探索 ChatGPT 推荐结果中的多样性、新颖性与流行度偏差

达里奥·迪帕尔马[1, *]，乔瓦尼·玛丽亚·比安科菲奥雷[1, *]，维托·沃尔特阿内利[1]，费德卢乔·纳尔杜奇[1]和托马索·迪诺亚[1]

[1]意大利巴里理工大学

摘要

ChatGPT 已成为一种用途广泛的工具，在多个领域展现出强大的能力。鉴于这些成果，推荐系统（RSs）领域的研究人员开始探索其在推荐场景中的应用，重点主要放在准确率上。尽管 ChatGPT 集成到推荐系统中已受到广泛关注，但对其在各个维度上的表现进行全面分析仍鲜有研究。具体而言，ChatGPT 提供多样化、新颖推荐结果的能力，以及探索流行度偏差等潜在偏差的能力，尚未得到充分考察。随着这类模型的应用不断扩大，理解这些方面对于提升用户满意度和实现长期个性化至关重要。

本研究通过评估 ChatGPT-3.5 和 ChatGPT-4 在多样性、新颖性及流行度偏差方面的能力，考察了 ChatGPT 提供的推荐结果。我们在三个不同的数据集上评估这些模型，并考察它们在 Top-N 推荐和冷启动场景中的表现。研究结果表明，ChatGPT-4 的表现可与传统推荐系统相媲美，甚至更胜一筹，并展现出在推荐结果中平衡新颖性与多样性的能力。此外，在冷启动场景下，ChatGPT 模型在准确率和新颖性方面均表现出色，表明它们对新用户尤其有益。本研究揭示了 ChatGPT 推荐结果的优势与局限，为这些模型提供超越准确率导向指标的推荐能力带来了新的视角。

关键词

ChatGPT、推荐系统（RSs）、大语言模型（LLMs）、多样性、新颖性、流行度偏差、冷启动

1. 引言

推荐系统（RSs） [1] 长期以来一直通过预测用户偏好并提供个性化内容，帮助用户在网络上发现有价值的信息。随着时间推移，这些系统已从矩阵分解方法发展到现代架构；现代架构拓展了最先进的深度学习模型 [2], ，这些模型最初是为时间序列预测 [3, 4],、自然语言处理 [5, 6, 7, 8], 和计算机视觉 [9, 10]等其他领域开发的。尽管在提高准确率方面取得了显著进展，当前用户建模和个性化领域的研究日益强调超越准确率的视角，例如多样性、新颖性和流行度偏差。这些因素不仅会影响系统的整体效能，还会影响用户满意度 [11],、长期参与度 [12], 和公平性 [13]。

随着 ChatGPT 于 2022年11月 发布，大语言模型（LLM）开始重塑推荐结果的呈现方式。与依赖精心构建的结构化训练数据的传统推荐系统不同，大语言模型能够生成自由文本，并可借助其海量知识提供更细致的解释和更广泛的项目覆盖度。因此，研究界正开始尝试大语言模型驱动的推荐流程 [14, 15, 16], ，并在提升推荐准确率方面取得了显著成果 [17, 18, 19, 20]。然而，现有关于基于ChatGPT的推荐系统的研究大多着眼于提升准确率，却忽视了对现实世界影响至关重要的超越准确率维度 [19, 21]。

GENNEXT@SIGIR'25：第一届下一代信息检索与推荐系统研讨会：语言智能体、生成模型与对话式人工智能，*2025*年*7*月*17*日，意大利帕多瓦
*通讯作者。 d.dipalma2@phd.poliba.it (D. D. Palma); giovannimaria.biancofiore@poliba.it (G. M. Biancofiore); vitowalter.anelli@poliba.it (V. W. Anelli); fedelucio.narducci@poliba.it (F. Narducci); tommaso.dinoia@poliba.it (T.D. Noia)
© 2025年本文作者版权所有。允许依据知识共享署名 4.0 国际（CC BY 4.0）使用。

忽视 ChatGPT 在超越准确率方面的表现，会让研究人员面对一个黑箱，难以判断它是否过度推荐热门物品、降低新颖性，或提供多样性较低的推荐结果，而这些都可能对用户满意度和长期个性化目标产生负面影响。早期研究主要关注如何利用 ChatGPT 对推荐结果进行重排序 [22],，另一些研究则开始探讨生成结果的惊喜性 [23] ，或研究 ChatGPT 如何生成推荐结果,以及其输出是否更接近基于内容的方法还是协同过滤方法 [24]。只有 Deldjoo [25] 研究了基于ChatGPT的推荐系统中的偏差，特别关注提供者公平性。尽管已有少数研究开始考察种族、性别和宗教等敏感属性相关的潜在偏差 [26], ，但 ChatGPT 中推荐多样性、新颖性和流行度偏差等方面仍基本未获探索。弥补这些空白对于确保个性化技术兼具有效性与公平性至关重要。

为此，我们分析 ChatGPT 的推荐行为，重点考察 ChatGPT-3.5 和 ChatGPT-4 在多项超越准确性的指标上的表现。具体而言，我们研究 ChatGPT 在正常条件下以及用户仅与少数物品交互的用户冷启动场景中，是否会生成多样且新颖的推荐结果，或表现出流行度偏差。我们的评估涵盖图书、电影和音乐三个不同领域，并以 Facebook Books [27, 28],MovieLens [29],和 Last.FM [30] 数据集作为基准，旨在回答以下研究问题（RQs）：
（研究问题1）ChatGPT 的推荐结果是否具有多样性？（研究问题2）ChatGPT 的推荐结果是否具有新颖性？（研究问题3）ChatGPT 是否受到流行度偏差的影响？（研究问题4）在准确率和超越准确率两个维度上，ChatGPT 在用户冷启动场景中的效果如何？

2. 相关工作

推荐系统中的多样性、新颖性和流行度偏差。为了提升用户参与度 [31], ，本文聚焦于推荐系统的超越准确性的度量，即多样性、新颖性和流行度偏差，并探究这些因素如何影响 ChatGPT 提供的推荐列表。

在推荐系统的发展过程中，研究人员曾意识到，仅依据准确性指标评估推荐结果是不够的。例如，Herlocker 等人 [32] 提出，推荐结果的表现应根据其对用户的有用性进行衡量。同样，Silveira 等人 [33], 在关于推荐系统评估的综述中指出，推荐结果可以依据效用、新颖性、多样性、意外性、惊喜性和覆盖度进行评估。Karimi、Jannach 和 Jugovac [34], 在对最先进的新闻推荐系统的评述中，确定多样性、新颖性和流行度是改善推荐结果最常见的质量因素。具体而言，多样性和新颖性通常被视为必须与预测准确性相平衡的质量因素 [35], ，也是推荐系统研究中讨论最多的超越准确性的目标 [36]。

随着人们对研究超越准确性指标的推荐系统日益感兴趣，越来越多的研究开始将这些指标作为改进目标。例如，Cheng 等人 [37] 和 Wu 等人 [38] 致力于构建不仅能够准确预测物品，还能使推荐结果达到较高多样性水平的推荐系统。Nakatsuji 等人 [39] 采用基于图的方法来识别新颖性更高的物品，而 Cai 等人 [40] 提出了一种缓解流行度偏差的方法。此外，Paparella 等人 [41]的研究强调了超越准确性评估推荐系统的重要性，并提出了一种多目标评估方法。

尽管许多研究使用超越准确率指标来评估和改进推荐系统，但现有文献缺少一个能够严格定义多样性、新颖性和偏差的统一框架，因此这些指标的含义模糊且彼此重叠。在我们的研究中，我们将这些概念定义如下：多样性是指推荐系统从目录中推荐多种物品的程度 [1], ，其依据是

[1]虽然我们承认多样性可以通过多种方式衡量，例如 Top-N 多样性（用户层面的差异 in

[42] 和 [43]。新颖性是指推荐项目让用户接触到相关体验的程度，而这些体验是用户不太可能独立发现的，其依据是 [44]。流行度偏差是指推荐系统倾向于优先推荐热门物品（即交互次数较多的物品），而非不太热门或小众物品的倾向，这与 [45]和 [42]一致。

基于ChatGPT的推荐。Gao 等人提出了首个将 ChatGPT 用于推荐的实例 [46], ，他们介绍了 ChatREC，这是一种 ChatGPT增强型推荐系统，可将推荐任务转化为与用户进行的交互式对话。作者提出了一种提示模板，将用户信息和用户-项目交互转换为向 ChatGPT 提出的查询。然而，该系统仅使用准确性指标（即召回率、精确率和归一化折损累计增益）进行评估。Manzoor 等人[47] 研究了 ChatGPT 在多轮对话式推荐背景设定下的表现，展示了其作为对话式推荐系统的潜力，并表明其优于传统方法。Hou 等人 [17] 将研究重点放在零样本设置下的 Chat GPT 上。他们通过设计专用的提示模板分析了 ChatGPT模型，并发现，在零样本推荐任务中，ChatGPT-4 的排序表现优于其他大语言模型。

Sanner 等人 [48] 研究了 ChatGPT 作为推荐系统执行 Top-N 推荐任务的能力，旨在找出生成相关推荐结果的最有效提示策略。作者得出结论，零样本设置能够生成最相关的推荐列表，其表现优于基于内容的基线方法。然而，他们的结论仅基于归一化折损累计增益这一评估指标，因此研究发现仅涉及推荐系统的一个维度。

Dai 等人 [19] 采用提示方法，通过评分预测、成对推荐和重排序策略，研究 ChatGPT 推荐物品的能力。他们在四个领域开展的实验表明，ChatGPT 具备推荐物品的能力。然而，这项研究仅从准确率角度考察了 ChatGPT 在 Top-N 推荐中的能力。

Li 等人 [49] 专注于将 ChatGPT 应用于图书推荐场景，设计了 BookGPT 以处理单物品推荐和评分预测任务。然而，由于作者仅关注图书领域，该研究未对 ChatGPT 在多个领域中的表现进行具有普适性的分析。

尽管上述研究均聚焦于利用 ChatGPT 提升推荐系统的性能，但它们主要基于准确性指标。为弥补这一空白，我们研究 Top-N 推荐任务，不再局限于准确率，而是从多样性、新颖性和流行度偏差等方面评估 ChatGPT 的表现，同时强调其在用户冷启动场景中的超越准确率能力。

3. 方法论

以下各节将讨论本研究所采用的方法论，概述用于从 ChatGPT 收集推荐结果的提示词设计，详细介绍实验所用的数据集，呈现用于比较的基线方法，并列出用于评估多样性、新颖性和流行度偏差的指标。

3.1. 提示词设计

GPT-3 [50] 的推出展示了大语言模型在获得清晰、针对特定任务的提示词后执行多种任务的能力，也说明了提示词如何影响模型的响应，并在塑造模型执行特定任务的表现方面发挥关键作用 [51]。

随着 ChatGPT 的广泛普及，关于提示工程的文献不断增多，研究范围从零样本和少样本等基础提示 [52] 扩展到思维链 [53], 、思维树 [54], 、Reflexion [55], 或图提示 [56]等更复杂的提示。我们手动设计了多种提示技术 [57], ，包括零样本、少样本、思维链和角色扮演

推荐列表）或时间多样性（随时间变化的多样性），我们关注总体多样性，因为它对物品曝光和长尾推广具有可衡量的影响。

（RP）提示，并参考 Xu 等人 [58] 和 Li 等人 [49], 的研究，确定最适合本研究的方法。

接下来，我们将介绍人工设计的提示词，并说明选择 RP 提示作为本研究主要技术的主要原因。具体而言，对于所有测试的提示词和每位用户，输入均由用户的历史记录构成，并以如下格式呈现为物品列表：$\{History\ of\ the\ User\}$：$Item_1$，$Item_2$，……，$Item_N$。

零样本提示 [59]。在零样本提示中，我们直接向 ChatGPT 提供用户的历史记录，并要求其给出 50 条推荐结果，如参考示例所示（见图 1）。然而，~71% 生成的列表包含的物品少于 50 个或含有重复条目，且 ~6% 未能正确执行任务。

零样本提示词

我喜欢{history of the user}，请给我 50 条推荐结果。

图 1： 一个零样本提示词示例，旨在根据用户的历史记录生成推荐结果。

少样本提示 [50]。在零样本提示之后，我们测试了少样本提示，通过提供少量推荐示例来帮助大语言模型更好地理解任务（见图 2）。尽管这些上下文示例减少了执行错误，~44% 生成的列表仍包含重复物品。

少样本提示

用户观看的电影：《肖申克的救赎》 用户评分：5（满分 5 分） 推荐电影：《绿里奇迹》 解释：这两部电影都是广受好评的剧情片，主题涉及困境中的希望与坚韧。由于用户给《肖申克的救赎》打了很高的分，他们可能也会喜欢《绿里奇迹》。

用户观看的电影：《盗梦空间》 用户评分：4 推荐电影：《星际穿越》 解释：这两部电影都是克里斯托弗·诺兰执导的科幻电影，以令人费解的情节和出色的视觉效果而闻名。喜欢《盗梦空间》的用户很可能也会喜欢《星际穿越》。

用户观看的电影：《黑暗骑士》 用户评分：5 推荐电影：

图 2： E 展示一个少样本提示示例：根据用户的使用情况生成附带解释的推荐结果
用户观看的电影
和评分。

思维链（CoT）提示 [53]。借助思维链，我们尝试将推荐任务分解为明确的步骤，迫使 ChatGPT 逐步进行推理。如图 3所示，我们明确规定了指令、用户的偏好以及识别最合适推荐结果的步骤。这种方法产生了过多的令牌，在生成 ~26 件物品后达到了上下文限制。

角色扮演提示 [60]。遵循 Xu 等人 [58] 和 Li 等人 [49], 的研究，我们还测试了角色扮演提示词，即让 ChatGPT 扮演推荐系统，根据用户的历史记录推荐物品（见图 4）。这一策略效果最佳，消除了重复推荐结果。

思维链提示

指令：根据用户的偏好推荐一本书。

用户的偏好：用户喜欢科幻，尤其是人物塑造出色、世界构建精妙的图书。他们之前喜欢弗兰克·赫伯特的《沙丘》和奥森·斯科特·卡德的《安德的游戏》。

思维链：为了推荐一本合适的图书，我需要考虑用户对科幻、出色的人物塑造和精妙世界构建的偏好。用户喜欢《沙丘》和《安德的游戏》，这表明他们欣赏复杂的情节和细致的背景设定。基于这些标准，我会挑选一本属于科幻类型、且以类似特质著称的图书。

1. 用户喜欢科幻，因此我会重点关注这一类型的图书。2. 用户看重人物塑造，因此我会考虑那些以人物刻画丰满著称的图书。3. 用户欣赏精妙的世界构建，因此我会寻找背景设定细致且引人入胜的图书。4. 鉴于他们喜欢《沙丘》和《安德的游戏》，我会寻找情节复杂且广受好评的图书。

推荐结果：“《黑暗的左手》”，作者厄休拉·勒古恩 n

解释：“《黑暗的左手》”是一部著名的科幻小说，以其细致入微的世界构建以及对复杂社会与文化问题的探讨而闻名。厄休拉·勒古恩对人物的塑造十分出色，故事以盖森星为背景，带来丰富而身临其境的阅读体验。这本书应该很符合用户的偏好，以及用户此前对《沙丘》和《安德的游戏》的喜爱。

图 3：用于书籍推荐的思维链提示示例，其中融入了用户偏好，并通过逐步推理得出推荐结果。

角色扮演式推荐提示

给定一位用户，请以推荐系统的身份，只提供排名前 50 的推荐结果名称。你知道用户喜欢以下物品：{history of the user}

图 4：E 角色扮演式推荐提示的示例，旨在生成一个包含 50 条推荐结果的排序列表 n-，基于用户的历史记录 基于用户的历史记录。

测试了 30 个手工设计的提示词并参考角色扮演提示相关研究后 [58, 61]，我们选择了这种方法，因为它能够减少重复内容和令牌用量。在此设置中，ChatGPT 充当推荐系统，根据用户的历史记录生成 50 条推荐结果（见图。4）。

3.2. 实验设置

本节介绍实验设置，包括用于评估 ChatGPT 推荐结果超越准确性的性能的数据集、基线方法和指标，重点关注多样性、新颖性和流行度偏差。

数据集。我们在三个知名推荐数据集上评估了 ChatGPT，分别是 Movie- Lens100k [29], Last.FM [30], 和 Facebook Books [2]。为提高数据质量，我们采用了迭代式 10-核心过滤策略 [62], 仅保留交互次数至少为十次的用户和物品。表 1

[2]https://2015.eswc-conferences.org/program/semwebeval.html

表1 $k-core \geq 10$预处理后的数据集统计信息。

数据集	交互	用户	物品	稀疏度	内容
MovieLens	42,456	603	1,862	96.22%	类型
Last.FM	49,171	1,797	1,507	98.18%	类型
FB Books	13,117	1,398	2,234	99.58%	类型、作者

表2超越准确率指标概览

方面	指标	描述
多样性	物品变异系数	物品覆盖率（物品变异系数）衡量出现的物品数量在用户的前-n 项推荐结果中，确保提供更广泛的内容选择。
	Gini	基尼指数：一种用于表示分布不平等程度的统计离散程度度量。基尼指数的取值范围为 0 到 1，其中数值越高表示更大的不平等。推荐结果的集中程度，例如集中于热门物品 [42]。我们报告 1 –基尼系数，数值越高表示推荐结果越不集-中 .
新颖性	EFD	预期自由发现（EFD）：一种基于逆向集合频率的新颖性度量，用于表达算法的推荐相关长尾项目的能力。推荐此类项目可以让用户接触到不那么显眼且独特的内容，从而丰富用户体验 [44]。
	EPC	预期流行度补数（EPC）：该指标量化“现在被看到的未见过物品数量”，促进此前未知内容的发现，并支持用户探索 [44]。
流行度偏差	APLT	长尾物品平均流行度（APLT）：衡量前 n 条推荐中的长尾物品平均流行度，确保突出较少见的非主流物品 [45, 63]。
	ARP	基于评分的平均流行度（ARP）：计算流行度推荐列表中物品的流行度，根据其数量计算每件物品在训练数据中的交互次数。通过考虑物品流行度，该指标有助于平衡推荐，避免热门物品过度曝光，从而应对偏差，并确保不同用户偏好得到更公平的分布用户偏好 [42]。

展示了预处理后的数据集统计信息。

基线模型。 为了衡量 ChatGPT 的有效性，我们通过实验将其表现与三个类别中的最先进基线方法进行比较：非个性化、协同过滤和基于内容的过滤方法。为确保公平比较，我们使用 Elliot 框架训练基线方法并优化其超参数 [64], ，并按照所有未评分项目评估协议 [65, 66]将数据集划分为 80% 的训练集和 20% 的测试集。实验所用代码已公开发布于：https://github.com/sisinflab/beyond-accuracy-recsys-chatgpt。下文按推荐类别介绍这些基线方法。

非个性化。 随机和 Most Popular 分别返回随机推荐结果和最受欢迎的推荐结果，作为参考点。

协同过滤。 为了比较 ChatGPT 推荐结果对超越准确率指标的影响，我们选取了以下协同过滤方法，每种方法各有侧重。具体而言，我们选取 RP^3_β [67]和 LightGCN [68]，因为它们已证明能够

在保持多样性的同时维持准确率 [69, 68]。选择 ItemKNN [70], UserKNN [71], 和 EASE^R [72], 是因为它们侧重于相关性和个性化 [72, 70]。最后，我们纳入 MF2020 [73] 和 NeuMF [74] ，以在模型复杂度与有效性之间取得平衡。 基于内容的过滤。 我们进一步将基于内容的模型纳入比较，这类模型优先采用显式特征表示，与协同模型形成有意义的对照。这样便可根据数据集选择最合适的模型类型来评估 ChatGPT。具体而言，我们纳入了 VSM [75] ，该方法将物品表示为高维空间中的向量，每个维度对应一个特征；还纳入了 AttributeItemKNN 和属性用户 KNN [76], ，它们依靠经词频-逆文档频率加权的属性向量计算相似度并生成推荐结果。

确保推荐结果的一致性。 ChatGPT模型仅根据提示中提供的用户画像生成推荐结果，不受预定义数据集的限制。因此，它们可能会产生幻觉 [77, 78, 79]，或推荐参考数据集中不存在的真实物品，从而导致物品名称不一致以及评估结果出现偏差。

为解决这一问题，我们采用后处理流程，利用格式塔模式匹配 [80]在数据集中识别最相似的匹配项，并接受相似度得分高于90%的物品（经经验确定）。无法匹配的物品会被标记为外部项目，这些项目源自大语言模型的预训练知识，并在评估中排除；为确保与传统推荐系统进行公平比较，我们仅选取目录内的物品。

由于这一最后步骤可能影响评估结果，我们确认所有推荐列表中的目录外项目始终出现在前10名之后，从而确保对排名敏感的指标不受影响，并保持评估的有效性。在我们的配置中，ChatGPT仅在排名第23位之后才会列出这些项目；图书、音乐和电影中目录外项目的数量分别为2,740、870和234。

最后，为了确保公平比较，我们按照推荐领域的广泛惯例，在截断位置为 10（即向每位用户提供前十名推荐）的情况下评估所有模型和 ChatGPT 的结果 [81, 73, 74]。**评估指标**。虽然我们的重点是 ChatGPT 推荐结果中超越准确率的方面，但纳入准确性指标以评估推荐结果是否与用户相关也很重要。为此，我们使用两种标准指标：精确率和召回率 [1, 82]。精确率和召回率的值越高，表示推荐系统提供的相关物品越多。此外，我们使用归一化折损累计增益（ nDCG） [83], 评估推荐结果的排序质量；该指标值越高，表示推荐列表的质量越好。

对于 超越准确率 指标，我们选取了一组指标来评估多样性、新颖性和受欢迎度偏差。所考察的具体指标详见表 2。

4. 实验结果

4.1. ChatGPT超越准确性的推荐性能

本节讨论表格3中的实证发现，重点关注（研究问题1）ChatGPT推荐结果的多样性、（研究问题2）其新颖性，以及（研究问题3）ChatGPT受流行度偏差影响的程度。评估涵盖三个数据集：Facebook Books、Last.FM 和 MovieLens，并将 ChatGPT-3.5 和 ChatGPT-4 与协同过滤和基于内容的过滤基线方法进行比较。表中标出了具有统计显著性的差异（在 $p < 0.05$下进行配对t检验）。

初步准确性分析。在考察多样性、新颖性和流行度偏差之前，我们首先验证 ChatGPT 的推荐结果是否实现了提供相关物品这一首要目标。我们使用归一化折损累计增益、召回率和精确率作为标准准确性指标。这些指标的值越高，推荐效果越好。

总体而言，ChatGPT 在推荐场景中展现出了相当的准确率。具体而言，在 Facebook Books 上，ChatGPT-4 的整体归一化折损累计增益最高（0.0932），显著优于表现最佳的基线模型 AttributeItemKNN（0.0479），也优于 ChatGPT-3.5（0.0668）。召回率和

精确率呈现出与归一化折损累计增益类似的趋势。

对于 Last.FM，尽管 ChatGPT-4（nDCG = 0.2832）未能超越表现最佳的协同过滤（CF）方法（RP^3_β: 0.3147），但仍跻身表现最出色的算法之列。ChatGPT-3.5 的表现不及 ChatGPT-4，但仍优于一些基线方法（例如 $EASE^R$、AttributeItemKNN）。

对于 MovieLens，尽管 ChatGPT-4 在所有准确性指标上的表现均优于 ChatGPT-3.5，nDCG 从 0.1475 提升至 0.1815，精确率从 0.1120 提升至 0.1551，但某些协同过滤算法（例如 RP^3_β: 0.2827 nDCG、0.2708 精确率）取得了显著更高的分数。尽管如此，ChatGPT 的准确率仍明显高于 VSM（nDCG 为 0.0174）和 AttributeItemKNN（nDCG 为 0.0326）等某些方法。

这些结果表明，ChatGPT-3.5 和 ChatGPT-4 在准确性指标上均取得了有效且合理的表现。这项初步评估确保后续对多样性、新颖性和流行度偏差的分析，是基于已达到准确率标准的推荐结果进行的。在接下来的章节中，我们将按照研究问题（RQs）展开分析。

（研究问题1。）ChatGPT 的推荐结果是否具有多样性？我们使用基尼系数和物品覆盖率（ItemCV）来评估多样性。基尼系数越低，表示推荐越集中于特定物品；覆盖率越高，则表示目录中有更多物品被推荐。

Facebook Books（表3）。ChatGPT-4 的基尼系数为 0.1050，物品覆盖率为 1,004，在这两项指标上均优于 ChatGPT-3.5（基尼系数 = 0.0713，物品覆盖率 = 853）。尽管一些基线方法（如 ItemKNN，基尼系数 = 0.5293，物品覆盖率 = 2,141）的多样性得分仍然更高，但总体而言，ChatGPT-4 和 ChatGPT-3.5 的排名均高于 MostPop 和 EASE 等基线方法^R。从物品覆盖率和基尼系数来看，ChatGPT 模型对特定物品的推荐高度集中，同时覆盖了近一半的物品总量（2,234 件中的 1,004 件）。

Last.FM（表3）。呈现出类似的趋势：ChatGPT-4 的基尼系数（0.2023）高于 ChatGPT-3.5（0.1927），表明推荐结果对特定物品的集中度较低。此外，GPT-4 覆盖了 1,507 件物品中的 944 件，而旨在权衡多样性与准确率的 RP^3_β覆盖度为 831，基尼系数为 0.1441，展现了 ChatGPT 推荐多样化物品的强大能力。

MovieLens（表3）。ChatGPT-4 的基尼系数为 0.0853，略优于 ChatGPT-3.5（0.0851）。不过，其项目覆盖率为 1,862 件物品中的 553 件，低于 RP 等方法$^3_\beta$（基尼系数为 = 0.1230，物品变异系数为 = 744）。这些结果表明，尽管 ChatGPT 的多样性得分低于某些基线方法，但在该数据集上的多样性得分仍具有可比性。

总结（研究问题1）。*ChatGPT* 的推荐结果在 *Facebook Books* 和 *Last.FM* 上具有中等多样性，而在 *MovieLens* 上的多样性有限；*GPT-4* 的表现始终优于 *GPT-3.5*。虽然其表现不及多样性最高的基线方法，但多样性优于某些协同过滤和基于内容的过滤方法。

（研究问题2。）ChatGPT 的推荐结果是否新颖？新颖性通过 EPC（预期流行度补数）和 EFD（预期自由发现）来衡量，两者的解读方式均为：数值越高，推荐结果越新颖。

Facebook Books（表3）。ChatGPT-4 的新颖性相对较高（EPC=0.0353，EFD=0.3486），超过了大多数基线方法，包括 ChatGPT-3.5（EPC=0.0250，EFD=0.2480），在这些指标上甚至超过了所有协同过滤和基于内容的过滤算法。

Last.FM（表3）。两个 ChatGPT 版本在 EPC 和 EFD 上的排名均高于平均水平，而 CF 和 CBF 方法（例如 RP^3_β：EPC=0.2110，EFD=1.9970，VSM：EPC=0.1593，EFD=1.4845 ）的表现处于相近水平。不过，ChatGPT-4（EPC 为 0.1918，EFD 为 1.8663）与 ChatGPT-3.5（EPC 为 0.1680，EFD 为 1.6436）之间的差异表明，GPT-4 更擅长推荐不那么主流的物品。

表3
三个数据集（Facebook Books、Last.FM 和 MovieLens）的综合结果。箭头表示更理想的指标值（↑数值越高，↓数值越低）。最佳值以粗体显示，次优值带下划线。结果按归一化折损累计增益排序。除非标有 *（ChatGPT-3.5）或 †（ChatGPT-4），否则基线方法的结果均具有统计显著性（配对t检验，$p < 0.05$）。“最佳协同过滤”和“基于内容的过滤”分别表示每个数据集上（按归一化折损累计增益计）表现最佳的协同过滤和基于内容的过滤基线方法。

Facebook Books

模型	准确率			多样性		新颖性		流行度偏差	
	归一化折损累计增益 ↑	召回率 ↑	精确率 ↑	基尼系数 ↑	物品变异系数 ↑	EPC ↑	EFD ↑	APLT ↑	ARP ↓
随机	0.0019	0.0034	0.0008	0.7753	2230	0.0007	0.0078	0.6874	5.7186
MostPop	0.0091	0.0137	0.0033	0.0045	17	0.0031	0.0228	0.0000	138.3632
LightGCN	0.0105	0.0171	0.0038	0.0053	112	0.0035	0.0269	0.0132	134.0763
NeuMF	0.0167	0.0243	0.0057	0.3336	1563	0.0065	0.0661	0.2444†	16.0072
EASE^R	0.0188	0.0313	0.0071	0.0111	228	0.0066	0.0547	0.0032	125.2026
ItemKNN	0.0288	0.0408	0.0086	**0.5293**	**2141**	0.0104	0.1099	**0.5974**	24.9652
MF2020	0.0317	0.0592	0.0133	0.0044	15	0.0116	0.0953	0.0000	114.0167
UserKNN	0.0320	0.0468	0.0098	0.1564	1372	0.0115	0.1065	0.0852	55.2988
RP$^3_\beta$ 最佳协同过滤	0.0379	0.0568	0.0120	0.3063	1888	0.0138	0.1357	0.3308	44.1225*
属性用户KNN	0.0402	0.0593	0.0133	0.0918	945	0.0152	0.1414	0.0466	64.2887
VSM	0.0458	0.0785	0.0172	0.2478	1389	0.0173	0.1913	0.5761	7.3705
AttributeItemKNN 最佳基于内容的过滤	0.0479	0.0705	0.0155	0.2824	1510	0.0182	0.2019	0.5879	**7.1801**
ChatGPT-3.5	0.0668	0.0936	0.0205	0.0713	853	0.0250	0.2480	0.1870	46.3236
ChatGPT-4	**0.0932**	**0.1283**	**0.0283**	0.1050	1004	**0.0353**	**0.3486**	0.2424	40.0319

Last.FM

模型	准确率			多样性		新颖性		流行度偏差	
	归一化折损累计增益 ↑	召回率 ↑	精确率 ↑	基尼系数 ↑	物品变异系数 ↑	EPC ↑	EFD ↑	APLT ↑	ARP ↓
随机	0.0044	0.0068	0.0052	0.8398	1507	0.0045	0.0478	0.5678	31.6844
NeuMF	0.1005	0.1133	0.0860	**0.5049**	**1492**	0.0804	0.7848	0.2418	**77.5480**
MostPop	0.1009	0.0895	0.0740	0.0081	27	0.0662	0.5907	0.0000	348.3308
LightGCN	0.1408	0.1329	0.1060	0.1114	635	0.1013	0.9372	0.2063	135.0381
AttributeItemKNN	0.2233	0.2013*	0.1481*	0.3854	1411	0.1584	1.5710	**0.3043**	87.8647
EASE^R	0.2278	0.1949*	0.1509	0.0331	283	0.1517	1.3761	0.0088	247.6099
VSM	0.2451*	0.2021*	0.1511	0.0826	653	0.1593	1.4845	0.0585	177.9949
属性用户KNN 最佳基于内容的过滤	0.2795†	0.2364	0.1818†	0.1724	923	0.1947†	1.8297†	0.0895	134.5766
UserKNN	0.2983	0.2538	0.1912	0.1491	846	0.2030	1.9060†	0.0550	152.7412
ItemKNN	0.3013	0.2595	0.1925	0.1634	962	0.2080	1.9854	0.1146	152.4739
MF2020	0.3097	0.2576	**0.1986**	0.0908	460	**0.2116**	1.9571	0.0051	181.8922
RP$^3_\beta$ 最佳协同过滤	**0.3147**	**0.2634**	0.1957	0.1441	831	0.2110	**1.9970**	0.0678	153.0884
ChatGPT-3.5	0.2448	0.1964	0.1408	0.1927	952	0.1680	1.6436	0.1391	99.3311
ChatGPT-4	0.2832	0.2313	0.1680	0.2023	944	0.1918	1.8663	0.1267	102.1045

MovieLens

模型	准确率			多样性		新颖性		流行度偏差	
	归一化折损累计增益 ↑	召回率 ↑	精确率 ↑	基尼系数 ↑	物品变异系数 ↑	EPC ↑	EFD ↑	APLT ↑	ARP ↓
随机	0.0087	0.0062	0.0129	0.6917	1776	0.0108	0.1230	0.5482	22.2227
VSM	0.0174	0.0099	0.0205	0.0529	409	0.0209	0.2305	0.3732	29.2857
AttributeItemKNN	0.0326	0.0220	0.0389	**0.3962**	**1395**	0.0375	0.4285	**0.5510**	**23.6326**
LightGCN	0.0411	0.0349	0.0500	0.3105	1136	0.0421	0.4637	0.3040	43.4723
NeuMF	0.1235	0.0999†	0.1324	0.2761	1172	0.1171*	1.2757*	0.0970	70.5342
MostPop	0.1488*	0.0841*	0.1424†	0.0083	40	0.1097	1.2750*	0.0000	182.4909
MF2020	0.2013	0.1298	0.1985	0.0173	94	0.1576†	1.7712	0.0000	162.5163
EASE^R	0.2076	0.1229	0.1872	0.0118	67	0.1522†	1.7352†	0.0000	173.2040
属性用户KNN 最佳基于内容的过滤	0.2152	0.1317	0.2045	0.0590	438	0.1743	1.9356	0.0117	127.1064
ItemKNN	0.2709	0.1819	0.2626	0.1036	666	0.2348	2.5547	0.0470	103.0248
UserKNN	0.2814	0.1769	0.2601	0.0589	428	0.2263	2.4958	0.0057	125.7174
RP$^3_\beta$ 最佳协同过滤	**0.2827**	**0.1898**	**0.2708**	0.1230	744	**0.2421**	**2.6613**	0.0643*†	100.4106
ChatGPT-3.5	0.1475	0.0807	0.1120	0.0851	591	0.1260	1.3981	0.0733†	90.7590
ChatGPT-4	0.1815	0.1109	0.1551	0.0853	553	0.1453	1.6010	0.0775*	95.7042

MovieLens（表3）。在 MovieLens 上，ChatGPT-4（0.1453 EPC、1.6010EFD）在 EPC 和 EFD 值方面优于 ChatGPT-3.5（0.1260 EPC、1.3981 EFD），其表现与其他方法相当（例如 NeuMF：0.1171 EPC、1.2767 EFD），但低于 RP$^3_\beta$（EPC 为 0.2421，EFD 为 2.6613），后者是最佳模型。

总结 y（研究问题2）。*ChatGPT* 在 *MovieLens* 上的推荐结果展现出高于平均水平的新颖性，且 *d*

在 *Facebook Books* 和 *Last.FM* 上具有较高的新颖性，*GPT-4* 的表现通常优于 *GPT-3.5*。结果表明，*ChatGPT* 也会根据用户的历史记录，为每位用户推荐新颖的物品。

(RQ3.) ChatGPT 是否受到流行度偏差的影响? 我们使用 APLT（长尾物品的平均流行度；数值越高，表明越倾向于推荐长尾（较不热门）物品）和 ARP（基于评分的平均流行度；数值越低，表示流行度偏差越小）来考察流行度偏差。

Facebook Books（表3）。ChatGPT-3.5 的推荐结果 APLT 为 = 0.1870，ARP 为 = 46；ChatGPT-4 则有所提升，APLT 为 = 0.2424，ARP 为 = 40。由于 APLT 更高且 ARP 更低，GPT-4 在推荐长尾和较不热门物品方面的能力优于 GPT-3.5。尽管这两个模型的表现仍远不及纯 MostPop 方法（ARP 为 = 138），但一些基线方法，例如 AttributeItemKNN（APLT 为 = 0.5879，ARP 为 = 7）和 VSM（APLT 为 = 0.5761，ARP 为 = 7），取得了更好的 APLT 和 ARP 值。

Last.FM（表3）。ChatGPT-3.5 的 APLT 为 0.1391，ARP 为 99；GPT-4 的 APLT 为 0.1267，ARP 为 102，在各模型中处于中等水平。这表明 GPT-4 覆盖的长尾比例较小，且倾向于推荐更热门的物品。尽管它优于某些基线方法，例如 RP^3_β（APLT = 0.0678，ARP = 153），但表现不及其他基线方法，例如 AttributeItemKNN（APLT = 0.3043，ARP = 87.8647）。

MovieLens（表3）。ChatGPT 在 GPT-3.5 和 GPT-4 上的 ARP 分别为 90 和 95，低于 MostPop（ARP = 182），但高于某些基于图的方法（例如 LightGCN：ARP = 43）或基于邻居的方法（例如 AttributeItemKNN：ARP = 23）。这表明其行为不像 MostPop 那样受流行度驱动，但仍受到热门物品的影响。APLT 也呈现出类似趋势，进一步表明 ChatGPT 不会推荐长尾物品，并存在一定程度的流行度偏差。

总结（RQ3）。尽管 *ChatGPT* 的数值与 *MostPop* 得出的结果相差甚远，但它仍表现出推荐热门物品、忽视长尾物品的倾向。具体而言，*GPT-4* 的 *ARP* 低于 *ChatGPT-3.5*，这表明它倾向于推荐不那么热门的物品。

总而言之，ChatGPT模型展现出强大的超越准确性的性能：在图书领域实现了新颖性与多样性的最佳平衡，在音乐领域取得了相近的结果，而在电影领域的表现则不尽理想。尽管它对热门物品有一定倾向，但与 MostPop 或其他受流行度偏差影响较大的基线方法相比，这种偏差要小得多。此外，GPT-3.5 到 GPT-4 在全部三个数据集上的改进，凸显了 GPT-4 在推荐方面的优势，尤其是在平衡超越准确性的权衡方面。

这些发现凸显了 ChatGPT 作为推荐系统的潜力，同时也指出了有待改进之处，尤其是进一步提升其在不同领域平衡相关性、多样性和新颖性的能力。

4.2. 用户冷启动场景

现在，我们来考察用户冷启动推荐。这里将其定义为每位用户最多提供十次交互的情形。表 4 展示了三个数据集（Facebook Books、Last.FM 和 MovieLens）上的结果，并将 ChatGPT-3.5 和 ChatGPT-4 与表现强劲的协同过滤（CF）和基于内容的过滤（CBF）基线方法进行比较。我们的核心问题是：

表4
使用 ChatGPT-3.5 和 ChatGPT-4 对比分析**用户冷启动**交互（每位用户最多十次交互）。箭头表示每项指标的较高（↑）或较低（↓）取值是否更理想。最佳值以粗体标出，次佳值加下划线。CF 和 CBF 分别代表协同过滤和基于内容的过滤推荐系统。除带有 *（ChatGPT-3.5 对应值）和 †（ChatGPT-4 对应值）标记的数值外，Facebook Books 基线方法的结果均具有统计显著性（配对t检验，$p < 0.05$）。最佳协同过滤和基于内容的过滤方法分别记为 Best-CF 和 CBF，其依据为归一化折损累计增益（nDCG）。

	模型	准确率			多样性		新颖性		流行度偏差	
		归一化折损累计增益 ↑	召回率 ↑	精确率 ↑	基尼系数 ↑	物品变异系数 ↑	EPC ↑	EFD ↑	APLT ↑	ARP ↓
Facebook Books	随机	0.0011	0.0018	0.0004	0.4315	1560	0.0004	0.0034†	0.6793	5.8468
	MostPop	0.0115*	0.0143	0.0029	0.0037	14	0.0031	0.0236	0.0000	139.1043
	AttributeItemKNN 基于内容的过滤	0.0335*	0.0500	0.0100*	0.1873	957	0.0119*	0.1283*	**0.5661**	**7.2114**
	RP^3_β CF	0.0346*	0.0500	0.0096	**0.1932**	**1044**	0.0115	0.1161*	0.3332	41.7332*
	ChatGPT-3.5	0.0487†	0.0779†	0.0152†	0.0538	445	0.0168†	0.1714†	0.2004	45.0357
	ChatGPT-4	**0.0538***	**0.0873***	**0.0171***	0.0846	597	**0.0186***	**0.1877***	0.2458	37.9698
Last.FM	随机	0.0000	0.0000	0.0000	0.2091	345	0.0000	0.0000	0.5184	34.3684
	MostPop	0.0529	0.0877	0.0211	0.0063	15	0.0167	0.1493	0.0000	363.5789
	属性用户KNN 基于内容的过滤	0.1724	0.2149	0.0605	0.1487	282	0.0705	0.8092	**0.5000**	**40.5553**
	RP^3_β CF	0.2389	0.3333	0.0895	0.1237	254	0.1043	1.1120	0.2605	108.0526
	ChatGPT-3.5	**0.2921**	0.3423	0.0946	0.1340	257	**0.1289**	**1.3989**	0.3081	75.6000
	ChatGPT-4	0.2791	**0.3465**	**0.0947**	**0.1513**	**283**	0.1204	1.3141	0.3526	69.8632
MovieLens	随机	0.0000	0.0000	0.0000	0.0683	129	0.0000	0.0000	0.5231	22.8000
	MostPop	0.0254	0.0385	0.0077	0.0049	12	0.0048	0.0566	0.0000	197.6769
	属性用户KNN 基于内容的过滤	0.0368	0.0513	0.0154	0.0429	100	0.0101	0.1080	0.1231	90.7769
	RP^3_β CF	0.0791	0.1026	0.0231	**0.0566**	**117**	0.0269	0.2878	**0.1846**	**65.1231**
	ChatGPT-3.5	0.1117	0.0897	0.0231	0.0333	87	0.0313	0.3384	0.0769	96.5615
	ChatGPT-4	**0.1405**	**0.1538**	**0.0385**	0.0349	89	**0.0455**	**0.4988**	0.0462	102.9385

研究问题4：在准确率及超越准确率的维度上，ChatGPT 在用户冷启动场景中的表现有多有效?

冷启动下的准确率。尽管用户交互有限，ChatGPT 的准确率仍具有竞争力，甚至优于传统基线。对于 Facebook Books，GPT-4 的归一化折损累计增益（ 0.0538 ）和召回率（ 0.0873 ）均高于包括 RP^3_β（归一化折损累计增益 = 0.0346）和 AttributeItemKNN（0.0335）在内的所有基线方法。GPT-3.5 也超过了这些基线方法，但表现略逊于 GPT-4。对于 Last.FM，ChatGPT 保持了稳健的表现（$nDCG \geq 0.2791$、$Recall \geq 0.3423$），大幅优于 MostPop（归一化折损累计增益 = 0.0529）和随机基线。尽管 RP^3_β 在归一化折损累计增益方面领先（0.2389），GPT-4 在召回率和精确率方面通常表现更佳。对于 MovieLens，GPT-4 达到了最高的归一化折损累计增益（0.1405），超过了 CF 和基于内容的过滤基线，而 ChatGPT-3.5（0.1117）的表现也具有竞争力。这些结果凸显了 ChatGPT 仅凭少量交互便能有效识别相关物品的能力。

冷启动中的超越准确率。

多样性。在全部三个数据集中，GPT-4 的基尼系数和项目覆盖率通常都高于 GPT-3.5（例如，在 Facebook Books 上，基尼系数从 0.0538 增至 0.0846），这表明 GPT-4 的推荐结果涵盖了范围更广的物品。尽管 RP^3_β 等基线方法在 MovieLens 和 Facebook Books 上实现了更高的覆盖度，GPT-4 在 Last.FM 上的表现最佳。

新颖性。在所有数据集中，ChatGPT 的 EPC 和 EFD 值均高于协同过滤和基于内容的过滤基线方法（例如，在 Facebook Books 上，GPT-4 的 EPC 为 = 0.0186，而 RP^3_β= 0.0115），这表明它倾向于推荐新颖的物品，而非依赖主流物品。

流行度偏差。与所有数据集中的基线方法相比，ChatGPT 对热门物品表现出中等程度的倾向。尽管如此，它仍远不及 MostPop（例如，Facebook Books 上的 $ARP \geq 139$），但与某些基线方法（如属性用户KNN）的表现相当，这表明仍有进一步采取缓解策略的空间。

总而言之，ChatGPT 在冷启动场景中表现出色，能够：（i）保持

较高的准确率，即使用户交互极少，其中 GPT-4 往往优于 GPT-3.5；（ii）在多样性、新颖性和流行度偏差之间取得平衡；（iii）相较于基线方法持续展现改进，彰显 ChatGPT 在有限交互下推断用户兴趣的能力。

5. 结论

在本研究中，我们探讨了 ChatGPT 推荐的多样性、新颖性和流行度偏差。研究结果表明，在 Facebook Books、Last.FM 和 MovieLens 数据集上，ChatGPT模型展现出出色的超越准确性的性能：在 Facebook Books 上实现了新颖性与多样性的最佳平衡，在 Last.FM 上表现相近，而在 MovieLens 上的结果则不够理想。

此外，我们还发现，虽然 ChatGPT 在新颖性与多样性之间实现了良好平衡，但也倾向于推荐热门物品，尤其是在 MovieLens 数据集中。

最后，我们将研究扩展至用户冷启动场景。在该场景下，尽管用户交互极少，ChatGPT 仍能保持较高的准确率，并在多样性、新颖性和流行度偏差之间取得平衡，同时持续优于基线方法，表现出很强的效果。

这些发现凸显了 ChatGPT 作为推荐系统所具备的超越准确性的能力。未来的研究将纳入更多数据集，以便在不同领域推广这些发现，并开展实验，将 ChatGPT 与 Gemini、LLaMA 和 DeepSeek 等其他大语言模型进行比较。

局限性

如今，大语言模型被用于增强推荐系统的能力。然而，这些模型通常在海量的互联网规模语料库上训练，其中可能包含用于基准测试的开放数据集中的部分数据。近期一项研究探讨了 MovieLens-1M [84]中的记忆现象，结果表明，GPT 和 LLaMA-3 等模型能够记住此类数据集，且模型越大，记忆率越高。例如，LLaMA-3.1 405B 的记忆率为 12.9%，GPT-4 的记忆率则为 80.76%。未来研究应重点探究推荐质量的提升与记忆能力之间的相关性。

参考文献

[1] F. Ricci、L. Rokach、B. Shapira（编），《推荐系统手册》，施普林格，美国，2022年。[2] S. Zhang、L. Yao、A. Sun、Y. Tay，基于深度学习的推荐系统：综述与新视角，《ACM计算调查》52（2019）5:1–5:38。[3] J. Chung、Ç. Gülçehre、K. Cho、Y. Bengio，门控循环神经网络在序列建模中的实证评估，CoRR abs/1412.3555（2014）。URL：http://arxiv.org/abs/1412.3555。 arXiv:1412.3555。[4] A. Gu、T. Dao，Mamba：利用选择性状态空间进行线性时间序列建模，CoRR abs/2312.00752（2023）。URL：https://doi.org/10.48550/arXiv.2312.00752。doi：10.48550/ARXIV.2312.00752。 arXiv:2312.00752。[5] J. Devlin、M. Chang、K. Lee、K. Toutanova，BERT：用于语言理解的深度双向Transformer预训练，载于：NAACL-HLT（1），2019年，第4171–4186页。

[6] A. D. Bellis、V. W. Anelli、T. D. Noia、E. D. Sciascio，PRONTO：基于提示的掩码语言模型语义包含模式检测，载于：G. Demartini、K. Hose、M. Acosta、M. Palmonari、G. Cheng、H. Skaf-Molli、N. Ferranti、D. Hernández、A. Hogan（编），《语义网——ISWC 2024——第23届国际语义网大会》，美国马里兰州巴尔的摩，2024年11月11日至15日，论文集，第二部分，计算机科学讲义第15232卷，施普林格，2024年，第227–246页。URL：https://doi.org/10.1007/978-3-031-77850-6_13。doi：10.1007/978-3-031-77850-6_13。[7] G. Servedio、A. De Bellis、D. Di Palma、V. W. Anelli、T. Di Noia，隐藏状态是否隐藏着某些信息？检验大语言模型事实性编码能力的极限，arXiv预印本 arXiv:2505.16520（2025年）。[8] D. Di Palma、A. De Bellis、G. Servedio、V. W. Anelli、F. Narducci、T. Di Noia，Llama模型也有感情：通过探测揭示Llama模型中的情感倾向与情绪表征，arXiv预印本 arXiv:2505.16491（2025年）。[9] P. Aghilar、V. W. Anelli、M. Trizio、E. Di Sciascio、T. Di Noia，无需训练且保持身份一致的图像编辑：用于时尚姿势对齐与归一化，《专家系统与应用》293（2025）128579。doi：https://doi.org/10.1016/j.eswa.2025.128579。[10] P. Aghilar、V. W. Anelli、A. Lops、F. Narducci、A. Ragone、S. Roccotelli、M. Trizio，视觉营销中的自适应用户建模：平衡品牌形象与运营效率，载于：《第33届ACM用户建模、自适应与个性化会议论文集》，UMAP 2025，美国纽约市，2025年6月16日至19日，ACM，2025年，第358–360页。URL：https://doi.org/10.1145/3699682.3730976。doi：10.1145/3699682.3730976。[11] Y. Ping、Y. Li、J. Zhu，超越准确率的度量：推荐系统中的多样性、新颖性和惊喜性对用户参与度的影响，《电子商务研究》（2024）1–28。[12] T. Duricic、D. Kowald、E. Lacic、E. Lex，超越准确率：基于图神经网络的推荐系统中多样性、惊喜性与公平性的综述，《Frontiers Big Data》6（2024）。[13] S. Karimi、H. A. Rahmani、M. Naghiaei、L. Safari，推荐系统中的提供者公平性与超越准确性的权衡，CoRR abs/2309.04250（2023）。[14] M. Attimonelli、A. D. Bellis、C. Pomo、D. Jannach、E. D. Sciascio、T. D. Noia，我们真的需要专门化吗？评估用于零样本推荐与搜索的通用型文本嵌入，载于：RecSys，ACM，2025年。URL：https://doi.org/10.1145/3705328.3748040。doi：10.1145/3705328.3748040。[15] D. Di Palma、G. Servedio、V. W. Anelli、G. M. Biancofiore、F. Narducci、L. Carnimeo、T. D. Noia，超越文字：ChatGPT能否支持最先进的推荐系统？，载于：IIR，CEUR研讨会论文集第3802卷，CEUR-WS.org，2024年，第13–22页。[16] M. Valentini，基于大语言模型的协作与竞争多智能体推荐系统，载于：C. Hauff、C. Macdonald、D. Jannach、G. Kazai、F. M. Nardini、F. Pinelli、F. Silvestri、N. Tonellotto（编），《信息检索进展——第47届欧洲信息检索会议》，ECIR 2025，意大利卢卡，2025年4月6日至10日，论文集，第五部分，计算机科学讲义第15576卷，施普林格，2025年，第204–211页。[17] Y. Hou、J. Zhang、Z. Lin、H. Lu、R. Xie、J. J. McAuley、W. X. Zhao，大语言模型是推荐系统的零样本排序器，载于：ECIR（2），计算机科学讲义第14609卷，施普林格，2024年，第364–381页。[18] D. Di Palma，检索增强推荐系统：利用大语言模型增强推荐系统，载于：RecSys，ACM，2023年，第1369–1373页。[19] S. Dai、N. Shao、H. Zhao、W. Yu、Z. Si、C. Xu、Z. Sun、X. Zhang、J. Xu，揭示ChatGPT在推荐系统中的能力，载于：RecSys，ACM，2023年，第1126–1132页。

[20] M. Attimonelli, D. Danese, D. Malitesta, C. Pomo, G. Gassi, T. D. Noia，Ducho 2.0：迈向一个更及时更新的统一框架，用于提取推荐系统中的多模态特征，载于：WWW（伴随卷），ACM，2024年，第1075–1078页。[21] J. Liu, C. Liu, R. Lv, K. Zhou, Y. Zhang，ChatGPT 是一个好的推荐系统吗？一项初步研究，CoRR abs/2304.10149（2023年）。[22] D. Carraro, D. Bridge，利用大语言模型重排序以提升推荐多样性，ACM信息系统汇刊（2024年）。网址：https://doi.org/10.1145/3700604。doi:10.1145/3700604。[23] Y. Tokutake, K. Okamoto，大语言模型能否评估推荐系统中的惊喜性？，J. Adv. Comput. Intell. Intell. Informatics 28（2024年）1263–1272。[24] D. Di Palma, G. M. Biancofiore, V. W. Anelli, F. Narducci, T. D. Noia，基于内容还是协同过滤？从ChatGPT推荐结果的列表间相似性分析中获得的见解，载于：UMAP（附属出版物），ACM，2025年，第28–33页。[25] Y. Deldjoo，理解基于ChatGPT的推荐系统中的偏差：提供者公平性、时间稳定性与时效性，ACM信息系统汇刊（2024年）。网址：https: //doi.org/10.1145/3690655。doi:10.1145/3690655。[26] J. Zhang, K. Bao, Y. Zhang, W. Wang, F. Feng, X. He，ChatGPT 能否公平地进行推荐？评估大语言模型推荐的公平性，载于：RecSys，2023年，第993–999页。[27] A. C. M. Mancino, A. Ferrara, S. Bufi, D. Malitesta, T. D. Noia, E. D. Sciascio，KGToRE：通过知识感知图神经网络模型提供量身定制的推荐，载于：RecSys，2023年，第576–587页。[28] S. Bufi, A. C. M. Mancino, A. Ferrara, D. Malitesta, T. D. Noia, E. D. Sciascio，KGUF：一种简单的知识感知图推荐系统，采用基于用户的语义特征筛选，载于：IRonGraphs，《计算机与信息科学通讯》第2197卷，施普林格，2024年，第41–59页。[29] F. M. Harper, J. A. Konstan, MovieLens数据集：历史与背景，ACM Trans. Interact. Intell. Syst. 5（2016年）19:1–19:19。[30] I. Cantador, P. Brusilovsky, T. Kuflik，信息异质性与推荐系统融合第二届研讨会（HetRec2011），载于：RecSys，ACM，纽约，美国，2011年，第387–388页。[31] G. M. Biancofiore, D. Di Palma, C. Pomo, F. Narducci, T. Di Noia，对话式用户界面与智能体，载于：《以人为中心的人工智能：图解科学探索》，施普林格，2025年，第399–438页。[32] J. L. Herlocker, J. A. Konstan, L. G. Terveen, J. Riedl，评估协同过滤推荐系统，ACM信息系统汇刊 22（2004年）5–53。[33] T. Silveira, M. Zhang, X. Lin, Y. Liu, S. Ma，你的推荐系统表现如何？推荐系统评估综述，Int. J. Mach. Learn. Cybern. 10（2019年）813–831。[34] M. Karimi, D. Jannach, M. Jugovac，新闻推荐系统——综述与未来方向，Inf. Process. Manag. 54（2018年）1203–1227。[35] A. Gunawardana, G. Shani, S. Yogev，评估推荐系统，载于：《推荐系统手册》，施普林格，美国，2022年，第547–601页。[36] M. Kaminskas, D. Bridge，多样性、惊喜性、新颖性与覆盖度：推荐系统中超越准确性的目标综述与实证分析，ACM Trans. Interact. Intell. Syst. 7（2017年）2:1–2:42。[37] P. Cheng, S. Wang, J. Ma, J. Sun, H. Xiong，学习推荐准确且多样的物品，载于：WWW，ACM，2017年，第183–192页。[38] W. Wu, L. Chen, Y. Zhao，基于用户个性实现推荐多样性的个性化，用户建模与用户自适应交互 28（2018年）237–276。[39] M. Nakatsuji, Y. Fujiwara, A. Tanaka, T. Uchiyama, K. Fujimura, T. Ishida，摇滚乐迷的古典音乐

？拓展用户兴趣的新颖推荐，载于：CIKM，ACM，2010年，第949–958页。[40] M. Cai, L. Chen, Y. Wang, H. Bai, P. Sun, L. Wu, M. Zhang, M. Wang，流行度感知的对齐与对比以缓解流行度偏差，载于：KDD，ACM，2024年，第187–198页。[41] V. Paparella, D. Di Palma, V. W. Anelli, T. D. Noia，拓宽视野：评估推荐系统在优先考虑准确率之外的潜力，载于：RecSys，ACM，2023年，第1139–1145页。[42] D. Jannach, L. Lerche, I. Kamehkhosh, M. Jugovac，推荐系统推荐什么：推荐偏差及可能对策分析，用户建模与用户自适应交互 25（2015年）427–491。[43] G. Adomavicius, J. Zhang，数据特征对推荐系统性能的影响，ACM管理信息系统汇刊 3（2012年）3:1–3:17。[44] S. Vargas, P. Castells，推荐系统新颖性与多样性指标中的排名和相关性，载于：RecSys，ACM，2011年，第109–116页。[45] H. Abdollahpouri, R. Burke, B. Mobasher，通过个性化重排序管理推荐系统中的流行度偏差，载于：FLAIRS，AAAI出版社，2019年，第413–418页。[46] Y. Gao, T. Sheng, Y. Xiang, Y. Xiong, H. Wang, J. Zhang，Chat-rec：迈向交互式可解释的大语言模型增强推荐系统，CoRR abs/2303.14524（2023年）。[47] A. Manzoor, S. C. Ziegler, K. M. P. Garcia, D. Jannach，作为对话式推荐系统的ChatGPT：以用户为中心的分析，载于：UMAP，ACM，2024年，第267–272页。[48] S. Sanner, K. Balog, F. Radlinski, B. Wedin, L. Dixon，对于基于语言和项目的偏好，大语言模型是具有竞争力的近冷启动推荐器，载于：RecSys，2023年，第890–896页。[49] Z. Li, Y. Chen, X. Zhang, X. Liang，BookGPT：由大语言模型赋能的图书推荐通用框架，Electronics 12（2023年）4654。[50] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei，语言模型是少样本学习器，载于：NeurIPS，2020年。[51] A. Kong, S. Zhao, H. Chen, Q. Li, Y. Qin, R. Sun, X. Zhou, E. Wang, X. Dong，通过角色扮演提示实现更好的零样本推理，载于：NAACL-HLT，计算语言学协会，2024年，第4099–4113页。[52] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa，大语言模型是零样本推理器，CoRR abs/2205.11916（2022年）。[53] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou，思维链提示可激发大语言模型的推理能力，载于：NeurIPS，2022年。[54] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, K. Narasimhan，思维树：利用大语言模型进行审慎问题求解，载于：NeurIPS，2023年。[55] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, S. Yao，Reflexion：通过语言强化学习的语言智能体，载于：NeurIPS，2023年。[56] Z. Liu, X. Yu, Y. Fang, X. Zhang，GraphPrompt：统一图神经网络的预训练与下游任务，载于：WWW，ACM，2023年，第417–428页。[57] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, A. Chadha，大语言模型提示工程系统综述：技术与应用，CoRR abs/2402.07927（2024年）。[58] L. Xu, J. Zhang, B. Li, J. Wang, M. Cai, W. X. Zhao, J. Wen，面向推荐系统提示大语言模型：综合框架与实证分析，CoRR abs/2401.04997（2024年）。

[59] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever 等，《语言模型是无监督多任务学习者》，OpenAI 博客 1（2019）9。[60] J. Jin, X. Chen, F. Ye, M. Yang, Y. Feng, W. Zhang, Y. Yu, J. Wang，《借助对话代理为推荐系统插上交互的翅膀》，载于：NeurIPS，2023。[61] A. Kong, S. Zhao, H. Chen, Q. Li, Y. Qin, R. Sun, X. Zhou，《通过角色扮演提示实现更好的零样本推理》，CoRR abs/2308.07702（2023）。[62] A. C. M. Mancino, S. Bufi, A. di Fazio, A. Ferrara, D. Malitesta, C. Pomo, T. D. Noia，《Datarec：用于推荐系统标准化且可复现的数据管理的 Python 库》，载于：第48届ACM信息检索研究与发展国际会议论文集，SIGIR 2025，意大利帕多瓦，2025年7月13日至18日，ACM，2025。网址：https://doi.org/10.1145/3726302.3730320。doi: 10.1145/3726302.3730320.[63] V. Paparella, V. W. Anelli, F. M. Nardini, R. Perego, T. D. Noia，《搜索与推荐中的帕累托最优解事后选择》，载于：CIKM，ACM，2023，第2013–2023页。[64] V. W. Anelli, A. Bellogín, A. Ferrara, D. Malitesta, F. A. Merra, C. Pomo, F. M. Donini, T. D. Noia，《Elliot：用于可复现推荐系统评估的全面而严谨的框架》，载于：SIGIR，ACM，美国纽约州纽约市，2021，第2405–2414页。[65] A. Ferrara, V. W. Anelli, A. C. M. Mancino, T. D. Noia, E. D. Sciascio，《Kgflex：利用稀疏特征分解和知识图谱实现高效推荐》，《ACM推荐系统汇刊》（2023）。[66] U. Javed, K. Shaukat, I. A. Hameed, F. Iqbal, T. M. Alam, S. Luo，《基于内容和基于上下文的推荐系统综述》，《国际新兴学习技术杂志》（iJET）16（2021）274–306。[67] B. Paudel, F. Christoffel, C. Newell, A. Bernstein，《面向交互式应用的可更新、准确、多样且可扩展的推荐》，《ACM交互式智能系统汇刊》7（2017）1:1–1:34。[68] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, M. Wang，《LightGCN：简化并增强用于推荐的图卷积网络》，载于：SIGIR，2020，第639–648页。[69] C. Cooper, S. Lee, T. Radzik, Y. Siantos，《推荐系统中的随机游走：精确计算与模拟》，载于：WWW（伴随卷），2014，第811–816页。[70] B. M. Sarwar, G. Karypis, J. A. Konstan, J. Riedl，《电子商务推荐算法分析》，载于：EC，ACM，美国纽约州纽约市，2000，第158–167页。[71] J. S. Breese, D. Heckerman, C. M. Kadie，《协同过滤预测算法的实证分析》，载于：UAI，1998，第43–52页。[72] H. Steck，《适用于稀疏数据的极浅自编码器》，载于：WWW，ACM，美国纽约州纽约市，2019，第3251–3257页。[73] S. Rendle, W. Krichene, L. Zhang, J. R. Anderson，《重新审视神经协同过滤与矩阵分解》，载于：RecSys，2020，第240–248页。[74] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T. Chua，《神经协同过滤》，载于：WWW，2017，第173–182页。[75] G. Salton, A. Wong, C. Yang，《用于自动索引的向量空间模型》，《ACM通讯》18（1975）613–620。[76] Z. Gantner, S. Rendle, C. Freudenthaler, L. Schmidt-Thieme，《MyMediaLite：一个免费的推荐系统库》，载于：RecSys，ACM，美国纽约州纽约市，2011，第305–308页。[77] Y. Chen, Q. Fu, Y. Yuan, Z. Wen, G. Fan, D. Liu, D. Zhang, Z. Li, Y. Xiao，《幻觉检测：稳健辨别大语言模型中的可靠答案》，载于：CIKM，2023，第245–255页。[78] F. Nie, J. Yao, J. Wang, R. Pan, C. Lin，《减少神经表层实现中幻觉的一种简单方法》，载于：ACL（1），2019，第2673–2679页。[79] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, P. Fung，《综述：

自然语言生成中的幻觉，ACM计算调查 55（2023年）248:1–248:38。[80] V. E. Giuliano, P. E. J. Jr., G. E. Kimball, R. F. Meyer, B. A. Stein，通过格式塔方法实现自动模式识别，信息与控制 4（1961）332–345。[81] A. V. Petrov, C. MacDonald，gsasrec：使用负采样训练的序列推荐中的过度自信问题缓解，载于：RecSys，2023年，第116–128页。[82] D. L. Olson, D. Delen，高级数据挖掘技术，施普林格，美国，2008年。[83] K. Järvelin, J. Kekäläinen，信息检索技术的累积增益评估，ACM信息系统汇刊 20（2002）422–446。[84] D. Di Palma, F. A. Merra, M. Sfilio, V. W. Anelli, F. Narducci, T. Di Noia，大语言模型会记忆推荐数据集吗？关于MovieLens-1M的初步研究，载于：第48届ACM信息检索研究与发展国际会议论文集，SIGIR 2025，意大利帕多瓦，2025年7月13日至18日，ACM，2025年。网址：https://doi.org/10.1145/3726302.3730178。doi：10.1145/3726302.3730178。