**Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models**

在大型语言模型时代重新思考对话式推荐系统的评估

**太长不看版：**

本文研究利用ChatGPT进行对话式推荐系统（CRSs）的应用，揭示了现有评估协议的不足之处，作者认为现有评估协议未能重视CRSs的交互性。对此，本文进一步提出一种基于LLMs的交互式评估方法，利用了基于LLMs的用户模拟器，称为iEvaLM。本文通过对两个公共CRS数据集的实验证明了该方法显著改善了性能。同时本文中ChatGPT展示了对其推荐结果的有说服力的解释生成能力，作者认为该研究对深入理解LLMs在CRSs中未开发的潜力以及为基于LLMs的CRSs研究提供更好的评估方法具有重要贡献。

**摘要：**

大型语言模型（LLMs）的最近成功表明其在开发更强大的对话式推荐系统（CRSs）方面具有巨大潜力，该系统依赖自然语言对话以满足用户需求。在本文中，作者对利用ChatGPT进行CRSs的研究进行了调查，揭示了现有评估协议的不足之处。现有协议在强调与人工标注的真实物品的匹配程度时，可能忽视了CRSs的交互性。为了克服这一限制，本文进一步提出一种基于LLMs的交互式评估方法，称为iEvaLM，该方法利用了基于LLMs的用户模拟器。该评估方法可以模拟各种系统-用户交互场景。通过对两个公共CRS数据集的实验证明，与目前的评估协议相比，该方法显著改进了性能。此外，本文强调可解释性的评估，对此，Chat-GPT生成了对其推荐结果的有说服力的解释。代码发布在https://github.com/RUCAIBox/iEvaLM-CRS。

**1 导论：**

CRSs旨在通过跨越多个回合的自然语言对话提供高质量的推荐服务。通常推荐模块根据对话背景中的用户偏好提供推荐，而对话模块则根据对话背景和物品推荐生成响应。近期的研究表明LLMs如ChatGPT有望成为出色的CRS，但缺少对此的全面研究。

本文对此进行了研究，结果令人惊讶：ChatGPT在这个实证评估中表现不佳。作者认为主要原因在于当前的评估协议依赖于手动注释的推荐和对话之间的匹配，可能过分强调基于对话背景的真实物品（ground-truth items）的匹配程度。同时基于固定对话的评估协议没有考虑到会话推荐的互动性质。因此，作者认为传统指标可能无法反映LLM的真实能力。

本文为此进一步提出了一种基于LLMs的交互式评估方法，称为iEvaLM。利用LLMs已表现出来的在角色扮演领域的出色能力，基于LLMs设计了用户模拟器，可以灵活适应不同的CRS，而无需进一步调整，使CRS能以更自然的方式与用户进行交互。为了进行全面的评估，本文还考虑了基于属性的问答和自由形式的闲聊这两种互动方式。

使用该评估标准，ChatGPT在REDIAL数据集上，通过五轮互动，Recall@10指标从0.174增加到0.570，甚至超过了当前领先的CRS baseline模型的Recall@50结果。在可解释性这一关键方面，ChatGPT展示出了为其推荐提供有说服力解释的能力。此外，现有的CRSs可以通过互动提升推荐性能，这往往被传统评估所忽视。在OPENDIALKG数据集上，现有的CRSs在基于属性的问答设置下表现得更差，而ChatGPT在两种设置下表现都更好。作者总结其主要贡献如下：

(1)第一次系统地对ChatGPT在大规模数据集上进行了有关对话推荐的能力的考察。

(2)对传统的评估协议下ChatGPT的局限性进行了详细分析，讨论了其在现有基准测试上失败的根本原因。

(3)提出了一种新的交互式方法，利用基于LLM的用户模拟器来评估CRS,并证明其有效性。

**2 背景和实验设置**

**2.1 任务描述**

CRSs旨在通过多轮交互，即基于模板的问答或基于自然语言的闲聊，来提供物品推荐。文中考虑基于自然语言的闲聊。CRSs由两个主要子任务组成：推荐和对话。在每一轮中，系统要么呈现一项推荐，要么启动一个新的对话，直到用户接受了推荐的项目或终止了对话。鉴于ChatGPT在对话方面的表现出色，将重点评估其在推荐子任务中的性能。

**2.2 实验设置**

**数据集：**在 RE-DIAL（Li et al., 2018）和 OPENDIALKG（Moon et al., 2019）数据集上进行实验。它们的统计数据如下：

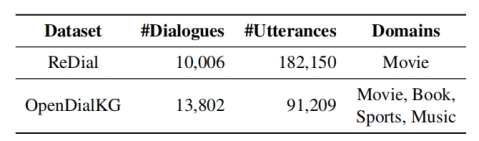


表1：数据集统计

**Baselines：**本文对ChatGPT与一些代表性的监督和无监督方法进行比较分析，包括KBRD（Chen et al., 2019）KGSF（Zhou et al., 2020）CRFR（Zhou et al., 2021a）BARCOR（Wang et al., 2022b）MESE（Yang et al., 2022）UniCRS（Wang et al., 2022c）text-embedding-ada-002（Neelakantan et al., 2022）。除text-embedding-ada-002是一种无监督方法外，其他方法是在CRS数据集上进行监督训练的。

**评估标准：**采用Recall@k评估推荐子任务，在REDIAL数据集上设置k = 1、10、50，在OPENDIALKG数据集上设置k = 1、10、25（由于ChatGPT有时可能会拒绝太多的项目，只评估Recall@1和Recall@10）。

**模型细节：**使用OpenAI API提供的公开可用的模型gpt-3.5-turbo，并在调用API时设置温度=0，使输出更具确定性。

**3 ChatGPT的对话式推荐**

作者在本章中首先讨论了如何将ChatGPT应用于CRSs，随后分析了其性能。

**3.1 方法**

本文提出了以下两种方法：

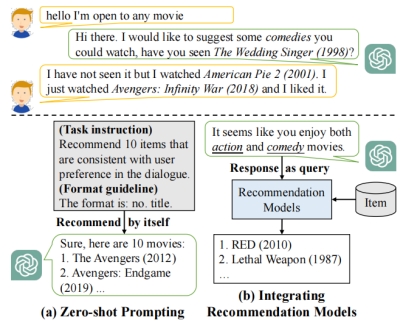


图 1: 将ChatGPT适应于CRS的方法。

**零样本设置：**本文首先在零样本提示设置下测试ChatGPT，提示包含任务说明（描述任务）和格式指导（规定输出格式）。

**集成推荐模型：**ChatGPT并非专门针对推荐进行优化，且它往往生成不在评估数据集中的物品，使直接评估预测结果变得困难。因此，本文纳入外部推荐模型以限制输出空间，通过将对话历史和生成的回复连接起来，作为这些模型的输入，直接预测目标物品或计算与物品候选项的相似度进行匹配。本文选择CRS模型MESE作为无监督方法。

**3.2 评估结果**

首先将ChatGPT的正确率与CRS baseline相比较，然后展示了由其所产生的解释，以评估其可解释性。

**3.2.1 正确率**

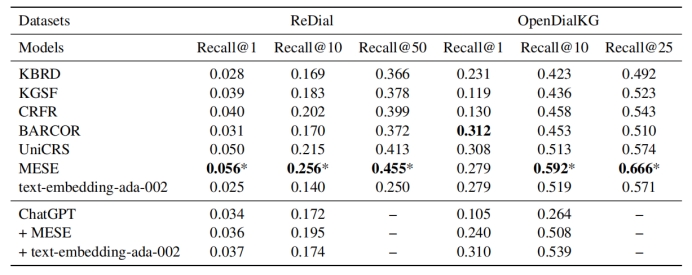


表 2: 现有CRS和ChatGPT的整体性能。带有\*标记的数字表示改进与最佳基线相比具有统计学显著性（t检验的p值<0.05）。

当使用零样本提示时，ChatGPT只在这些基线方法中取得了平均性能，并且远远落后于性能最佳的方法。当结合外部推荐模型时，它的性能可以有效提高，特别是在OPENDIALKG数据集上。作者认为这表明ChatGPT生成的回复可以帮助外部模型理解用户的偏好，但在REDIAL数据集上性能远远不足。

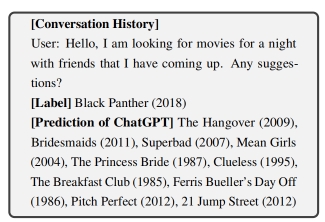
**3.2.2 可解释性**

为更好地理解ChatGPT进行推荐的方式，作者要求其生成解释来检查其内部工作原理。然后，作者请两位评注员来判断解释与对话内容的相关程度（不相关，部分相关或高度相关）。ChatGPT能够理解用户偏好并给出合理的解释，却与其在准确性方面的表现差相矛盾，导致作者研究失败。

**3.3 ChatGPT为何失败**

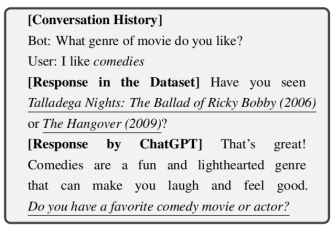
作者认为有以下两个原因：

1. **缺乏明确的用户偏好。**因示例通常轮次较少，CRS可能无法收集足够的证据来准确推理用户的意图。并且这些对话主要以闲聊形式收集，很难反映真实的用户偏好。



作者随机抽取了100个具有少于三个轮次的失败示例，并邀请两名注释员确定用户偏好是否模糊不清。在Cohen's Kappa系数为0.75的前提下51%的示例被标注为缺乏明确的用户偏好。作者进一步提出因为ChatGPT没有进行微调，只基于对话上下文进行预测，这一问题对其影响更大。

1. **缺乏积极的澄清。**评估的一个主要限制是必须严格遵循现有的对话流程。在真实场景中，CRS会在需要时提出积极的澄清，而现有的评估协议并不支持这一点。但由于有这么多item符合当前的要求，在进行推荐之前询问澄清是合理的。



两个注释员对ChatGPT生成的响应进行分类（澄清、推荐或闲聊）。我们发现其中36%是澄清，表明需要考虑澄清在评估中的重要性。（Cohen's Kappa系数为0.81）

总之，作者认为现有的评估协议虽然基于对话级的评估，允许系统与用户互动，但它仅限于预定义的对话流程或基于模板的话语无法捕捉现实世界对话的复杂性和细微差别。

**4 CRSs的一种全新评价方法**

作者提出了一种替代性的评估方法iEvaLM，采用基于LLM的用户模拟进行交互式评估。并通过实验展示了它的有效性和可靠性。

**4.1 概览**

iEvaLM与现有的CRS数据集无缝集成。关键思想是基于LLM的出色角色扮演能力进行接近真实的用户模拟。将真实项目视为用户偏好，并通过指示来构建基于LLM的模拟用户的角色。在交互之后，本文不仅通过将预测结果与真实项目进行比较来评估准确性，还通过向基于LLM的评分器查询生成的解释来评估可解释性。

**4.2 交互形式**

1.基于属性的问答：系统的行为被限制为从k个预定义的属性中选择一个来询问用户或进行推荐。在每一轮中，首先让系统确定这k+1个选项，然后用户以基于模板的回答回应。

2.自由形式的闲聊：没有任何限制，系统和用户都可以自由主动发起。

**4.3 用户模拟**

本文采用LLM进行用户模拟。模拟用户可以采用以下三种行为之一：

• 谈论偏好。当系统对用户的偏好进行澄清或征询时，模拟用户会回复有关目标物品的信息。

• 提供反馈。当系统推荐一组物品时，模拟用户会检查每个物品，并在找到目标物品时提供积极的反馈，否则提供消极的反馈。

• 结束对话。如果系统推荐了目标物品之一，或者交互达到一定的轮数，模拟用户会结束对话。

作者使用现有数据集中的真实物品来构建模拟用户的逼真角色，利用OpenAI API中提供的text-davinci-003模型构建评估器，通过手动指令设置其行为：首先将真实物品填充到角色模板中，然后使用一组手动制作的规则来定义它们的行为。在每个回合中将对话附加到指令中作为输入，将max\_tokens设置为128，温度设置为0，交互轮数的最大值被设置为5。

**4.4 表现测定**

本文采用了主观和客观指标来衡量推荐的性能以及用户体验。对于客观指标，本文采用召回率来评估交互过程中的每个推荐动作；对于主观指标，使用说服力来评估交互过程中最后一个推荐动作的解释质量，旨在评估用户是否可以被说服接受推荐（取值范围为{0, 1, 2}）。使用OpenAI API提供的text-davinci-003作为评分器，将对话、解释和评分规则连接为提示，以减少人力需求。

**5 评估结果**

**5.1 用户模拟器质量**

作者利用现有数据集中的带注释对话来评估。邀请五名标注者扮演系统的角色，根据随机抽样的100个示例的第一轮带注释对话，与每个用户模拟器进行互动，完成其余对话。然后再请另外两名标注者进行成对评估，一个来自模拟器生成的话语，另一个来自DialoGPT或数据集。当两名标注者都认为某个方法的话语更好时，计为该方法获胜；否则，计为平局。标注者之间的Cohen's Kappa为0.73。评估指标为评估单轮和多轮自由形式闲聊的设定中生成话语的自然性和有用性。自然性意味着话语流畅且可能是由人类生成的，而有用性则意味着话语与用户的偏好一致。

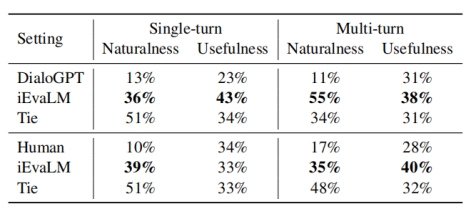


表 4: 在单轮和多轮设置中，以自然性和实用性为指标进行性能比较。每个数值表示特定模型在两两比较中获胜或并列的百分比。

结果显示，模拟器在自然性方面显著优于DialoGPT，模拟器的有用性也优于其他方法。

**5.2 CRS的表现**

本文使用不同的评估方法来比较现有CRS和ChatGPT的性能。对于ChatGPT使用ChatGPT + text-embedding-ada-002，因为它在传统评估中表现出色。

**5.2.1 主要结果**

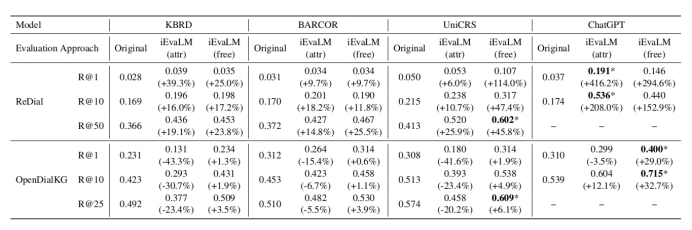


表 5: 不同评估方法下CRS和ChatGPT的性能，其中“attr”表示基于属性的问答，“free”表示自由形式的闲聊。“R@k”表示Recall@k。

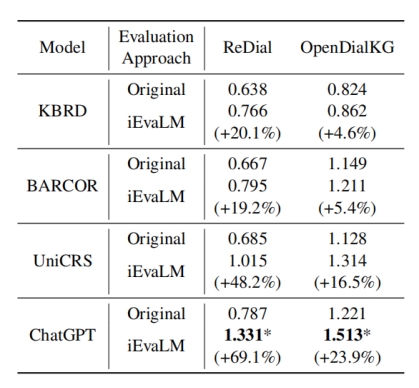


表 6: 解释的说服力。只考虑iEvaLM中的自由形式聊天设置。

评估结果见表格 5和表格 6。大多数模型在准确性和可解释性方面相较于传统方法有所提升。在现有的CRS中，性能排名为UniCRS > BARCOR > KBRD（UniCRS和BARCOR都利用预训练模型以增强对话能力，UniCRS还将知识图谱（KGs）融入提示中）。作者认为这表明现有的CRS具有与用户交互来提供更好推荐和用户体验的能力，往往被传统评估所忽视。

对于ChatGPT，在召回率和说服力方面均有显著的性能提升，甚至超过了两个数据集上大多数CRS的Recall@25或Recall@50。这表明ChatGPT具有更强的交互能力，并且可以根据用户偏好提供高质量和有说服力的推荐，也证明了iEvaLM在评估CRS准确性和可解释性方面的有效性。

作为通用CRS，ChatGPT展现出更大潜力。现有的CRS在OPENDIALKG数据集的基于属性的问题回答设置中表现较差，远远不如传统设置中的表现，作者认为可能是情景设置不一致导致的。这些结果表明了传统评估的局限性，它仅关注单一对话情境，而本文的评估方法则允许多种情境的评估。

**5.2.2 评价的可靠性**

作者随机选择100个示例，其中包含由ChatGPT生成的解释，并要求评分器和两个标注者根据相同的指示分别对它们进行评分。标注者之间的Cohen's Kappa为0.83。下图显示了两个得分分布的相似性，表明本文基于LLM的评分器作为人类评估者的替代物是可靠的。

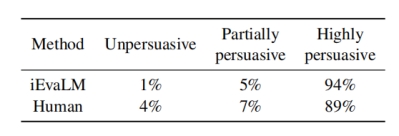


表 7: 使用基于LLM的评分器和人工评分，对REDIAL数据集中100个随机例子的说服力进行评估分配得分。

作者同时随机选择了100个实例，并请五个标注者和用户模拟器与不同的CRS进行自由形式的闲聊。结果如表8所示。

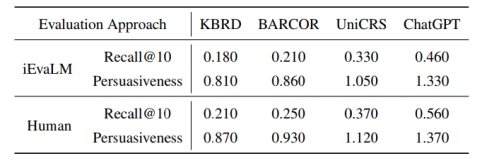


表 8: 对REDIAL数据集中的100个随机示例应用模拟用户和真实用户进行评估

可以看到，我们的用户模拟器的排名与真实用户的一致，并且绝对得分也是可比较的。这表明我们基于LLM的用户模拟器能够提供令人信服的评估结果，并作为可靠的人类评估者的替代方法。

**6 总结**

本文系统地考察了ChatGPT在现有基准数据集上进行对话推荐的能力，并提出了一种替代的评估方法iEvaLM。首先证明了ChatGPT的性能不令人满意。通过对失败案例的分析，发现问题的根源是现有的评估协议过分强调基于对话上下文的地面真实项目的匹配。为了解决这个问题，提出了一种使用基于LLM的用户模拟器的交互式评估方法。

通过使用这种新方法进行实验，得出以下结论：（1）ChatGPT非常强大，在评估中比目前领先的CRS在准确性和可解释性方面表现更佳；（2）现有的CRS也在交互中得到了改进，这是传统评估忽视的一个重要方面；（3）ChatGPT在不同的设置和数据集下展示出了成为通用型CRS的巨大潜力。同时，作者还证实了该评估方法的有效性。

总之，作者的工作对于理解和评估诸如ChatGPT这样的LLM在对话式推荐中的作用做出了贡献，为在LLM时代进一步研究这一领域铺平了道路。