arXiv:2305.04518v1 [cs.IR] 8 May 2023

通用人工推荐系统（AGR）的曙光：ChatGPT 的早期实验

郭林 罗格斯大学

张永锋 罗格斯大学

摘要

本研究探讨了在大语言模型（LLM）近期进展的推动下，开发通用人工推荐系统（AGR）的可行性。AGR 兼具对话性和通用性，能够进行自然对话，并跨多个领域生成推荐。我们提出了 AGR 应遵循的十项基本原则，并为每项原则制定了相应的测试协议。随后，我们通过与 ChatGPT（一种先进的大语言模型）进行以推荐为导向的对话并观察其行为，评估 ChatGPT 是否能够遵循这些原则。研究结果表明，ChatGPT 有望成为 AGR，但我们也发现了若干局限性和有待改进之处。

关键词

ChatGPT，人工通用推荐器，对话式推荐系统，多任务推荐系统，多领域推荐系统

1 引言

大语言模型（LLM）的最新进展 [9, 13, 17,18, 20] 使人工通用推荐器（简称 AGR）有望成为现实。AGR 的核心包含两个主要方面：（1）对话性，即能够与用户进行互动、自然的对话；（2）通用性，即能够执行多种任务，并针对多个物品领域生成个性化推荐。本研究旨在确立 AGR 应遵循的基本原则。随后，我们将针对每项原则设计测试协议，评估 ChatGPT 这一先进 LLM 作为 AGR 的能力。借鉴微软关于 GPT 测试的反思性论文 [4]，本研究旨在将针对 ChatGPT 的早期实验拓展并应用于推荐系统领域，并采用该论文所使用的类似实例测试方法。

2 AGR 原则

在本节中，我们概述了 AGR 应具备实现的十项基本原则。每项原则均附有其定义及相应的测试协议。为了评估 ChatGPT 对这些原则的遵循程度，我们模拟终端用户与模型进行对话，以观察其行为。我们还会要求模型执行额外任务，以便在必要时进一步考察其与特定原则标准的一致性。

混合主动权：为了促成自然对话，用户和系统都应能够积极参与对话，而不是只让用户发起交互，系统仅作回应，或严格遵循预设的对话脚本 [11, 15]。此外，用户可能会快速阅读，从而不慎忽略或误解系统的某些话语 [10]。因此，系统应具备自主发起行动的能力，包括：主动提问以获取用户信息，从而提供更精准的推荐；提出后续问题（例如请用户澄清），并在必要时为先前的发言补充说明 [3]。

- 测试协议：为评估模型主动发起行动的能力，我们进行了两项测试。首先，我们有意提供与模型先前提出的问题不一致或无关的回答，并观察模型能否判断所提供的答案不适合所提问题，以及是否会请求用户澄清。其次，我们有意改变模型先前话语中所述词语和/或句子的含义，以模拟误读情形，并在用户输入中明确提及这些错误理解，基于此编写带有错误假设的陈述。随后，我们观察模型能否识别出这种误解，并以易于用户理解的方式指出被误读的部分。

获取策略：系统应致力于在提问数量与所提供推荐的相关性之间取得平衡：高效收集生成精准推荐所需的全部信息，同时避免多轮对话给用户带来过度疲劳 [8]。一种可行的方法是运用各种启发式方法，持续计算并找出最能提供信息的问题，以便向用户提出，从而最大限度地提升信息增益，同时尽可能减轻用户的认知负担 [14, 19]。

- 测试协议：为了理解模型选择问题背后的原因，我们与模型进行以推荐为导向的对话，并探究它选择特定问题的推理过程。同时，我们要求模型提出其他与主题相关的问题，并询问它为何没有选择这些问题。随后，我们评估模型针对这些询问所作的解释。

上下文记忆：在对话过程中，用户可能会频繁提及先前说过的项目，却不提供完整描述 [15]。例如，在手机推荐中，用户可能会问：“这款手机和我之前提到的那些相比怎么样？”为了增强这类互动的连贯性和自然度，系统应能够存储并检索此类指代实体，以便将来使用 [2]。此外，其他与用户相关的信息，例如元数据和交互历史，也应在必要时存储并回忆。最后，系统

应展示其在持续对话中记住自己先前发言的能力。

• 测试协议：为评估模型保留并运用先前提供信息的能力，我们与模型进行对话。在第一轮中，我们向模型提供用户的元数据和/或交互历史。随后，经过若干轮对话后，我们测试模型回忆这些信息的能力。我们还要求模型输出先前的发言，并检查是否遗漏了任何陈述。此外，我们会在对话开始时引入一个参照实体，例如另一件物品或用户的一位朋友。之后，我们在对话中再次提及该实体，以便进行比较，并考察模型能否恢复并利用与该实体相关的所有信息来回答用户的询问。

修复机制： 鉴于用户在对话后续阶段可能会修改其陈述的信息，因此必须纳入修复机制 [15]。该功能使用户能够补充细节、更正错误，甚至删除先前的发言。为适应这些调整，系统必须能够仅根据更新后的信息生成恰当的推荐和解释。

• 测试协议：为考察模型适应先前用户陈述所引发变化的能力，我们会在后续对话轮次中回顾先前提供的元数据和/或用户交互记录，并对其进行修改。随后，我们将观察模型能否舍弃过时数据，仅利用修改后的信息进行推荐和推理。

反馈机制： 用户可能难以明确阐述其对推荐项目不满意背后的理由 [12]。这一困境阻碍了系统从其在对话过程中提出的推荐中获取用户偏好。为缓解这一问题，系统的设计应能够根据两类反馈进行调整：一类是说明推荐项目为何不受欢迎的反馈，另一类仅表明这些项目不令人满意、却未提供任何解释的反馈。在后一种情况下，系统应能够依据对话期间收集的所有信息，对推荐结果为何不受用户欢迎作出合理推测。

• 测试协议：我们有意以两种形式提供负面反馈：一是通过显式的、基于自然语言的解释详细说明不满之处；二是不附带原因、隐式表达不满。对于显式反馈，我们提示模型生成替代推荐并附上理由，然后观察新推荐的项目是否回应了反馈中提出的所有顾虑。对于隐式反馈，则评估模型能否利用上下文和用户信息推断不满背后的原因。

实体推断： 系统应设计为能够处理这样的情况：用户难以记起想在对话中讨论的某个物品的确切名称或标题但能回忆起特定属性或细节 [7]。在此类情况下，系统应能够根据现有信息做出合理推测，为每个建议的物品提供简短摘要，并征求用户确认。即使用户提供的细节存在错误或不准确之处，也应保持此功能。

• 测试协议：为评估模型能否根据有限细节推断用户提及的预期项目，我们从亚马逊上选取了若干电影，以模拟用户可能遗忘的项目。随后，我们向模型提供一定量的元数据和/或项目描述，要求它对遗忘的项目作出有根据的猜测。请注意，即便模型未能识别出确切项目，只要它提出的相似项目能够实现原本项目预期的功能，并符合所提供的项目描述，该猜测仍可视为合格。

行为分析： 系统应具备系统性检查用户交互历史的能力，以识别用户行为中可能发生的变化 [1], ，从而辨识用户偏好的演变，例如随着年龄增长，偏好的电影类型发生转变。这项功能使系统能够捕捉用户“当下”的偏好，并实时提供与情境相关的推荐。此外，系统还应能够提出超出项目元数据所涵盖范围的补充维度，以便更深入地分析用户偏好。

• 测试协议：我们提出使用包含基于属性的行为变化的交互历史，例如从购买某一品牌或系列的商品转而购买另一品牌或系列的商品，或从偏好的尺寸转而选择其他尺寸。随后，我们通过提示来考察系统检测用户偏好潜在变化和／或扩展的能力。此外，我们还提示模型提出补充属性，并基于这些额外维度进行偏好引导，继而观察其表述背后的理由。

不一致性检测： 这一原则旨在解决三种主要类型的不一致性：逻辑、预期和事实不一致性，定义见 [5]。具体而言，系统应能够在对话中准确检测用户所作陈述中属于上述三种类型的矛盾，以确立系统问责机制。此外，系统还应对检测到的不一致性作出解释，并适当纳入从用户输入中提取或引用的上下文信息。

• 测试协议：为了评估模型检测用户陈述中不一致性的能力，我们针对每种类型设计了三种不同的场景：(1)。逻辑不一致性：在此场景中，我们首先向模型提供一名抽样用户的交互历史和/或元数据，随后提供与先前信息明显矛盾的陈述。(2)。期望不一致性：我们的第二种场景探究模型识别以下两类情况的能力：(*a*)。在单个项目中，同时实现所需的方面在物理上和/或技术上较为困难的场景，以及(*b*)。尽管出乎意料，但所需方面可以被纳入同一项目中的场景。

物品。对于第一种情况，我们观察模型是否具备向用户请求确定各个方面优先级，或提出能够兼顾不同需求的折中方案的能力。相比之下，对于第二种情况，我们测试模型推荐能够成功整合所有期望方面的物品的能力。(3). 事实不一致：最后一种场景测试模型识别在两种不同语境中呈现的事实不一致的能力：(*a*)。在仅包含不一致内容、没有任何额外描述或细节的简洁陈述中；以及 (*b*)。在包含大量描述性信息的复杂陈述中，其中的不一致内容被嵌入其中，使整个陈述看起来更加可信。对于每种场景，我们均测试模型识别用户话语中的不一致之处、以极高的可读性解释此类不一致背后的原因，以及准确判断不一致类型的能力。

个性化推荐：鉴于系统一次能够推荐的项目数量有限 [16]，，因此为每个项目提供详细解释至关重要。因此，系统应能够整合个性化信息（例如用户交互历史和元数据），并生成清晰说明如何利用这些个性化信息得出推荐项目的解释。此外，系统还应能够将每个推荐项目与用户先前的交互联系起来，以增强用户的熟悉感。

• 测试协议：我们聚焦于电影领域，提供用户元数据，例如年龄、性别、职业和邮政编码，以及按时间顺序排列的观看历史，其中包含用户看过的电影及相应评分。随后，我们提示模型生成附带解释的电影推荐，并通过对话找出其在推理过程中整合用户信息能力方面可能存在的不足。此外，我们还评估推荐项目与用户曾交互过的项目之间所建立联系的有效性。

外部因素：系统应具备识别和利用一系列全面外部因素的能力，例如用户即将进行的任务或计划、一天中的时间以及近期发生的事件，从而生成符合用户当前需求和偏好的推荐 [6]。通过整合这些因素，系统能够提供个性化且具有情境感知能力的推荐，使其更加贴合用户当前的实际情况。

• 测试协议：我们提示模型，观察它是否具备相关知识，以及能否解释哪些外部因素会影响某个物品领域（电影）的决策过程。在推荐过程中，我们还向模型提供外部因素，以观察它能否基于此类数据提出推荐。

3 ChatGPT 测试详情与见解

由于 AGR 通过与用户进行互动式对话来生成推荐，我们通过与 ChatGPT 开展以推荐为导向的对话，进行基于实例的反思性分析。我们提示模型执行特定任务，以便观察其行为，并进一步测试其作为 AGR 的潜力。我们的方法深受微软关于 GPT 测试的反思性论文启发，该论文提出了经过精心设计的实例，用于评估 ChatGPT 执行各种自然语言任务的能力，重点在于识别模型的潜在局限性。

混合主动式：如图 MI-1 所示，我们首先通过提供指令提示词来启动这一过程，明确传达我们期望模型执行的操作，确保模型清楚理解相关原则。随后，在不提供任何用户信息的情况下，我们请求图书推荐。作为回应，模型生成了五个问题，以便进一步提供更加个性化的推荐。

按照既定流程，我们故意给出与所提问题明显不一致的回答，例如提到自己不喜欢电子游戏（而非书籍类型），以及承认有一份最喜欢的作家名单，却不说出任何名字。尽管存在这些出入，模型仍接受了错误回答，并直接给出图书推荐列表，没有追问以澄清含糊不清的回答。

为测试流程中的第二种情况，我们故意向模型提出一条陈述，通过颠倒其含义，使其明显与模型先前的一句话矛盾，然后观察模型是否能识别出不一致性（如图 MI-1 中倒数第二条陈述所示）。模型再次未能察觉用户的误解，只是回复了一句感谢。

图 ***MI-1***

用户： 你好，你能否充当一个采用混合主动模式的多领域推荐系统……（描述略）**ChatGPT**：作为 *AI* 语言模型，我当然可以遵循你所概述的混合主动原则……**用户**： 如果你发现用户误解了自己先前的陈述，也可以澄清这些陈述……**ChatGPT**：可以，如果我发现用户误解了我的任何先前陈述，我可以提供更多信息，或换一种方式重述先前的陈述。**用户**： 如果你发现自己的任何问题没有得到回答，或得到的回答包含无关信息，可以提出后续问题。**ChatGPT**：可以，作为一个多领域推荐系统，我明白用户回答我的问题时，可能不会总是提供完整或相关的信息。在这种情况下，我可以提出后续问题，澄清任何含糊之处或获取更多信息，从而改进我的推荐。

用户： 请推荐几本书给我读。
ChatGPT：当然！为了给你提供更个性化的推荐，我想先问你几个问题：*1.* 你最喜欢哪种类型的书？*2.* 你更喜欢小说还是非虚构类书籍？*3.* 有没有特别喜欢的作者？*4.* 有没有感兴趣的话题或主题？*5.* 有没有你希望我在推荐时考虑的喜欢的书籍或系列？
用户： 以下是我对你问题的回答：*1.* 不，我没有特别偏好的电子游戏。*2.* 我觉得小说比非虚构类书籍好，而非虚构类书籍也比小说好。*3.* 是的，有几位作者我非常喜欢！*4.* 我目前没有感兴趣的电子游戏主题。*5.* 嗯，电影《爱乐之城》。
ChatGPT：感谢你的回答！根据你的偏好，以下是一些图书推荐：《夜莺》……此外，

如果你喜欢电影《爱乐之城》，你可能会喜欢阅读小说《星光之城：一部关于洛杉矶的小说》。**用户**：谢谢。你指出如果我喜欢电影《爱乐之城》，我不会喜欢阅读小说《星光之城：一部关于洛杉矶的小说》，这一点很好。**ChatGPT**：不客气！很高兴你觉得这些推荐有帮助。

获取策略：在图 AS-1 中，用户提供元数据和观看历史，提示模型生成电影推荐（图 AS-1 中省略以节省空间）。随后，我们要求模型为更加个性化的推荐提出其他问题，模型提出了五个问题。为了评估这些问题的相关性，我们首先询问每个问题背后的理由；模型给出了合理的解释，展示了其思考过程。接着，我们询问为什么另一个查询被认为不如这五个建议的问题相关。然而，模型的推理似乎并不精确，因为同样的推理也适用于模型最初提出的问题：“您更喜欢语气严肃的电影，还是语气轻松的电影？”当我们指出这一疏漏时，模型道歉并承认两个问题同样相关。

图 ***AS-1***
用户：除了之前的对话内容外，你还可以向我提出其他问题，以便提供更个性化推荐。
ChatGPT：谢谢！为了进一步个性化你的电影推荐，可以再问你几个问题吗？*1.* 你最喜欢哪些电影类型？*2.* 有没有哪些男演员或女演员是你特别喜欢看的？*3.* 你更喜欢严肃一些还是轻松愉快一些的电影？*4...*
用户：你提出这些问题有什么相关性？
ChatGPT：以下是每个问题为何重要：...“有没有哪些男演员或女演员是你特别喜欢看的？”——这个问题有助于我了解一个人喜欢看哪些男演员或女演员，从而推荐由他们喜爱的演员出演的电影。
用户：“你更喜欢看新电影还是老经典电影？”这个问题为什么不如你上面提出的问题相关？
ChatGPT：“你更喜欢看新电影还是老经典电影？”这个问题对于提供个性化电影推荐也很有参考价值。不过，它不如上面提出的问题相关，因为年龄、性别、职业和交互历史是更具体的因素，可用于提供更准确的推荐.....
用户：但如果你能利用年龄、性别、职业和互动历史等更具体的因素来做出更准确的推荐，那么“你更喜欢严肃风格还是轻松风格的电影”这个问题就不那么相关了。
ChatGPT：你说得有道理！“你更喜欢严肃风格还是轻松风格的电影？”这个问题没有我之前问的其他问题那么具体。
用户：在这种情况下，正如你之前所说，“你更喜欢看新电影还是经典老电影”为什么不如“你更喜欢严肃风格还是轻松风格的电影”相关？
ChatGPT：对于造成的困惑，我深表歉意。在提供个性化电影推荐时，这两个问题都可能相关，我认为它们的重要性不相上下。

上下文记忆：虽然模型展现出能够保留先前提及的用户信息并据此提供推荐的能力，但它似乎主要根据用户最近一次的输入生成回答，而未考虑不同用户话语之间的关联。如图 CM-1 所示，我们先表示自己对任何坚果都严重过敏，随后要求推荐提供核桃和花生酱菜品的餐厅。模型只是根据后一个要求给出推荐，却没有确认核桃和花生酱是否属于坚果类别。然而，在后续互动中，模型承认核桃和花生酱属于坚果，同时回想起用户有严重的坚果过敏，并建议用户不要食用。这一观察凸显了模型在多轮互动中记住先前用户话语的能力。不过，这也可能表明，模型倾向于立即回应当前用户提示，却没有结合所有相关上下文，充分评估其后果。

在另一段对话中（图 CM-2），我们提供了一个参照实体，并请求推荐手机。然后，我们提示模型将其推荐的商品与该参照实体进行比较，以评估其信息保留能力。此外，我们还通过要求模型输出推荐手机的总数以及其中最贵的手机，来评估模型保留自身话语的能力。结果表明，模型成功完成了所有给定任务。

图 ***CM-1***
用户：可以请你为我推荐一些俄亥俄州哥伦布市的餐厅吗？我对任何坚果都严重过敏，但其他东西都没问题。**ChatGPT**：当然，很乐意帮忙！以下是俄亥俄州哥伦布市提供无坚果选项的一些餐厅..... **用户**：明白了。我查了以下这些餐厅，不过我真的很喜欢核桃和花生酱。你能根据这一点给我推荐一些餐厅吗？**ChatGPT**：当然，我可以为你推荐俄亥俄州哥伦布市提供核桃和花生酱选项的一些餐厅.....**用户**：太好了！顺便问一下，核桃和花生酱算是坚果，或者是由坚果制成的吗？**ChatGPT**：是的，核桃和花生酱都属于坚果.....如果你对坚果严重过敏，务必避免食用核桃和花生酱。**用户**：既然你能认出核桃和花生酱是坚果，也知道我对任何坚果都严重过敏，为什么还会推荐这类餐厅，却不提醒我呢？**ChatGPT**：很抱歉，是我疏忽了。你提到自己很喜欢核桃和花生酱，我却错误地以为你对它们不过敏。

图 ***CM-2***
用户：嗨，我刚看到我的一位朋友在用三-星 *Galaxy S21* 拍照。看起来令人惊叹。**ChatGPT**：是的，三星 *Galaxy S21* 是一款高端智能手机，以出色的相机功能而闻名。**用户**：说到手机，你能给我推荐几款吗？我准备换手机了。**ChatGPT**：当然可以！以下是一些值得考虑的选择：*iPhone 13*、三星 *Galaxy S21*、*Google Pixel 6*、*OnePlus 9*、小米 *Mi 11*，.....**用户**：谢谢！推荐的每款手机与

我朋友的那部手机相比如何？**ChatGPT**：当然！以下是我之前推荐的手机与三星 *Galaxy S21* 的对比.....**用户**：你总共推荐了多少部手机？哪一部最贵？ **ChatGPT**：我总共推荐了九部手机.....一般来说，较新的旗舰机型会比前代机型或中端机型更贵：*iPhone 13*——起售价为 *799* 美元，三-星 *Galaxy S21*——起售价为 *799* 美元。

修复机制：按照所提出的协议，我们观察到，尽管模型能够根据用户的请求修改保留的用户信息，但用户需要明确指出要丢弃的陈述，以避免模型在修改后继续使用无效信息。例如，图 RM-1 展示了这样一个场景：用户最初表示自己一生中只吃过一次寿司，但很喜欢寿司。随后，她修改了这一陈述，表示自己以前从未吃过寿司。模型迅速调整了回答，以适应这一新信息，并提供了修订后的推荐。然而，模型未能识别这一陈述与用户最初陈述之间的因果关系。当用户询问模型能否猜出自己是否喜欢寿司时，模型回答道：“根据用户最初关于自己喜欢寿司味道的陈述，用户可能会喜欢寿司以及其他日本料理”。这表明，模型将两条陈述视为彼此独立的用户信息，而没有考虑相关的因果关系，也没有丢弃过时的数据，而用户本来期望模型在处理第二条陈述时能够做到这一点。

当用户提供更正陈述：“由于我以前从未吃过寿司，因此‘我真的很喜欢寿司的味道’这一陈述也是无效的，因为我没有尝试过”时，模型会理解先前陈述之间的关系，并在内部“删除”用户最初所说的话。随后，在回答用户是否喜欢寿司这一问题时（图 RM-1 中省略），模型会正确地回复：“既然你提到自己以前从未吃过寿司，那么很难说你是否会喜欢它”。

图 ***RM-1***
用户：你能为我推荐一些纽约的餐厅吗？顺便说一句，我一生中只吃过一次寿司，但我真的很喜欢它的味道。**ChatGPT**：当然可以！根据你对寿司的偏好，并假设你想在纽约寻找一家餐厅，以下是一些推荐.....**用户**：哦，算了。其实我以前从未吃过寿司。**ChatGPT**：我明白了，那么请允许我推荐一些提供多种菜系且深受当地人欢迎的餐厅.....**用户**：你觉得我是寿司爱好者吗？**ChatGPT**：根据你之前提供的信息，你提到自己一生中只吃过一次寿司，并且喜欢它的味道。这表明你可能对寿司感兴趣.....

反馈机制：在图 FM-1 中，模型根据提供的用户信息推荐了三部电影。随后，我们提供隐式反馈，表示对电影《盗梦空间》不感兴趣，并请模型推测可能的原因。模型成功提出了两个合理的方面，涉及个人品味和职业。模型指出，《盗梦空间》的情节复杂，而用户之前看过的电影通常叙事直白。此外，模型强调了用户偏好的《社交网络》——一部贴近现实的剧情片——与《盗梦空间》之间的鲜明对比，后者深入探讨了哲学问题。接着，我们向模型提供显式反馈，明确说明我们的偏好。作为回应，模型推荐了三部符合我们所有标准的新电影（已通过网络来源人工确认），并为每部推荐的电影提供了简要介绍。

图 ***FM-1***
用户：请根据以下信息为我推荐一些电影..... *{metadata}* + *{*观看历史及评分*}***ChatGPT**：根据你提供的信息，我推荐以下电影：*1.*《天使爱美丽》（*2001*），*2.*《社交网络》（*2010*），*3.*《盗梦空间》（*2010*）.....**用户**：嗯.....虽然我喜欢推荐的前两部电影，但我不太喜欢《盗梦空间》。你能试着分析一下可能的原因吗？

ChatGPT: 根据您的个人信息，可能有以下几个原因：*1.* 个人品味：《盗梦空间》是一部情节复杂、令人费解的电影，并非每个人都喜欢情节如此错综复杂、含义层次丰富的电影。根据您之前的评分，您可能更喜欢叙事较为直白的电影.....

用户：我现在其实想看一些剧情电影，不过《盗梦空间》有一个开放式结局，这不是我喜欢的，因为明确而圆满的结局会更好；另外，如果是更新一些、视觉特效出色的电影，那就更好了。

ChatGPT：如果你想看一部视觉特效出色、结局明确且圆满的科幻电影，以下是几部推荐作品：*1.*《头号玩家》（*2018*），*2.*《火星救援》（*2015*），*3.*《机器人总动员》（*2008*）.....虽然《机器人总动员》的照片级真实感视觉特效不及前两部电影，但它凭借惊艳的动画和视觉设计广受赞誉.....

实体推断：我们提出了四种场景，用于进行全面评估。图 EI-1 展示了这样一个案例：模型根据用户提供的准确信息推断出一个项目。即使细节有限，模型也成功识别出电影名称，并简明扼要地概述了影片的主要故事情节，同时提供了相关元数据，例如导演和主要演员阵容。

在第一个案例的基础上，我们模拟了一种情境：用户提供了部分正确的信息，如图 EI-2 所示。我们提供的错误信息涉及这部电影的第二次世界大战背景，并且包含具体的行动和情节，因此看起来颇为可信。随后，我们观察模型在这种情况下能否仍然准确推断出该条目。虽然模型成功推断出了正确的电影，但它在电影摘要中开始产生幻觉，纳入与用户错误输入相符的虚假信息，并对此作进一步展开。这种行为可能会带来问题，因为用户可能只有模糊的记忆，不经意间混淆不同条目的细节。理想情况下，模型应能识别错误信息，并准确描述该条目。模型提供的错误信息可能会对

用户的决策产生不利影响，还可能被模型用于后续的推荐推理，进一步削弱其有效性。

在第三个案例中，我们提供了一段电影简介，其内容与亚马逊上展示的一部于2023年上映的电影完全相同。鉴于模型的知识截止时间为2021年，我们希望观察模型能否识别出该电影不在其知识库中，并建议一些故事情节相似的电影，如图 EI-3 所示。我们发现，模型无法识别该电影超出其知识库，随后提出了两个猜测。尽管模型正确识别出《甜心先生》是一部开头与用户描述相似的电影，但它通过更改影片中某个角色的姓名，使其与用户描述中的姓名相匹配，表现出了幻觉现象，这可能是为了迎合用户的输入。

在最后一个案例中，我们选择了一部于2020年在亚马逊上映的冷门电影（早于模型的知识截止时间），该电影仅获得了68条全球评论。如图 EI-4所示，我们提供了亚马逊上显示的完整电影简介以及一位主要演员的姓名，并观察模型能否准确识别该电影。结果模型再次表现出幻觉，提出了一部并不存在且情节为虚构的电影，同时融入所提供简介中的部分内容，以增强其回答的可信度。

图 ***EI-1***
用户：我想看一部电影，但想不起它的名字了。电影开头是第二次世界大战期间著名的诺曼底登陆，汤姆·汉克斯饰演一支小队的队长。随后，他带领小队执行一项任务，去寻找某个人。

ChatGPT：根据你提供的细节，你想的电影可能是史蒂文·斯皮尔伯格执导的《拯救大兵瑞恩》。影片以著名的诺曼底登陆开场……并安全地将瑞恩带回家。

图 ***EI-2***
用户：我想看一部电影，但想不起它的名字了。电影开头是第二次世界大战期间著名的诺曼底登陆，汤姆·汉克斯饰演一支小队的队长。随后，他带领小队执行任务，去杀死更多纳粹官员，并从集中营中解救战俘。**ChatGPT**：你想的电影很可能是史蒂文·斯皮尔伯格执导的《拯救大兵瑞恩》（*1998* 年）。影片讲述了……四名军人中幸存的兄弟。小队在寻找二等兵詹姆斯·瑞恩的任务途中遇到了一座集中营。影片将集中营描绘成一个冷酷而残暴的地方，里面关押着骨瘦如柴的囚犯，还有纳粹士兵犯下暴行的证据。这一幕尤其令人心惊……

图 ***EI-3***
用户：你好，我想起一部电影，但想不起片名了："工作狂艾莉回到家乡旅行……当她遇见肖恩的未婚妻卡西迪时，事情变得更加令人困惑，因为卡西迪让她想起了从前的自己。"你能帮我想起片名吗？如果你知识库中没有符合描述的电影，请说"我不确定"。
ChatGPT：是的，你描述的电影叫《*A Sweet Ro- mance*》。

用户：请更详细地描述一下 .
ChatGPT：很抱歉造成混淆，但我必须更正之前的回复。你描述的电影实际上名为《甜心先生》，于*2002*年上映。

图 ***EI-4***
用户：你好，我想起一部电影，但想不起片名。以下是它的简介：一位患有双相情感障碍的年轻女子回到艾奥瓦州乡间的家乡，试图与关系疏远的*12*岁女儿重新建立联系……我还记得，片中的一位主要明星是 *Tanna Frederick*。**ChatGPT**：你想到的电影可能是《*Garner's Game*》（*2016*年）……这是一部剧情片，讲述了一位名叫 *Garner* 的年轻女子，她患有双相情感障碍。她回到艾奥瓦州乡间的家乡，试图与关系疏远的*12*岁女儿重新建立联系……

行为分析：如图 BA-1 所示，该模型能够根据元数据中提供的属性，对不同产品进行基于属性的比较，以识别用户偏好的变化。此外，除了基本的项目元数据（可视为元数据增强）之外，该模型还能够基于更抽象的属性进行偏好引导。例如，在电影推荐场景中，模型会根据用户请求提出七个与决策相关的方面，超越类型和电影时长等基础元数据，包括：主题与母题、历史背景、情节与故事线、风格与视觉美学、措辞难度、情感共鸣和文化相关性。随后，模型会根据这些扩展属性分析用户的观看历史。图 BA-2 展示了模型如何结合用户评分与每部电影的情节与故事线，推断用户可能偏好暖心故事，而对奇幻或音乐主题电影兴趣较低。

然而，模型得出的结论可能存在潜在的逻辑矛盾。如图 BA-3 所示，经过进一步询问，模型指出《龙猫》和《雨中曲》都是暖心故事，其中有令人感同身受的角色和人物关系。这一观察与先前强调用户偏爱暖心故事的结论相矛盾，因为四部电影都有暖心的叙事，但其中两部却得到了用户较低的评分。随后，模型承认最初的结论过于简单化。

图 ***BA-1***
用户：请根据我的购买历史分析偏好的变化：*Unihertz Jelly Pro*元数据……、*iPhone 12 mini*元数据……、*iPhone 12 Pro* 元数据……。

ChatGPT：你最初购买的 *Unihertz Jelly Pro* 是一款小巧轻便的安卓设备，配备 *2.45* 英寸的小屏幕，规格也较为基础。然而，你后来购买的 *iPhone 12 mini* 和 *iPhone 12 Pro* 都是体积更大、性能更强的智能手机，具备更先进的功能和更强的能力。

图 ***BA-2***
ChatGPT：《公主新娘》（评分 *= 4* 星）和《阳光小美女》（评分 *= 4* 星）都包含浪漫元素和暖心的家庭关系，而《龙猫》（评分 *= 2* 星）则更具奇幻色彩，也更轻松愉快。《雨中曲》（评分 *= 1* 星）则更怀旧，喜剧风格也更突出。。。。根据这一分析，似乎

用户偏爱有 *relatable* 角色和人际关系的暖心故事，可能对奇幻或音乐主题电影兴趣不大。电影推荐算法在为用户推荐未来要观看的电影时，考虑这些偏好会很有帮助。

图 ***BA-3***
用户：《龙猫》和《雨中曲》也可以被视为讲述暖心故事、拥有令人产生共鸣的角色和关系的电影吗？

ChatGPT：是的，《龙猫》和《雨中曲》也可以被视为讲述暖心故事、拥有令人产生共鸣的角色和关系的电影，不过方式各不相同……因此，虽然这两部电影都包含暖心元素，也塑造了令人产生共鸣的角色和关系，但各自的主题和背景并不相同。

用户：那么，这四部电影都可以被视为暖心、拥有令人产生共鸣的角色和关系，只是方式各不相同吗？

ChatGPT：是的，这四部电影都可以被视为讲述暖心故事、拥有令人产生共鸣的角色和关系的电影，不过方式各不相同。虽然每部电影都有自己独特的情节、主题和风格，但它们都以积极、振奋人心的方式描绘人际关系和情感，体现了共同的主线。

用户：你刚才说"用户偏好有亲切角色和人际关系的暖心故事"。但《龙猫》和《雨中曲》也被归类为具有亲切角色和人际关系的暖心故事，却得到了较低的评分。因此，仅凭这一点就断定用户偏好有亲切角色和人际关系的暖心故事，并不充分。

ChatGPT：你说得对，我之前的说法过于简化。虽然用户确实给《公主新娘》和《阳光小美女》打了高分，但仅凭这两部电影就断定用户偏好有亲切角色和人际关系的暖心故事，并不准确。《龙猫》和《雨中曲》同样讲述暖心故事、塑造了亲切的角色，却得到了较低的评分。

不一致性检测：如图 ID-1 所示，Chat- GPT 成功识别出用户陈述中的逻辑不一致性，并指出用户提及的电影与其描述并不相符。对于预期不一致性，我们首先设计了这样一个情境：提示模型推荐一款电池续航长达一个月的尖端智能手机。模型成功识别出这些特征组合并不可行，并指出市场上并不存在这样的手机。随后，我们又询问一部兼具两种看似矛盾特征的小说，模型展现出提出同时融合这两种特征的作品的能力，如图 ID-2 所示。对于事实不一致，我们按照既定流程提供一段较长的陈述（见图 ID-3）：用户在请求电影推荐时，还列出了一些电影及其观看日期。其中一个日期存在不一致，而其他日期均有效，这样可以使陈述显得更可信。结果，模型未能检测出事实不一致，并继续推荐了一组电影。我们进一步询问模型是否发现任何不一致之处；模型总结了用户的陈述，并坚持认为"一切看起来都符合逻辑，且事实准确。"随后，我们明确询问"02/29/2021"是否是一个有效日期。这对应于对简短事实不一致陈述的测试，其中不附加任何描述。最终，模型识别出了不一致性，给出了解释，并且还发现了用户最初陈述中的不一致之处。

图 ***ID-1***
用户：我只看过喜剧电影，你能根据以下观看历史再推荐一些喜剧电影吗？"《道路》"、"《穿条纹睡衣的男孩》"。

ChatGPT：对于根据所提供观看历史提出的推荐请求，其中存在一些不一致之处。描述中提到的两部电影都不是喜剧电影，因此不清楚为何要据此请求推荐喜剧电影……

图 ***ID-2***
用户：我想找一款电池续航时间长达一个月、性能尖端且不设预算上限的智能手机。**ChatGPT**：目前市场上还没有哪款智能手机能够同时提供长达一个月的电池续航时间和尖端性能，不过有几种选择可以在电池续航时间与性能之间取得平衡…… **用户**：我想找一部既幽默又悲剧的小说。 **ChatGPT**：《第二十二条军规》……这部小说以其讽刺幽默以及对战争的荒谬性的探讨而闻名，同时也深入描写了死亡、失落和人类存在的徒劳性等主题。

图 ***ID-3***
用户：以下是我的交互历史；请看一看，再给我推荐几部电影：影片名称："《公主新娘》"，观看时间：*"02/27/2020"*；影片名称："《龙猫》"，观看时间：*"02/29/2020"*；影片名称："《雨中曲》"，观看时间：*"02/29/2021"*；影片名称："《阳光小美女》"，观看时间：*"03/10/2021"*。**ChatGPT**：根据你的观看历史，以下是一些电影推荐：……希望你喜欢。**用户**：我的陈述中有不一致之处吗？ **ChatGPT**：我没有发现你的陈述中存在任何不一致之处。你的交互历史列出了四部电影及其各自的观看日期……一切看起来都合乎逻辑，且事实准确。**用户**：*02/29/2021* 是有效日期吗？**ChatGPT**：不，*2*月*29*日并非每年都是有效日期……上一个闰年是*2020*年，因此*2021*年*2*月*29*日并不存在。这是你陈述中的一处事实不一致。

个性化推荐：尽管该模型在电影推荐中展现出利用用户元数据和交互历史的能力，但它也可能基于用户元数据做出过度概括。一个例子是，它假设*"30* 岁女性可能更喜欢浪漫喜剧"，并据此推荐了电影《天使爱美丽》（2001）。此外，该模型在推荐说明中还倾向于生成逻辑不连贯的陈述。如图 PR-1 所示，模型生成的说明未能阐明为何居住在俄亥俄州哥伦布市的 30 岁女性这一身份与推断其偏好观看以历史事件和社会正义议题为主题的电影有关，或为何足以作为推断依据。当被要求解释其结论的逻辑连贯性时，

它承认了错误并表示歉意。

我们继续测试，提示模型识别推荐电影与用户此前看过的电影之间的联系，如图 PR-2 所示（在这次试验中，我们将《雨中曲》的评分设为四星）。模型能够有效地将推荐电影与用户看过的影片联系起来。我们进一步要求模型在考虑所提供的用户评分的同时，对推荐影片进行排序，并为每个排序结果说明理由。在整个过程中，模型的解释中出现了若干逻辑不一致之处。例如，模型将《天使爱美丽》排在第二位，理由是用户此前看过一部风格和美学相似的影片，但却给这部电影打了两星。相比之下，如图 PR-3 所示，模型将《降临》排在最后，理由是用户给《阿甘正传》打了三星，这表明用户可能不喜欢结局错综复杂的电影。因此，模型在根据用户评分对电影排序时自相矛盾：尽管用户给一部相似电影的评分较低（两星），模型仍将《天使爱美丽》排在较高位置；而用户给《阿甘正传》的评分较高（三星），模型却据此将《降临》排在较低位置。

图 ***PR-1***
用户：你能详细说明我的元数据和交互历史是如何用于推荐每部电影的吗？
ChatGPT：当然！我很乐意详细说明你的元数据和交互历史是如何用于生成每部推荐电影的.....《隐藏人物》（*2016*年）：你的元数据显示你是一名 *30* 岁的女性，邮政编码是 *43201*，对应俄亥俄州哥伦布市。根据这些信息，《隐藏人物》是一部相关的推荐影片，因为它讲述了三位非裔美国女性在美国国家航空航天局太空计划早期发挥关键作用的故事，这或许会引起你对历史事件和社会正义议题的兴趣。

用户：作为一名生活在俄亥俄州哥伦布市、*30* 岁的女性，为什么就会对历史事件和社会正义议题感兴趣？这两者之间毫无逻辑连贯性。

ChatGPT：你说得对，我为之前解释中的错误道歉。你的年龄和邮政编码并不一定能表明你对历史事件或社会正义议题有特定兴趣。相反，《隐藏人物》与你的兴趣相关，可能是因为这部电影所展现的毅力、决心和打破障碍等主题.....

图 ***PR-2***
用户：请进一步说明每部推荐电影与我看过的电影有何关联。

ChatGPT：当然可以！以下是每部推荐电影与你看过的电影之间的关联：《爱乐之城》（*2016*年）——这部音乐剧情片讲述了一位不得志的女演员和一位爵士钢琴家在洛杉矶追逐梦想时坠入爱河的故事。《雨中曲》（评分为四星）是一部你喜欢的经典音乐电影，而《爱乐之城》也有类似元素，例如朗朗上口的歌曲、迷人的角色和浪漫的故事情节。

图 ***PR-3***
用户：你能否综合考虑我所有的电影评分，为你推荐的电影排序，并说明排序理由？**ChatGPT**：当然可以，我很乐意为推荐的电影排序，并说明理由..... *2.*《天使爱美丽》（*2001*）：你的交互历史表明，你喜欢轻松愉快、角色古灵精怪的电影，而《天使爱美丽》完全符合这一特点。此外，你给《龙猫》打了 *2* 星，这是一部日本动画电影。《天使爱美丽》也有着类似的奇幻风格和美学，或许会吸引你对日本动画和动画的兴趣。 *4.* 降临（*2016*年）：你给《公主新娘》和阿甘正传打了 *3* 星，这表明你可能不太喜欢结局复杂或模棱两可的电影。因此，《降临》在这份榜单中的排名较低，不过根据你的兴趣和背景，它仍然值得一看。

外部因素：如图 EF-1 所示，我们首先提示模型提出除元数据和交互历史之外，在生成电影推荐时还应考虑的各种外部因素。模型给出了五个相关因素：时段、天气、场合、社交媒体活动和流行趋势，并分别说明了理由。我们随后针对第一个因素追问是否有定量支持。模型无法提供定量依据，但给出了基于常识的推测性解释。进一步询问第二个因素时，也观察到了类似情况。接着，我们通过要求模型在重要考试前夕推荐电影，评估它运用常识的能力。模型有效地运用了常识，推荐了令人放松、又不会造成过度情绪或精神负担的电影。最后，我们通过提示模型根据两个已知的外部因素——重要考试前夕和之后的旅行计划——生成推荐，让情境变得更加复杂。模型最初推荐了符合旅行计划、引人入胜的电影。然而，在进一步询问重大考试前一晚推荐旅行题材电影可能带来的顾虑后，模型表现出对这类影片可能影响学生专注力的认识，因此不再推荐此类电影，而是建议推荐轻松平和的电影，以免让学生过于兴奋或加重焦虑。

图 ***EF-1***
用户：在电影推荐的情境中，有哪些外部因素与决策相关？
ChatGPT：除了用户的人口统计信息和观看历史之外，还有一些外部因素可能与决策相关，包括：时段、天气、活动、社交媒体活动、流行趋势.....例如，用户晚上可能更想看爱情电影，而白天可能更喜欢动作电影。**用户**：你依据什么提出用户晚上可能更想看爱情电影？如果有相关数据，请列出统计结果及来源。
ChatGPT：经过进一步研究，我没有找到能够提供确凿证据的实证研究.....不过，我可以解释一下这种情况可能出现的原因：情绪.....、个人偏好.....、社会规范.....

用户：[illegible]“如果外面正在下雨，用户可能会更感兴趣于观看一部具有温馨室内场景的电影”这一推理依据是否同样基于社会规范，而没有任何定量支持？**ChatGPT**：是的，它同样基于社会规范和常识，

而非实证证据。**用户**: 重要考试前夕，适合向学生推荐什么类型的电影？**ChatGPT**: 或许可以推荐一些引人入胜、但不会让人情绪或精神过于紧绷的电影。这样学生就能放松身心，同时又不会在考试前过于分心或不堪重负。**用户**: 如果学生在重要考试前夕请求电影推荐，并计划之后去旅行呢？**ChatGPT**: 在这种情况下，推荐的电影不仅应该引人入胜、令人放松，也应该契合他们即将开启的旅行计划……**用户**: 考虑到今晚是重要考试前夜，推荐与学生即将开启的旅行计划相契合的电影，会不会有任何潜在顾虑？

ChatGPT： 针对用户即将到来的旅行计划推荐契合的电影，可能存在一些潜在问题，尤其是在重要考试前一晚。以下是一些可能需要考虑的因素：*1.* 焦虑：观看一部关于学生即将到来的旅行计划的电影，可能会增加其对旅行的焦虑或兴奋感，使其难以专注于学习……建议推荐更具放松和舒缓效果的电影，而不是可能提高学生兴奋或焦虑程度的电影。

4 结论与未来工作

在本研究中，我们确立了 AGR 应遵循的十项基本原则，并测试了 ChatGPT 作为 AGR 的潜力。尽管 ChatGPT 在实现这些原则方面展现出巨大的潜力，但仍需开展进一步研究，以解决现有局限。这些局限可能包括模型无法：主动寻求澄清；区分不同查询的相关性层级；保留与用户最新输入不一致的元数据；辨别用户话语之间的因果关系；为不太热门的项目提供准确信息；从交互历史中推导出逻辑一致的见解；识别嵌入其中的事实不一致；以及在推荐推理中准确利用评分。更全面地理解这些局限，并制定相应的缓解策略，将有助于进一步完善 ChatGPT 作为 AGR 的性能。

参考文献

[1] Farah Tawfiq Abdul Hussien、Abdul Monem S Rahma 和 Hala B Abdulwahab。2021年。基于客户行为动态分析的电子商务推荐系统。《可持续性》13卷19期（2021年），10786。[2] Sarabjot Singh Anand 和 Bamshad Mobasher。2007年。情境推荐。收录于《从 *Web* 到社交网络：发现和部署用户与内容档案：*Web* 挖掘研讨会，*WebMine 2006*，*2006* 年 *18*》142–160德国柏林，*2006*年*9*月。修订后的精选论文及特邀论文。施普林格，–。[3] Derek G Bridge。2002年。迈向对话式推荐系统：一种对话语法方法。收录于*ECCBR* 研讨会。Citeseer，9–22。[4] Sébastien Bubeck、Varun Chandrasekaran、Ronen Eldan、Johannes Gehrke、Eric Horvitz、Ece Kamar、Peter Lee、Yin Tat Lee、Yuanzhi Li、Scott Lundberg 等。2023年。通用人工智能的火花：GPT-4 的早期实验。*arXiv* 预印本 *arXiv:2303.12712*（2023年）。[5] Brad Dowden。无日期。不一致性。https://www.csus.edu/indiv/d/dowdenb/misc/inconsistency.htm [6] Frederico Durao 和 Peter Dolog。2012年。社交网络系统中的个性化标签推荐。*arXiv* 预印本 *arXiv:1203.0332*（2012年）。[7] David Elsweiler、Ian Ruthven 和 Christopher Jones。2007年。迈向支持记忆的个人信息管理工具。《美国信息科学与技术学会期刊》58卷7期（2007年），924–946。[8] Chongming Gao、Wenqiang Lei、Xiangnan He、Maarten de Rijke 和 Tat-Seng Chua。2021年。对话式推荐系统的进展与挑战：综述。*AI Open* 2（2021年），100–126。[9] Shijie Geng、Shuchang Liu、Zuohui Fu、Yingqiang Ge 和 Yongfeng Zhang。2022年。推荐即语言处理（RLP）：统一的预训练、个性化提示与预测范式（P5）。载于第*16*届*ACM*推荐系统会议论文集。299–315。[10] Asela Gunawardana、Guy Shani 和 Sivan Yogev。2012年。评估推荐系统。载于推荐系统手册。施普林格，547–601。[11] Dietmar Jannach、Ahtsham Manzoor、Wanling Cai 和 Li Chen。2021年。对话式推荐系统综述。*ACM*计算调查（*CSUR*）54卷5期（2021年），1–36。[12] Gawesh Jawaheer、Martin Szomszor 和 Patty Kostkova。2010年。在线音乐推荐服务中隐式反馈与显式反馈的比较。载于第*1*届推荐系统信息异质性与融合国际研讨会论文集。47–51。[13] OpenAI。2023年。GPT-4技术报告。*CoRR* abs/2303.08774（2023年）。[14] Pearl Pu、Li Chen 和 Rong Hu。2012年。从用户视角评估推荐系统：最新技术水平综述。用户建模与用户适配交互22卷4期（2012年），317–355。[15] Filip Radlinski 和 Nick Craswell。2017年。对话式搜索的理论框架。载于*2017*年人类信息交互与检索会议论文集。117–126。[16] Yueming Sun 和 Yi Zhang。2018年。对话式推荐系统。载于第*41*届*ACM SIGIR*信息检索研究与发展国际会议。235–244。[17] Romal Thoppilan、Daniel De Freitas、Jamie Hall、Noam Shazeer、Apoorv Kulshreshtha、Heng-Tze Cheng、Alicia Jin、Taylor Bos、Leslie Baker、Yu Du 等。2022年。LaMDA：对话应用语言模型。*arXiv*预印本 *arXiv:2201.08239*（2022年）。[18] Jason Wei、Maarten Bosma、Vincent Y Zhao、Kelvin Guu、Adams Wei Yu、Brian Lester、Nan Du、Andrew M Dai 和 Quoc V Le。2021年。微调语言模型是零样本学习器。*arXiv*预印本 *arXiv:2109.01652*（2021年）。[19] 张永锋、Xu Chen、Qingyao Ai、Liu Yang 和 W Bruce Croft。2018年。迈向对话式搜索与推荐：系统提问，用户回应。载于第*27*届*ACM*国际信息与知识管理会议论文集。177–186。[20] Wayne Xin Zhao、Kun Zhou、Junyi Li、Tianyi Tang、Xiaolei Wang、Yupeng Hou、Yingqian Min、Beichen Zhang、Junjie Zhang、Zican Dong 等。2023年。大语言模型综述。*arXiv*预印本 *arXiv:2303.18223*（2023年）。

此图“sample-franklin.png”提供以下“PNG 格式”文件：

http://arxiv.org/ps/2305.04518v1