 (参数量未知) (2) BERT (参数量未知)	—	PRAG[40]	(1) DistilGPT2 (0.35B) (2) T5 (未说明具体版本)	—
InteRecAgent[27]	(1) GPT-4 (API) (2) LLaMA2-7B	—	Kovacs等人[41]	(1) Mistral 7B (英语版) (2) Vigogne 2 7B (法语版)	本地部署仅在CPU运行两LLM推理时间均约为8 s
MuseChat[28]	(1) GPT-3.5 (API) (2) Vicunna-7B	GPU: 16 NVIDAI V100 32 GB 音乐推荐模块batch: 34/GPU 句子生成模块epoch: 3	Ashaduzzaman等人[42]	GPT3.5-Turbo (API)	—
			Li等人[43]	GPT-2 (1.5B)	—
XRec[29]	(1) GPT-3.5-turbo (API) (2) LLaMA2-7B	—	Chun等人[44]	(1) LongChat-7B-32k (7B) (2) LongChat-13B-16k (13B) (3) Vicuna-33b-v1.3 (33B)	—
RecExplainer[30]	Vicuna-v1.3-7B	GPU: 8 NVIDAI V100 32 GB epoch: 10, batch: 64 训练方式: DeepSpeed (ZeRO-2)	LANE[45]	GPT-3.5 (API)	—
G-Refer[31]	(1) LLaMA2-7B (2) LLaMA3-8B (3) Qwen (0.5B、1.5B、3B、7B)	GPU: 8 NVIDAI A100 40 GB epoch: 2 batch (per GPU): 32 (LLaMA2-7B)、16 (LLaMA3-8B)	—		

在表 8 中, 若是微调方式, 则在备注中添加微调使用硬件设备, 训练轮数及训练时长等信息. 若是非微调方式, 则在备注中加入调用时间、是否为本地部署以及 API 调用等信息. 据此, 我们可以观察到大部分微调方式对计算资源有较高要求, 不微调直接调用 API 非常方便, 但经济成本较高且存在延迟问题.

本节系统整理了现有的使用 LLM 用于推荐系统可解释的工作, 首先提出我们的分类体系, 然后对领域内现有的工作进行了整理与分类, 阐述各工作的基本思路后, 我们量化对比了各类方法后发现最具实用场景的是推荐模型无关的带微调的 LLM 辅助型推荐系统范式, 这类工作在同时保持原先推荐模型高精度的推荐情况下, 又提供了高质量的推荐解释, 具有较大的研究前景.

2 评价指标

推荐系统中解释文本的质量评估是衡量可解释性效果的关键环节. Hulstijn 等人[46]系统性地提出了推荐解释的评价指标体系, 他们将其划分为主观指标和客观指标两大类. 其中, 主观指标主要通过用户调查问卷及深度访谈等方式获取, 而客观指标则是在系统运行过程中实时统计获得. 在具体评价标准方面, 解释质量通常从可理解性、详细程度、完整性、可操作性、准确性、可信度等多个维度进行评估.

然而, 现有评价体系存在两个主要局限性: 一方面, 这些评价指标的定义较为抽象, 缺乏具体可操作的衡量标准; 另一方面, 这些指标主要针对工业应用场景设计, 难以直接迁移到学术研究场景. 为此, 本节将聚焦学术研究场景, 重点探讨基于文本的可解释推荐系统的评价方法. 具体而言, 我们将评价方法分为两大类: 人工评价和自动评价. 其中自动评价又可进一步细分为基于传统指标的评价和基于 LLM 的评价方法. 另外, 还有一类指标虽用得不是很广泛, 但有其独特价值, 我们将其归类到拓展指标中.

2.1 人工评价

当前, 在推荐系统可解释性研究中, 多数研究者仍然是通过构建人工评价体系来验证其提出的可解释性框架的有效性. 具体而言, 现有研究主要从以下 3 个核心维度展开评估.

首先是推荐接受度维度, ELMAR[25]提出了整体接受度和推荐位置接受度的评价指标, Kovacs 等人[41]则在此基础上补充了遵循推荐的可能性和用户推荐满意度等指标; 再者是文本质量维度, 不同研究各有侧重, 如 ECCR-LLM[18]关注解释的正确性、信息量和合理性, Logic-Scaffolding[37]强调相关性、可读性和事实性, UARM[47]则着重评估流畅度、连贯性和说服力; 最后是系统体验维度, Kovacs 等人[41]提出的评价框架包含了透明度、易用性等指标, 而 Exp3rt[21]和 Ashaduzzaman 等人[42]则分别从总体质量和用户系统满意度等综合角度进行评估.

这些多维度的评价体系为文本解释质量的全面评估提供了系统化的方法论支持. 但是这些指标也存在局限性, 如评价标准缺乏明确定义导致可操作性不足, 主观性较强的评价方式可能影响结果的可比性, 加之冗长的评价周期难以满足实时推荐场景下对解释效果的即时反馈需求等, 不管在工业界还是学术界都难以使用.

2.2 自动评价

第 2.1 节介绍的人工评价体系不管是工业界还是学术界评价都显得耗时耗力, 接下来我们将介绍自动评价体系, 这类评价体系并不需要人工参与, 但评价效果却与人工评价相差不多, 在实际的应用场景中用得较为广泛. 详细指标如表 9 所示, 我们将其分成 3 个类别, 分别是传统评价指标、结合 LLM 指标以及拓展指标. 在每个类别下, 展示了各指标的名称、期望方向及作用. 其中期望方向中“↑”表示越高越好, “↓”则表示越低越好.

2.2.1 传统指标

传统研究在评估生成解释时, 基本思路是与用户自我解释 (如评论) 做匹配, 匹配度越高则生成解释质量越优, 这种评估方法主要从文本质量和可解释质量两个方面构建评价体系.

在文本质量评估方面, 多数研究采用 BLEU[48]和 ROUGE[49]两个指标来衡量生成文本与目标文本的形式相关性. 如 UARM[47]、LLM2ER-EQR[26]、ECCR-LLM[18]、CIER[19]、Li 等人[43]、Yu 等人[50]的研究均直接使用了这两个指标. 针对生成解释可能出现的同质化问题, NETE[51]引入的 USR 指标有效地对其进行了评估, 此指标在后续的研究中也被广泛使用.

在可解释质量评估方面, PETER[52]、CIER[19]、Li 等人[43]的研究采用了 FMR、FCR、DIV 等指标, Ariza-Casabona 等人[53]通过引入 FP、FR 指标进一步完善了可解释质量评估体系. 此外, 为了验证生成解释的

多样性, LLM2ER-EQR[26]和 PRAG[40]均采用了 Distinct 指标, 以衡量生成文本的词汇丰富性和重复性表达.

这些传统指标虽能在一定程度上解决相应问题, 但也存在明显局限, 例如 BLEU 和 ROUGE 仅衡量表面相似度, 无法反映语义一致性; USR、FMR、FCR、DIV 及 FP、FR 等依赖人工设定的特征, 且关注点局限于特征的形式, 容易忽略文本在流畅性、可读性和个性化等方面的语义特征; Distinct 只衡量文本表面的多样性, 无法评估语义层次的信息量, 若过度追求该指标, 可能会牺牲解释的连贯性与实用性.

表 9 解释文本评价指标

类别	指标	期望方向	作用
传统指标	BLEU	↑	通过计算生成解释和参考解释的*n*-gram匹配度衡量生成质量
	ROUGE	↑	评估生成文本与参考文本之间的相似度
	USR (唯一句子比例)	↑	衡量生成解释的唯一性
	FRM (特征匹配比例)	↑	从特征层面评估解释的有效性
	FCR (特征覆盖比例)	↑	从语料库层面评估特征的覆盖范围
	DIV (特征多样性)	↓	衡量不同用户-物品对的解释中特征的多样性
	FP (特征精度)	↑	衡量解释生成模型提取相关特征的准确性
	FR (特征召回率)	↑	衡量解释生成模型提取特征的完整性
	Distinct	↑	评估生成的解释在词汇和短语层面的多样性
结合LLM指标	GPTScore	↑	无需训练的自定义多维度文本评估
	BERTScore	↑	计算候选句子与参考句子之间的token级余弦相似度来衡量文本相似度
	BARTScore	↑	计算文本之间的生成概率来评估质量
	BLEURT	↑	给定真实解释通过生成解释的概率评估相似度
	LLM直接评价	↑	设置提示词对生成文本直接评分
拓展指标	FHR、D-FHR (特征幻觉率)	↓	评估句子、文档级别上衡量特征幻觉的程度
	perplexity score	↓	反映语言模型在预测序列中每个词时的不确定性
	COR (概念重复率)	↓	衡量生成解释的个性化和多样性, 从词级别的角度评估不同用户对同一物品生成解释的概念重叠程度
	CMR (概念匹配率)	↑	用于衡量生成解释与用户偏好和物品特征的一致性, 评估解释中包含的关键概念数量
	aspect score	↑	评估推荐系统生成的解释与用户真实偏好之间一致性
	情感匹配分数	↑	评估生成的解释与真实解释在情感一致性
	AP (说服力精准度)	↑	衡量解释中包含的说服性词对的比例
	AB Divergence	↓	衡量参考文本和生成文本的概率分布差异
	L2 Distance	↓	比较参考文本和生成文本的概率分布距离
	Fisher-Rao Distance	↓	衡量生成文本和参考文本之间的相似度

2.2.2 结合 LLM 指标

Zhang 等人[54]证明了 LLM 作为可解释推荐评价器的巨大潜力, 他们从说服力、透明度、准确性、满意度这四个方面分别使用 LLM 及人工的方式对解释文本进行评价, 得出 LLM 有与人工相当的评价能力这一结论, 这说明使用 LLM 评估解释文本具有巨大潜力.

在 XRec[29]中, 作者指出 BLEU 和 ROUGE 较弱的语义识别能力缺陷. 为了解决该问题, 他们提出新颖的基于 LLM 的评价指标, 这类指标通过输入生成解释-真实解释文本对, 然后经由 LLM 输出相似度评分从而达到评价生成解释文本质量的目的. 具体而言, 作者使用 GPTScore[55]、BERTScore[56]、BARTScore[57]、BLEURT[58]等方法评估解释文本质量. 对于 GPTScore, 该方法获取文本对文本级别的相似度后输出相似度分数. BERTScore 则计算 token 级别的文本对语义相似度, 并相应地输出准确率 (Precision)、召回率 (Recall) 和 $F1$ 分数 ($F1$-score); BARTScore 从文本生成角度出发, 利用 BART 模型在给定真实解释情况下生成解释的概率评估相似度, 概率越高, 表明越相似; BLEURT 通过对比学习捕捉文本语义差异, 进而输出综合性评分. G-Refer[31]则采用与 XRec 相同的思路, 在解释评价方面取得了较好的评价效果.

上述方法需要输入文本对, 但并非所有可解释推荐数据集中都包括用户对于物品的评论数据. 为此, RecExplainer[30]采取了一种替代方案, 不依赖真实解释文本, 而是直接运用预训练 LLM 的世界知识, 并结合

合适的评分提示词进行评估. 具体而言, 该方法通过主观分析, 设置 4 级评分体系来评估解释生成能力. 除此之外, 该篇文章作者还结合了人工评价, 同样采用四级评分体系, 证实人工评分与 LLM 评分具有高度一致, 表明该评估体系具有可靠性.

这些结合 LLM 的指标虽能够从语义层面较好地评估解释质量, 但也存在明显不足, 例如过度依赖预训练 LLM 的质量和规模、计算资源要求高、对提示词敏感等. 在事实性与深层逻辑评估方面, 它们的表现往往不理想, 容易忽略数字或专有名词等事实性错误. 此外, 一些此类指标还可能受到预训练模型固有偏见或“幻觉”的影响, 从而导致评估结果出现不公平甚至错误.

2.2.3 拓展指标

在梳理文献的过程中, 我们发现除了传统的评价指标以及结合 LLM 的指标外, 部分研究还提出了一些创新性的评价指标. 为便于参考, 我们将这些创新指标整理如下.

为评估 LLM “幻觉”问题对解释质量的影响, Ariza-Casabona 等人[53]提出了 FHR (句子级别) 和 D-FHR (文档级别) 指标, 这些指标专门用于衡量生成内容的真实性和可靠性. 与此同时, Petruzzelli 等人[14]在评估模型生成解释的质量时, 不仅引入了困惑度分数 (perplexity score) 来衡量文本的流畅性, 还结合了生成文本的平均长度和标准差, 以进一步量化解释的详实程度和稳定性. 此外, 在基于结构化知识的评估方面, LLM2ER-ERQ[26]设计了概念重叠率 (COR) 和概念匹配率 (CMR) 两项新指标, 从而能够更有效地评估生成解释的一致性.

在推荐解释的个性化评估方面, DRE[35]通过提取用户评论中的关键方面, 计算生成解释与真实用户偏好的方面分数 (aspect score, ASP) 重叠度, 从而衡量解释的个性化匹配能力. 类似地, Sent-XRec[59]不仅使用 BERTScore 评估生成解释与参考文本的语义相似度, 还引入情感匹配分数来量化解释在情感倾向 (正向/负向/混合) 上的对齐程度. 此外, Chen 等人[60]提出了说服力精准度 (AP) 指标, 通过统计解释中说服性词汇的比例来评估其解释效果.

在更复杂的多维度评估体系中, MuseChat[28]综合运用了 BERTScore、AB 散度 (AB divergence)、L2 距离 (L2 distance) 和费希尔-拉奥距离 (Fisher-Rao distance) 等指标, 从语义相似度、分布差异和向量空间对齐等多个角度全面评估生成内容的质量. 这些创新指标不仅丰富了可解释推荐系统的评估维度, 也为后续研究提供了更细粒度的分析工具.

这些拓展指标虽各具优势, 但也存在一定局限. 例如, FHR 的计算依赖于特征提取的准确性, 若特征标注不完整, 容易产生误判; D-FHR 容易忽略单个解释中的局部幻觉问题, 从而影响评价效果. 困惑度分数只能反映文本表面的统计特性, 难以衡量语义准确性、事实一致性等深层质量, 同时对生成内容的多样性与个性化缺乏区分度. COR 与 CMR 主要关注概念重叠, 却忽略了语义深度和上下文关联性, 并且对外部知识图谱或概念集的质量依赖较强, 在强调概念匹配的同时, 往往忽视了文本的流畅性与语义连贯性.

方面分数依赖评论数据的质量和完整性, 若评论稀疏或噪声较多, 则可能导致评估偏差. 情感匹配分数依赖外部情感分析工具 (如 BERT) 的准确性, 可能引入额外误差, 并且在区分混合或复杂情感场景时能力有限, 同时计算成本较高. 说服力精准度依赖词汇统计, 容易忽略上下文语义及用户的个性化需求.

AB 散度的计算依赖预训练语言模型, 容易受模型偏差影响, 对于短文本或文本重叠度较低的情况, 评估效果有限, L2 距离对文本的局部语义变化不敏感, 易受向量表示质量影响, 难以全面捕捉文本的复杂语义关系和上下文信息. 费希尔-拉奥距离的计算复杂度较高, 并且对数据分布的假设较为严格, 在实际应用中还会受到数据质量和模型输出稳定性的制约.

2.3 总结

在第 2.1 和 2.2 节中, 我们分别探讨了人工评价和自动评价方法. 值得注意的是, 由于这些指标各自存在局限性, 在实际研究中往往不会单独使用, 而是相互结合, 以充分发挥各自优势, 弥补不足, 从而提升评价结果的客观性和全面性. 例如, Peng 等人[38]通过 USR、FCR 和 DIV 指标评估生成文本的可解释质量的同时, 采用 BLEU 和 ROUGE 指标衡量文本生成质量. 类似地, Abu-Rasheed 等人[39]在人工评估中应用 Likert 量表进行主观评价, 并辅以 ROUGE 等自动化指标进行客观分析.

与推荐系统的推荐精度评估不同, 推荐解释质量的评价目前仍缺乏统一、完善的指标体系. 因此, 采用人工评价与自动评价相结合的混合评价模式仍是当前最可靠的解决方案. 这一现状也为该领域研究者提供

了重要启示: 开发更先进、更标准化的推荐解释质量评价指标, 将成为一个极具价值的研究方向. 未来研究可重点关注如何建立兼顾解释性、可信度和用户感知的多维评价框架.

3 数据集

无论是普通推荐还是可解释推荐, 都依赖大量数据. 这些数据来源各异, 有些公开共享, 可以直接获取使用; 有些则因隐私保护或商业机密等原因无法公开. 为了更好地支持相关研究, 我们对可解释推荐领域常用的数据集进行了梳理, 首先根据数据的可获取性将其分为公开数据集和私有数据集两类, 具体见第 3.1.1 节, 该节不仅展示了数据集的公开性信息, 还列出了数据集的名称、是否包含评论、主要特点及局限性等内容, 并特别附上了使用这些数据集的相关研究工作.

3.1 公开数据集

可解释推荐所使用的数据集与传统协同过滤推荐的数据集存在显著差异. 与仅需要用户—物品历史交互数据的传统推荐不同, 可解释推荐通常需要利用额外的辅助信息 (如用户评论、物品描述等) 来生成具有解释性的推荐结果.

Hasan 等人[61]认为用户评论作为自由形式的文本, 其能帮助推荐系统理解用户偏好和物品特征, 有效解决传统推荐的数据稀疏、可解释性差等问题. 评论数据对可解释推荐尤为重要, 因为它作为用户的自我表达, 能够真实而直接地反映用户的偏好. 正因如此, 在现有的大部分推荐系统研究中, 评论数据被广泛用作重要的信息源. 特别是在当前利用 LLM 进行可解释推荐的研究中, 评论数据常被作为基准真值 (ground truth) 用于模型微调和对齐, 以确保生成的解释与用户真实的表达意图保持一致. Kim 等人[21]也在他们的文章中提及评论数据的价值, 他们指出用户对物品的评论蕴含着丰富的主观偏好信息, 利用这些信息可以增强推荐及解释效果.

因此, 为了更深入地分析公开数据集的使用情况, 我们根据工作中是否使用评论对数据集进行了分类, 相关工作汇总结果如表 10 所示.

接下来将从两个方面对研究中所使用的数据集进行阐述: 对于涉及评论数据的研究, 我们将详细说明评论数据的运用方法; 而对于不包含评论数据的研究, 将重点解释其参考文本等数据生成过程与依据.

表 10 基于评论角度的公开数据集的相关工作

是否使用评论数据	模型/现有工作
使用评论数据	HDRec[22]、LLM2ER-EQR[26]、XRec[29]、G-Refer[31]、DRE[35]、Lin等人[36]、UARM[47]、Chen等人[60]、ALEX[62]
未使用评论数据	LLM4CDR[14]、InteRecAgent[27]、RecExplainer[30]、CrossDR-Gen[33]、LLME4RS[34]、LANE[45]、Yu等人[50]、Park等人[63]

3.1.1 使用评论数据的推荐解释

基于评论数据的可解释推荐系统研究取得了显著进展, 这些工作主要利用用户评分及文本评论来生成个性化推荐解释. 针对不同的评论数据处理方法和模型架构, 这些研究可以系统地分为以下 3 类.

(1) 研究聚焦于从评论数据中构建用户和物品档案. XRec[29]在 Amazon Book、Yelp 和 Google 数据集上, 利用 LLM 从评论中提取用户意图, 并结合物品元数据据形成物品档案, 再整合用户历史交互生成个性化解解释. G-Refer[31]使用与 XRec 相同的数据集, 它通过评论数据构建用户-物品二分图来融合协同过滤信号和语义信息, 对 最终由 LLM 生成解释. DRE[35]则专注于在 Amazon 的 3 个子类别数据集 (电子产品、服饰和家居用品) 上, 通过分析用户历史评论和其他用户评论来构建用户偏好画像和物品特征描述.

(2) 研究侧重于从评论中提取结构化概念来增强推荐. LLM2ER-EQR[26]在 Amazon Movie & TV、Yelp 和 TripAdvisor 数据集上, 通过从评论数据中匹配 *n*-gram 短语概念图来构建用户偏好和物品属性的概念集合, 再结合预训练的因果语言模型生成解释.

(3) 研究直接将评论作为解释生成的监督信号. Lin 等人[36]在 TripAdvisor 数据集上将评论文本作为基准真值来训练解释生成模型. HDRec[22]则在 Amazon Sports、Beauty 数据集上, 通过分层蒸馏从评论中提取更细致的交互原理, 生成专门的真实解释和评论摘要, 然后用这些数据对齐 LLM 输出, 以此获得令用户满意的推荐解释.

综上, 使用评论数据进行可解释推荐的研究涵盖电商、本地服务和旅游等多个领域, 验证了基于评论的可解释推荐方法的广泛适用性. 具体对比见表 11.

3.1.2 未使用评论数据的推荐解释

可解释推荐系统的研究不仅限于使用评论数据, 许多工作通过挖掘物品元数据、用户交互行为和结构化信息来生成推荐解释. 这些研究主要可分为两类.

表 11 可解释推荐数据集分类

类别	数据集		是否包含评论	主要特点	局限性	使用的模型/工作
公开数据集	Amazon	Movies &TV	是	包含用户评分、电影/TV节目元数据 (如导演、演员)	可能存在虚假评论; 覆盖范围以英语内容为主	LLM4CDR[14]、LLM2ER-ERQ[26]、Rec-Explainer[30]、LR-Recsys[32]、Peng等人[38]、CIER[19]、PRAG[40]、Li等人[43]
		Books		涵盖书籍评分、详细评论、出版信息	评论者偏差, 如偏好畅销书	LLM4CDR[14]、Exp3rt[21]、XRec[29]、G-Refer[31]
		Beauty		包含产品成分、用户使用体验	评论数量不均衡, 其中热门产品居多	HDRec[22]、InteRecAgent[27]、LANE[45]
		其他		多样化子类别, 如电子产品、家居	—	LLM4CDR[14]、Pleaser[20]、HDRec[22]、RecExplainer[30]、DRE[35]
	Yelp		是	包含用户对餐厅、商家的评分和文本评论, 覆盖多领域	数据可能存在地域偏差(以北美为主); 部分评论内容较短或低质量	CIER[19]、LLM2ER-ERQ[26]、XRec[29]、G-Refer[31]、LR-Recsys[32]、Peng等人[38]、PRAG[40]、Li等人[43]
	TripAdvisor		是	涵盖酒店、景点、餐厅的详细评论和评分, 用户交互丰富	存在虚假评论问题; 数据分布不均衡 (热门地点评论更多)	CIER[19]、LLM2ER-ERQ[26]、LR-Recsys[32]、Peng等人[38]、PRAG[40]、Lin等人[43]
	Steam		是	包含游戏评分、用户评测和社区互动数据	偏向硬核玩家群体, 普通用户评论较少	InteRecAgent[27]、RecExplainer[30]、LANE[45]
	MovieLens		否	纯评分数据集, 包含用户对电影的评分和标签, 无文本评论	缺乏上下文信息 (如用户为何喜欢某电影)	Chat-REC[24]、InteRecAgent[27]、LANE[45]
	Movie-Lens	+TMDB	否	TMDB为MovieLens提供电影元数据 (如导演、演员、剧情简介)	缺乏用户评论, 无法分析文本情感; TMDB元数据需额外处理	LLME4RS[34]
		+IMDB	是	结合MovieLens的评分数据和IMDB的文本评论, 支持评分与评论联合分析	IMDB评论可能存在语言偏差 (以英语为主);数据整合复杂度高	Logic-Scaffolding[37]
	Google Reviews		是	包含用户对地点 (如餐厅、商店)的评分和详细评论, 覆盖范围广	数据质量参差不齐 (如简短或广告性评论); 地域覆盖不均衡	XRec[29]、G-Refer[31]
	IMDB		是	电影领域的专业评论和评分数据集, 包含详细用户评论和电影元数据	评论者多为电影爱好者, 可能缺乏普通观众视角	Exp3rt[21]
	Foursquare	TKY	是	东京地区的用户签到和评论数据	数据局限于东京, 泛化性有限;隐私问题导致数据匿名	CrossDR-Gen[33]
		NKY		纽约地区的用户签到和评论数据	数据时效性可能不足 (如旧评论占比高)	CrossDR-Gen[33]
私有数据集	DrRank		否	专业医生排名数据	存在数据共享壁垒	Zeng等人[23]
	MuseChat数据集			视频音乐对话数据		MuseChat[28]
	Citation网络数据集			导师推荐数据		Ashaduzzaman等人[42]
	Saks Global金融数据集			养老基金推荐数据		ELMAR[25]
	新加坡南洋理工课程			课程推荐解释数据		Chun等人[44]
	智能家居能源数据			智能家居能源数据		Kovacs等人[41]
	收集人工生成样本			手工收集		Abu-Rasheed等人[39]

(1) 研究利用物品元数据增强推荐解释. Rec-Explainer[30]在 Amazon Video Games、Amazon Movies & TV 及 Steam 数据集上, 不仅使用协同过滤信号, 还整合物品名称和推荐模型嵌入来生成解释. 类似地, LANE[45]在 MovieLens、Amazon Beauty 和 Steam 数据集上, 用物品标题的语义编码替代传统物品编号 (item ID), 结合提示模板和语义对齐模块优化解释生成. LLM4-CDR[14]则专注于跨域推荐场景, 通过 Amazon Movies、CDs 和 Books 等数据集的物品描述特征 (如类别、品牌) 构建提示语来微调模型以此来生成解释.

(2) 研究侧重于结合外部数据. LLME4RS[34]同时使用 MovieLens 和 TMDB 数据, 后者提供丰富的电影元数据 (如剧情简介、演职员信息), 使模型能基于类型相似度、用户评分模式等生成解释. InteRecAgent[27]在 Amazon Beauty、Steam 和 MovieLens 数据集上, 通过整合物品结构化信息提升解释质量. MuseChat[28]则利用 YouTube8M 的视频元数据构建推荐解释.

这些研究展现了不依赖评论数据仍能构建有效可解释推荐系统的多种途径, 通过利用物品元数据、领域知识和结构化信息, 在不同应用场景下也实现了较高质量的推荐解释.

3.2 私有数据集

在推荐系统研究中, 私有数据集通常面临规模有限且缺乏用户评论数据的挑战, 这使得传统基于评论的分析方法难以直接应用. 为弥补这一数据缺失问题, 近期研究探索了利用 LLM 生成与数据集特性相匹配的参考解释文本的创新方法. 这些生成文本可以作为替代性语义信息源, 为推荐系统提供额外的解释维度.

基于私有数据集, 利用 LLM 构建参考解释文本是一种有效的方法, 以 ECCR-LLM[18]为例, 该研究利用 Amazon 电商平台数据, 重点聚焦手机、咖啡和手袋这 3 大商品品类进行模型构建. 为支持互补推荐任务, 他们对原始数据进行了多维度预处理, 包括物品属性、兼容关系、用户共购行为以及筛选后的高相关互补数据, 其中所有参考解释文本均采用少样本提示的 ChatGPT 生成.

未使用参考解释文本的工作因其相对简单的实现方式和较高的实用性, 在实际应用中占据了重要地位. 这类研究通常基于特定领域的专有数据集展开, 通过挖掘数据内在特征构建推荐模型. 从金融投资到医疗健康, 众多学者利用不同领域的私有数据开展了富有成效的探索. 例如, ELMAR[25]采用合作金融机构的私有数据集进行金融推荐; Kovacs 等人[41]使用瑞士家庭能源系统监测数据, Abu-Rasheed 等人[39]构建了人工标注样本数据集, Zeng 等人[23]则利用专业医生排名数据集 DrRank 实现医生推荐. 这些研究展现了未使用参考解释文本方法在不同应用场景中, 解释直接利用物品属性也有一定的有效性.

3.3 总结

LLM 应用于推荐可解释的核心在于利用评论数据揭示用户偏好. 评论作为用户意图的直接表达, 不仅能增强推荐准确性, 更为解释生成提供了自然语言依据. 对于缺乏评论的私有数据, 研究者通常借助 LLM 生成模拟评论作为补充. 评论数据的质量直接影响解释的可信度, 这使其成为可解释推荐系统不可或缺的关键要素. 未来需进一步探索评论数据的深度挖掘方法, 以提升解释的个性化和说服力.

4 应用场景

在我们综述的文献中, 既涵盖学术领域的理论创新, 也涉及多个实际应用场景的研究. 本节重点整理了 LLM 在推荐系统可解释性方面的实践应用, 具体包括教育、金融、医疗、电子商务和新闻媒体等领域. 这些研究不仅展现了 LLM 技术广泛的应用前景, 还为提升推荐系统可解释性提供了丰富的实践案例.

4.1 教育领域

在教育领域, 可解释推荐系统旨在提升学习成效、激发学习兴趣、适配个性化学习路径并增强用户信任, 其解释需求深度融合了教学规律与认知科学原理. 此类解释不仅承担推荐逻辑的功能, 还需发挥引导知识教学的作用, 要求具备依据学习者知识水平与认知风格动态调整复杂度与呈现方式的适配能力. 知识准确性在该场景中具有刚性约束, 任何事实性或概念性偏差均可能对学习过程造成负面影响, 因而 LLM 生成的解释必须严格对齐课程标准与学科规范.

现有研究多通过融合领域知识与智能生成技术实现这一目标. Abu-Rasheed 等人[64]结合知识图谱与专家经验构建系统, 使解释可关联教材核心概念, 并以文本、标签、雷达图等多模态形式满足差异化学习需求, 其 LLM 驱动的智能体还可通过问答交互纠正学习者的理解偏差. 针对教育推荐的“可迭代性”特征, Hendrawan 等人[65]提出 Grapevine 系统允许学习者通过关键词调整推荐方向, 在系统的解释模块中, LLM 动态生成解释并实时响应, 形成“反馈-调整-再解释”的闭环.

在具体技术路径的实现上, Li 等人[66]将知识追踪 (KT) 与检索增强生成 (RAG) 相结合, 使 LLM 能基于学习者答题历史与课程知识库生成针对性解释, 实现资源与认知缺口的精准匹配; Ma 等人[67]则利用课程注册时序数据对 LLM 进行微调, 使解释不仅反映课程内容关联, 还体现教学进度的逻辑性, 例如在推荐先修课程时明确说明其对后续学习的支撑作用.

4.2 金融领域

金融领域的可解释推荐系统以辅助理性决策、管

理风险、满足合规要求与建立用户信任为核心目标, 其解释机制需高度契合该领域对专业性、准确性与透明度的严苛标准. 与其他领域相比, 金融推荐中的解释不仅用于帮助用户理解推荐逻辑, 更是规避决策风险、符合监管规范的重要保障. 以基金推荐为例, 解释需同时覆盖历史收益表现、波动率等量化指标, 以及重仓行业的政策风险等潜在隐患, 这种“风险-收益双维度”信息直接影响投资决策质量.

该领域的解释约束主要体现在 3 个方面. 首先, 专业性与合规性是首要前提, 解释必须使用精准的金融术语并严格遵循监管要求, 其次, 风险透明度要求解释在呈现产品优势的同时, 客观揭示潜在风险及成因. 最后, 可追溯性要求关键数据来源清晰可查, 以便审计与验证.

现有研究在满足上述约束方面进行了针对性探索. Yu 等人[50]通过融合公司概况、新闻事件与市场数据设计提示模板, 使 LLM 生成的解释既能揭示资产价格波动的逻辑, 又可借助思维链推理呈现风险传导路径, 增强可信度. ELMAR[25]则结合相似用户画像技术, 在冷启动场景中参考同类投资者的风险偏好生成解释, 并探索内置合规检查模块, 确保输出严格符合金融监管条文.

4.3 医疗领域

在医疗领域, 可解释推荐系统的设计旨在支持诊断决策、促进健康行为改变、提升治疗依从性, 并构建稳固的医患信任机制. 其解释机制直接关涉健康结局与医疗安全, 因此在专业性、准确性与个性化方面的要求显著高于其他领域. 与教育、金融场景不同, 医疗推荐的解释不仅需呈现推荐逻辑, 还应作为医患沟通的有效媒介, 既要满足专业医务人员对医学合理性与科学严谨性的评估标准, 又需确保患者能够理解建议的理论依据与可执行方案. 例如, 在心血管疾病预防情境下, 解释应同时涵盖“药物与当前血压水平匹配的医学依据” (专业层面) 与“每日服药时间及注意事项” (患者层面).

医疗解释的独特约束体现于多重维度. 高度的专业性与准确性构成首要准则, 解释必须严格遵循临床实践指南及循证医学证据, 且术语使用需精确对标国际疾病分类 (ICD) 或解剖学标准. 例如, 在推理过程中需明确区分“心肌梗死”与“心绞痛”的病理差异与诊断逻辑. Afanasieva 等人[68]提出的 CaRiFaM 算法通过三维推荐体系实现此要求, 其中由 ChatGPT 生成的解释不仅与心脏病防治指南保持一致, 还结合患者的特定检查指标 (如血脂水平、血压值), 以确保医学表述的严谨性与可溯性. 此领域的推荐解释还需具备动态更新知识的能力. Wali 等人[69]开发的系统正是如此, 他们将通过智能手表直接传输心率和血压等数据实时接入系统, 使解释内容始终与病人动态保持一致.

针对医疗场景特性的适配策略呈现出多样化的技术路径. ALEX[62]采用 LLM 作为解释生成层, 对接 BERT 模型基于公共卫生数据的预测结果, 并利用提示词工程、少样本学习及针对性约束对解释进行专业性验证与偏差修正, 从而确保疫情监测推荐的解释符合流行病学规范. Wali 等人[69]则提出“分类器+LLM”架构, 先使用 CatBoost 模型实现心脏病风险分级, 再由 LLM 驱动的对话代理将风险结果转化为个性化健康建议, 其解释包含基于 SHAP 值识别的关键风险因素, 从而帮助用户直观理解模型预测背后的主要驱动因素.

4.4 电子商务领域

电子商务领域的可解释推荐系统旨在促进用户转化、提升满意度与黏性、帮助用户挖掘潜在需求, 并建立对平台的信任. 其解释机制需深度契合消费场景的决策特性, 既要迅速抓住用户注意力, 又要精准传递商品价值, 从而降低决策成本. 与其他领域不同, 电商推荐的解释不依赖严谨的专业术语体系, 而是强调以通俗化表达直击用户痛点.

该领域的独特约束体现在多维度的平衡中. 通俗性与吸引力是首要要求, 解释需采用消费者易懂的语言, 避免行业黑话, 并通过场景化表达激发购买欲望, 例如将“高性价比”转化为“同配置价格低 20%, 多数用户反馈可使用 3 年以上.”, Xiang 等人[70]提出的 Transformer 模型通过分析用户历史浏览的关键词, 使生成的解释自动关联使用场景, 提升了与用户需求的契合度. 个性化与相关性同样是关键约束, 要求解释紧密绑定用户的显性偏好与隐性需求. Wang 等人[71]提出的 LLM-PKG 借 LLM 生成产品知识图谱, 经验证映射至企业产品, 支持个性化推荐, 提升用户参与和交易. 可信度与说服力的构建也至关重要, 现有方法多通过引用用户生成内容 (UGC) 增强解释的真实性, 这类基于群体反馈的表述往往比平台自夸更易获得信任. 例如, Liu 等人[72]的 LLM-CRS 框架通过 LLM 与 CRS 协作, 结合商品推荐与对话生成, 提升推荐可信度与交互说服力.

在适配上述多重约束的过程中, 现有技术呈现出清晰的演进趋势. Xu 等人[73]利用 LLM 的跨领域知识迁移能力, 在用户行为数据有限的冷启动场景中提供显著优势, 能为推荐生成可理解的解释, 有效缓解了冷启动问题. LLM-PKG 框架[71]则依托知识图谱的结构化特性, 使解释既能呈现“商品-用户”的关联路径, 又能追溯生成依据, 提升了可审计性.

4.5 新闻媒体领域

在新闻媒体领域, 可解释推荐系统旨在满足用户多样化信息需求、提供多元视角、打破信息茧房、提升用户参与度, 并强化媒体公信力. 其解释机制需高度契合新闻传播所固有的时效性、客观性与关联性特征. 相较于电商、教育等领域, 新闻推荐解释不仅应阐明“为何推荐该内容”, 更需揭示“内容与用户既有认知之间的关联机制”.

该领域的特性主要体现为多维度的动态要求. 时效性是首要条件, 新闻事件的快速演进决定了解释必须具备实时更新能力. Liu 等人[74]提出的 RecPrompt 框架通过双 LLM 协同机制满足这一需求: 其一, 基于用户阅读历史总结兴趣主题并生成对应新闻列表; 其二, 依据推荐结果与用户真实反馈的匹配度优化解释模板, 从而确保推荐理由与用户兴趣保持一致. 在该领域中, 深度关联与背景补充则要求解释超越单纯的“兴趣匹配”. 例如, Gao 等人[75]提出的 GNR 范式, 通过生成式叙事将焦点新闻与历史报道及相关领域分析进行时序化整合, 形成具有背景深度的解释内容. 立场中立与偏见规避是新闻推荐解释的核心伦理要求, 需要防止 LLM 训练数据中潜在倾向性的放大效应. RecPrompt[74]通过自动提示工程减少人工干预, 并利用 LLM 基于用户历史行为提炼兴趣主题, 数据驱动的优化机制可降低训练数据偏见对推荐的影响.

现有方法在适配上述特性时也呈现出差异化的技术路径. GNR[75]采用主题级与语义级双层表征体系, 以捕捉新闻间的深层语义关联, 并结合生成式叙事技术呈现事件发展的时间线逻辑; RecPrompt[74]则通过构建“推荐器-调优器”闭环机制, 使解释能够在用户反馈驱动下持续优化, 实现对新闻动态与用户认知变化的双重响应.

4.6 总结

综上, LLM 可解释推荐方法在教育、金融、医疗、电子商务及新闻媒体领域的应用呈现出显著的场景特异性, 其核心差异源于各领域对解释目标、约束条件及技术路径的不同要求. 这种差异性不仅体现在解释内容的呈现形式上, 更深刻反映在领域特性对方法选择的底层约束中. 这种深刻的领域依赖性决定了构建成功的 LLM 可解释推荐系统, 必须将领域知识深度融入模型设计、解释生成及效果评估的全过程.

5 现有挑战及未来方向

5.1 现有挑战

可解释推荐系统在提升用户信任和增强推荐透明度方面发挥着重要作用, 但仍面临诸多挑战.

(1) 在解释的准确性与可控性方面, 由于生成的解释可能包含不准确或虚构的内容, 其可信度容易受到影响. 此外, 部分解释仅用于合理化推荐结果, 而非真正反映推荐决策过程, 且现有方法多依赖相关性而非因果关系, 导致难以提供深层次的解释.

(2) 计算效率与实时性也是当前面临的重要挑战. LLM 的微调成本较高, 限制了其在大规模推荐系统中的部署. 同时, 在复杂的推荐场景下 (如长序列或多模态数据), LLM 受限于上下文窗口长度, 难以高效生成解释, 从而影响系统的实时性.

(3) 多模态融合与跨域泛化能力仍有待提升. 现有方法主要依赖文本信息, 未能充分利用图像、视频等多模态数据来增强解释的丰富性. 此外, 在跨域推荐场景中, 解释方法的泛化能力不足, 导致其难以适应不同领域的推荐需求.

(4) 评估标准的缺失进一步制约了可解释推荐系统的发展. 目前, 业界缺乏统一的解释质量评估指标, 例如解释的一致性、用户理解度等, 这使得不同方法的优劣难以客观衡量, 也阻碍了可解释推荐技术的进一步优化和应用落地.

5.2 未来方向

可解释推荐系统的发展虽然面临诸多挑战, 但也催生了一系列值得深入探索的研究方向.

(1) 在提升解释质量与可控性方面, 研究者可通过引入外部知识 (如知识图谱) 来约束 LLM 生成解释的真实性, 同时结合词组级细粒度解释和反事实分析等技术, 进一步增强解释的可信度和可解释性. 动态解释生成与个性化适配同样是解决该挑战的重要策略, 通过根据用户偏好自动调整解释的详细程度, 并构建“解释-反馈-优化”的闭环机制, 结合强化学习 (RLHF) 优

化解释策略, 能够更好地满足不同用户的需求.

(2) 高效可解释推荐框架的探索至关重要, 例如采用语义嵌入和零样本提示降低计算成本, 并优化长序列推荐的解释生成机制, 以克服 LLM 上下文窗口的限制.

(3) 多模态融合与跨域解释的研究也备受关注, 通过整合图像、音频、视频等丰富信息, 可以生成更具表现力的解释, 同时设计通用解释框架以适配电商、医疗、金融等不同领域的推荐需求.

(4) 评估与伦理考量同样是该领域不可忽视的重要议题. 当前亟需建立统一的解释质量评估标准, 涵盖解释一致性、用户理解度等关键指标, 同时关注解释的公平性、隐私保护等伦理问题, 以确保可解释推荐系统的健康发展.

6 结语

本文系统综述了 LLM 在推荐系统可解释性研究中的应用进展, 从领域研究现状、评价指标、数据集和应用场景这 4 个方面进行了全面梳理. 研究发现, LLM 通过其强大的语言生成和语义理解能力, 提升了推荐解释的质量和个性化程度, 尤其在基于评论数据的场景中表现突出.

现有研究可分为基于 LLM 的推荐系统和 LLM 辅助型推荐系统两大类, 其中推荐模型无关的带微调 LLM 辅助型范式展现出较高的实用价值. 然而, 该领域仍面临解释准确性、计算效率、多模态融合和评估标准等挑战. 未来研究应聚焦于知识增强的解释生成、动态个性化适配、跨领域通用框架以及标准化评估体系的构建, 以推动可解释推荐系统的进一步发展与应用落地.

参考文献

1 纪守领, 李进锋, 杜天宇, 等. 机器学习模型可解释性方法、应用与安全研究综述. 计算机研究与发展, 2019, 56(10): 2071–2096. [doi: 10.7544/issn1000-1239.2019.20190540]

2 成科扬, 王宁, 师文喜, 等. 深度学习可解释性研究进展. 计算机研究与发展, 2020, 57(6): 1208–1217. [doi: 10.7544/issn1000-1239.2020.20190485]

3 苏宇. 优化算法可解释性及透明度义务之诠释与展开. 法律科学 (西北政法大学学报), 2022, 40(1): 133–141.

4 Sheu RK, Pardeshi MS. A survey on medical explainable AI (XAI): Recent progress, explainability approach, human interaction and scoring system. Sensors, 2022, 22(20): 8068. [doi: 10.3390/s22208068]

5 Yeo WJ, Van Der Heever W, Mao R, *et al*. A comprehensive review on financial explainable AI. Artificial Intelligence Review, 2025, 58(6): 189. [doi: 10.1007/s10462-024-11077-7]

6 Ribeiro MT, Singh S, Guestrin C. “Why should I trust you?”: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016. 1135–1144.

7 Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 4768–4777.

8 Zanon AL, da Rocha LCD, Manzato MG. Balancing the trade-off between accuracy and diversity in recommender systems with personalized explanations based on linked open data. Knowledge-based Systems, 2022, 252: 109333. [doi: 10.1016/j.knosys.2022.109333]

9 Zhang YF, Chen X. Explainable recommendation: A survey and new perspectives. Foundations and Trends® in Information Retrieval, 2020, 14(1): 1–101. [doi: 10.1561/1500000066]

10 Chatti MA, Guesmi M, Muslim A. Visualization for recommendation explainability: A survey and new perspectives. ACM Transactions on Interactive Intelligent Systems, 2024, 14(3): 19. [doi: 10.1145/3672276]

11 Feng YJ, Feuerriegel S, Shrestha YR. Contextualizing recommendation explanations with LLMs: A user study. arXiv:2501.12152, 2025.

12 Lu HY, Ma WZ, Wang YF, *et al*. User perception of recommendation explanation: Are your explanations what users need? ACM Transactions on Information Systems, 2023, 41(2): 48. [doi: 10.1145/3565480]

13 Silva I, Said A, Marinho LB, *et al*. Leveraging large language models for recommendation and explanation. Proceedings of the 10th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS 2023) Co-located with 17th ACM Conference on Recommender Systems (RecSys 2023). Singapore: CEUR-WS.org, 2023. 74–81.

14 Petruzzelli A, Musto C, Laraspata L, *et al*. Instructing and prompting large language models for explainable cross-domain recommendations. Proceedings of the 18th ACM Conference on Recommender Systems. Bari: ACM 2024. 298–308. [doi: 10.1145/3640457.3688137]

15 Lubos S. Improving recommender systems with large language models. Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization. Cagliari:

ACM, 2024. 40–44. [doi: 10.1145/3631700.3664919]
16 Okoso A, Otaki K, Koide S, *et al*. Impact of tone-aware explanations in recommender systems. ACM Transactions on Recommender Systems, 2025, 3(4): 55.
17 Said A. On explaining recommendations with Large Language Models: A review. Frontiers in Big Data, 2025, 7: 1505284. [doi: 10.3389/fdata.2024.1505284]
18 Li ZL, Liang Y, Wang M, *et al*. Explainable and coherent complement recommendation based on large language models. Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. Boise: ACM, 2024. 4678–4685. [doi: 10.1145/3627673.3680028]
19 Liu SJ, Ding RX, Lu WH, *et al*. Coherency improved explainable recommendation via large language model. Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia: AAAI Press, 2025. 12201–12209.
20 Li ZH, Zou LX, Ma C, *et al*. Efficient and explainable sequential recommendation with language model. Information Processing & Management, 2025, 62(4): 104122. [doi: 10.1016/j.ipm.2025.104122]
21 Kim J, Kim H, Cho H, *et al*. Review-driven personalized preference reasoning with large language models for recommendation. Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. Padua: ACM, 2025. 1697–1706.
22 Zhang LY, Ling WY, Daizhou SW, *et al*. HDRec: Hierarchical distillation for enhanced LLM-based recommendation systems. Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad: IEEE, 2025. 1–5.
23 Zeng ZY, Li DY, Yang YQ. Towards explainable doctor recommendation with large language models. arXiv:2503.02298, 2025.
24 Gao YF, Sheng T, Xiang YL, *et al*. Chat-REC: Towards interactive and explainable LLMs-augmented recommender system. arXiv:2303.14524, 2023.
25 da Silva EA, Marinho LB, de Moura ES, *et al*. A tool for explainable pension fund recommendations using large language models. Proceedings of the 18th ACM Conference on Recommender Systems. Bari: ACM, 2024. 1184–1186.
26 Yang MY, Zhu MY, Wang Y, *et al*. Fine-tuning large language model based explainable recommendation with explainable quality reward. Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver: AAAI Press, 2024. 9250–9259. [doi: 10.1609/aaai.v38i8.28777]
27 Huang X, Lian JX, Lei YX, *et al*. Recommender AI agent: Integrating large language models for interactive recommendations. ACM Transactions on Information Systems, 2025, 43(4): 96. [doi: 10.1145/3731446]
28 Dong ZK, Liu XL, Chen B, *et al*. MuseChat: A conversational music recommendation system for videos. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 12775–12785.
29 Ma QY, Ren XB, Huang C. XRec: Large language models for explainable recommendation. Proceedings of the 2024 Findings of the Association for Computational Linguistics: EMNLP 2024. Miami: Association for Computational Linguistics, 2024. 391–402. [doi: 10.18653/v1/2024.findings-emnlp.22]
30 Lei YX, Lian JX, Yao J, *et al*. RecExplainer: Aligning large language models for explaining recommendation models. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Barcelona: ACM, 2024. 1530–1541. [doi: 10.1145/3637528.3671802]
31 Li YH, Zhang XN, Luo LH, *et al*. G-Refer: Graph retrieval-augmented large language model for explainable recommendation. Proceedings of the 2025 ACM Web Conference. Barcelona: ACM, 2025. 240–251. [doi: 10.1145/3696410.3714727]
32 Wang YY, Li P, Chen MM. The blessing of reasoning: LLM-based contrastive explanations in black-box recommender systems. arXiv:2502.16759, 2025.
33 Zeng J, Tao HJ, Wen JH, *et al*. Explainable next POI recommendation based on spatial-temporal disentanglement representation and pseudo profile generation. Knowledge-based Systems, 2025, 309: 112784. [doi: 10.1016/j.knosys.2024.112784]
34 Lubos S, Tran TNT, Felfernig A, *et al*. LLM-generated explanations for recommender systems. Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization. Cagliari: ACM, 2024. 276–285.
35 Gao S, Wang YF, Fang JB, *et al*. DRE: Generating recommendation explanations by aligning large language models at data-level. arXiv:2404.06311, 2024.
36 Lin CS, Tsai CN, Su ST, *et al*. Predictive prompts with joint training of large language models for explainable recommendation. Mathematics, 2023, 11(20): 4230. [doi: 10.3390/math11204230]
37 Rahdari B, Ding H, Fan ZW, *et al*. Logic-Scaffolding: Personalized aspect-instructed recommendation explanation generation using LLMs. Proceedings of the 17th ACM International Conference on Web Search and Data Mining.

Merida: ACM, 2024. 1078–1081. [doi: 10.1145/3616855.3635689]

38 Peng YC, Chen H, Lin CS, *et al*. Uncertainty-aware explainable recommendation with large language models. Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN). Yokohama: IEEE, 2024. 1–8. [doi: 10.1109/IJCNN60899.2024.10651104]

39 Abu-Rasheed H, Weber C, Fathi M. Knowledge graphs as context sources for LLM-based explanations of learning recommendations. Proceedings of the 2024 IEEE Global Engineering Education Conference (EDUCON). Kos Island: IEEE, 2024. 1–5.

40 Xie ZH, Singh S, McAuley J, *et al*. Factual and informative review generation for explainable recommendation. Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: AAAI Press, 2023. 13816–13824.

41 Kovacs A, Meteier Q, Angelini L, *et al*. Explanation modalities for a recommender system optimizing energy management in a solar-powered smart home. Proceedings of the 30th International Conference on Intelligent User Interfaces. Cagliari: ACM, 2025. 43–47.

42 Ashaduzzaman M, Nguyen T, Tsai CH. Explaining social recommendations using large language models. In: de la Iglesia DH, de Paz Santana JF, López Rivero AJ, eds. New Trends in Disruptive Technologies, Tech Ethics, and Artificial Intelligence. Cham: Springer, 2024. 73–84.

43 Li L, Zhang YF, Chen L. Personalized prompt learning for explainable recommendation. ACM Transactions on Information Systems, 2023, 41(4): 103.

44 Chun HW, Ong RK, Khong AWH. Reasonable sense of direction: Making course recommendations understandable with LLMs. Proceedings of the 67th IEEE International Midwest Symposium on Circuits and Systems (MWSCAS). Springfield: IEEE, 2024. 1408–1412. [doi: 10.1109/MWSCAS60917.2024.10658914]

45 Zhao HK, Zheng SM, Wu LK, *et al*. LANE: Logic alignment of non-tuning large language models and online recommendation systems for explainable reason generation. arXiv:2407.02833, 2024.

46 Hulstijn J, Tchappi I, Najjar A, *et al*. Metrics for evaluating explainable recommender systems. Proceedings of the 5th International Workshop on Explainable and Transparent AI and Multi-agent Systems. London: Springer, 2023. 212–230. [doi: 10.1007/978-3-031-40878-6_12]

47 Sun PJ, Wu L, Zhang K, *et al*. An unsupervised aspect-aware recommendation model with explanation text generation. ACM Transactions on Information Systems, 2022, 40(3): 63. [doi: 10.1145/3483611]

48 Papineni K, Roukos S, Ward T, *et al*. BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia: Association for Computational Linguistics, 2002. 311–318.

49 Lin CY. ROUGE: A package for automatic evaluation of summaries. Proceedings of the 2004 Text Summarization Branches Out. Barcelona: Association for Computational Linguistics, 2004. 74–81.

50 Yu XL, Chen Z, Lu YB. Harnessing LLMs for temporal data—A study on explainable financial time series forecasting. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track. Singapore: Association for Computational Linguistics, 2023. 739–753. [doi: 10.18653/v1/2023.emnlp-industry.69]

51 Li L, Zhang YF, Chen L. Generate neural template explanations for recommendation. Proceedings of the 29th ACM International Conference on Information & Knowledge Management. ACM, 2020. 755–764.

52 Li L, Zhang YF, Chen L. Personalized transformer for explainable recommendation. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021. 4947–4957. [doi: 10.18653/v1/2021.acl-long.383]

53 Ariza-Casabona A, Boratto L, Salamó M. A comparative analysis of text-based explainable recommender systems. Proceedings of the 18th ACM Conference on Recommender Systems. Bari: ACM, 2024. 105–115. [doi: 10.1145/3640457.3688069]

54 Zhang XY, Li YS, Wang JY, *et al*. Large language models as evaluators for recommendation explanations. Proceedings of the 18th ACM Conference on Recommender Systems. Bari: ACM, 2024. 33–42. [doi: 10.1145/3640457.3688075]

55 Fu JL, Ng SK, Jiang ZB, *et al*. GPTScore: Evaluate as you desire. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Mexico City: Association for Computational Linguistics, 2024. 6556–6576.

56 Zhang TY, Kishore V, Wu F, *et al*. BERTScore: Evaluating text generation with BERT. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: OpenReview.net, 2020. 1–43.

57 Yuan WZ, Neubig G, Liu PF. BARTScore: Evaluating

generated text as text generation. Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates Inc., 2021. 27263–27277.

58 Sellam T, Das D, Parikh A. BLEURT: Learning robust metrics for text generation. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 7881–7892. [doi: 10.18653/v1/2020.acl-main.704]

59 Shimizu R, Wada T, Wang Y, *et al*. Disentangling likes and dislikes in personalized generative explainable recommendation. Proceedings of the 2025 ACM Web Conference 2025. Sydney: ACM, 2025. 4793–4809.

60 Chen H, Wang B, Yang K, *et al*. Persuasive-oriented explanation generation and evaluation of personalized recommendation. Proceedings of the 8th International Conference on Big Data and Computing. Shenzhen: ACM, 2023. 74–80.

61 Hasan E, Rahman M, Ding C, *et al*. Review-based recommender systems: A survey of approaches, challenges and future perspectives. ACM Computing Surveys, 2024. [doi: 10.1145/3742421]

62 Jiang Y, Qiu RH, Zhang Y, *et al*. Balanced and explainable social media analysis for public health with large language models. Proceedings of the 34th Australasian Database Conference on Databases Theory and Applications. Melbourne: Springer, 2023. 73–86.

63 Park J, Kim S, Lee S. A user preference and intent extraction framework for explainable conversational recommender systems. Proceedings of the 2023 ACM SIGCHI Symposium on Engineering Interactive Computing Systems. Swansea: ACM, 2023. 16–23. [doi: 10.1145/3596454.3597178]

64 Abu-Rasheed H, Weber C, Fathi M. Experimental interface for multimodal and large language model based explanations of educational recommender systems. arXiv:2402.07910, 2024.

65 Hendrawan RA, Brusilovsky P, Lekshmi Narayanan AB, *et al*. Explanations in open user models for personalized information exploration. Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization. Cagliari: ACM, 2024. 256–263.

66 Li ZX, Yazdanpanah V, Wang JD, *et al*. TutorLLM: Customizing learning recommendations with knowledge tracing and retrieval-augmented generation. arXiv:2502.15709, 2025.

67 Ma BX, Khan MAZ, Yang TY, *et al*. How good are large language models for course recommendation in MOOCs? arXiv:2504.08208, 2025.

68 Afanasieva TV, Platov PV, Komolov AV, *et al*. Leveraging ChatGPT and long short-term memory in recommender algorithm for self-management of cardiovascular risk factors. Mathematics,2024,12(16):2582.[doi:10.3390/math12162582]

69 Wali T, Bolatbekov A, Maimaitijiang E, *et al*. A novel recommender framework with chatbot to stratify heart attack risk. Discover Medicine, 2024, 1(1): 161. [doi: 10.1007/s44337-024-00174-9]

70 Xiang YF, Yu HY, Gong YL, *et al*. Text understanding and generation using transformer models for intelligent e-commerce recommendations. Proceedings of the 9th International Symposium on Advances in Electrical, Electronics, and Computer Engineering (ISAEECE 2024). Changchun: SPIE, 2024. 1329156. [doi: 10.1117/12.3034062]

71 Wang MH, Guo YC, Zhang DF, *et al*. Enabling explainable recommendation in e-commerce with LLM-powered product knowledge graph. arXiv:2412.01837, 2024.

72 Liu YX, Zhang WN, Chen YF, *et al*. Conversational recommender system and large language model are made for each other in e-commerce pre-sales dialogue. Proceedings of the 2023 Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics, 2023. 9587–9605.

73 Xu XN, Wu YC, Liang PH, *et al*. Emerging synergies between large language models and machine learning in ecommerce recommendations. arXiv:2403.02760, 2024.

74 Liu DR, Yang BM, Du HH, *et al*. RecPrompt: A self-tuning prompting framework for news recommendation using large language models. Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. Boise: ACM, 2024. 3902–3906.

75 Gao S, Fang JB, Tu Q, *et al*. Generative news recommendation. Proceedings of the 2024 ACM Web Conference. Singapore: ACM, 2024. 3444–3453. [doi: 10.1145/3589334.3645448]

(校对责编: 张重毅)