arXiv:2603.07605v1 [cs.IR] 8 Mar 2026

AI Box 美团

推荐系统的深度研究

Kesha Ou[1*]、**Chenghao Wu**[1*]、**Xiaolei Wang**[1*]、**Bowen Zheng**[1*]、
Wayne Xin Zhao[1†]、**Weitao Li**[2],**Long Zhang**[2],**Sheng Chen**[2],**Ji-Rong Wen**[1]中国人民大学高瓴人工智能学院，[2]美团。keishaou@gmail.com，batmanfly@gmail.com，

摘要

推荐系统的技术基础已从协同过滤发展到复杂的神经模型，并在近期进一步演进至大语言模型。尽管技术取得了这些进步，已部署的系统往往仍无法充分满足用户需求，只是简单地呈现商品列表，将探索、比较和综合信息的负担完全交给用户。本文认为，这种传统的"工具型"范式从根本上限制了用户体验，因为系统扮演的是被动过滤器，而非主动助手。为解决这一局限，我们提出了一种面向推荐的全新深度研究范式，以全面的、以用户为中心的报告取代传统的商品列表。我们通过**RecPilot**实例化了这一范式：这是一个多智能体框架，由两个核心组件构成：用户轨迹模拟智能体，能够自主探索物品空间；以及自进化报告生成智能体，能够将发现的信息综合为连贯、可解释的报告，以支持用户决策。该方法将推荐重新定义为一种主动的、由智能体驱动的服务。在公共数据集上的大量实验表明，RecPilot不仅能够在用户行为建模方面取得出色性能，还能生成极具说服力的报告，显著降低用户评估商品时的负担，从而验证了这一新型交互范式的潜力。

1 引言

推荐系统旨在帮助用户从海量项目中发现相关内容，从而缓解信息过载问题。为了准确捕捉和建模用户偏好，这些系统经历了多个关键技术发展阶段。早期方法主要基于协同过滤（CF）技术 [1]。随着深度学习的快速发展，各种神经网络架构被广泛采用，以增强用户表示学习并实现更稳健的用户—物品匹配。这些架构包括多层感知机（MLP） [2], 循环神经网络（RNN） [3], 卷积神经网络（CNN） [4], 和 Transformer模型 [5]。近年来，大语言模型（LLM） [6] 的兴起开辟了新的发展方向，推动了生成式推荐范式 [7],、基于大语言模型的推荐框架 [8],以及智能推荐智能体 [9]的发展。

得益于这些进展，近年来推荐系统中的用户体验有了显著提升。然而，展望未来的技术趋势时，值得提出一个更根本的问题：现有推荐系统是否已足够成熟？通过这些系统，用户体验是否已得到充分优化？为了探讨这一问题，我们转换视角——暂且放下推荐算法研究人员的身份，转而从与这些系统交互的个体用户的角度出发。从这一视角来看，我们的亲身经验——

*贡献相同。
†通讯作者。

而这些经验很可能也为许多用户所共有——促使我们给出一个更为谨慎的回答。以电子商务平台为例。在这类环境中，用户的主要意图通常是找到并购买符合特定需求的商品。为了在现有推荐系统中实现这一目标，用户必须浏览曝光物品列表，点击可能相关的物品以查看详情，并综合收集到的信息，最终做出购买决策。尽管技术不断进步，用户在这一过程中仍未得到充分满足：挑选商品依然十分耗费精力，而对于高价商品而言，这一负担尤其沉重。

仔细审视这一负面发现后，我们发现了一个根植于推荐系统定位的核心局限：当前推荐系统的主流交互范式仍然侧重于曝光相关物品。尽管在推荐系统设计的早期阶段，这种做法或许合情合理，但随着技术发展，它可能会成为进一步改进的阻碍。具体而言，这种范式可能已经固化为一种主流思维定式，狭隘地塑造着我们思考和设计推荐系统的方式——这种思维定式预设用户必须主动参与决策过程。例如，它默认用户会比较呈现的物品并仔细审视其详情，然后才做出最终选择。更具体地说，现有推荐系统的功能更像是工具，而非助手——也就是说，它们帮助用户获取信息，却不会统筹完整的推荐流程，以直接满足用户的根本意图，或至少显著减少用户为此付出的努力。

为了重新思考下一代推荐系统的范式，鉴于搜索与推荐是两种高度相关的信息寻求方式，我们从深度研究 [10] 在信息检索领域近期取得的成功中汲取灵感。在传统搜索系统中，用户会看到大量网页，必须逐一阅读以收集相关信息。相比之下，深度研究系统会运用智能体自主与搜索引擎交互，代表用户从检索到的网页中收藏信息，并将结果综合成一份全面的报告供用户阅读。从技术角度看，这种信息综合方法同样可以应用于推荐系统。具体而言，系统而非用户可以承担耗时费力的物品探索过程，发挥类似于现有深度研究系统中搜索引擎的作用。此外，系统不必只生成一份曝光物品列表，还可以生成一份内容更全面的报告，直接呈现最相关的物品及满足用户需求所需的全部信息。这种方法将使推荐系统从被动的、由用户驱动的筛选工具转变为主动的、由智能体驱动的服务，大幅降低认知负荷和交互成本，同时通过综合生成的报告提供可解释且值得信赖的决策。

为此，我们提出一种适用于推荐系统的深度研究范式——一种新颖的交互方式，以综合报告取代传统的商品列表。为实现这一范式，我们引入**RecPilot**，这是一个代表用户探索商品池并生成面向研究的报告以支持决策的多智能体框架。RecPilot由两个核心组件构成：用户轨迹模拟智能体和自进化报告生成智能体。用户轨迹模拟智能体以上下文信息和历史用户行为为输入，在商品池中主动模拟探索轨迹。我们将这一模拟过程表述为生成式建模问题，并采用基于无模型过程奖励的强化学习来增强泛化能力。报告生成智能体利用模拟轨迹，生成旨在支持用户决策的可解释报告。具体而言，该智能体将用户兴趣分解为用户在决策过程中可能考虑的多个方面，为每个方面生成一个排序列表，并将这些列表整合为综合排序列表，以促进快速且明智的选择。为确保个性化，我们使用结构化评价标准和基于经验的记忆来刻画用户偏好，并引入一种自进化优化机制，随着时间推移从这两个维度持续改进报告生成。

我们在公开推荐数据集上评估了所提方法的有效性。实验结果表明，该方法在建模已观察到的用户行为方面表现更优，Recall@5最高提升了52%，验证了其代表用户探索候选物品的可靠性。此外，生成的报告在多个维度上均表现出色。尤其是，与最佳基线方法相比，这些报告往往能够提供超越表层偏好匹配的新颖商品推荐（占比达77%），凸显了该方法在发掘潜在用户偏好方面的独特价值。

我们的主要贡献总结如下：

• 据我们所知，本研究首次尝试从根本上革新推荐系统的交互界面——从传统的以商品为中心的列表，转向以用户为中心的决策支持报告。

• 为实现这一新范式，我们提出了一种多智能体框架，由其代表用户自主探索商品池，并将探索结果整合为连贯、可解释的报告，以促进知情决策。

• 在公开推荐数据集上开展的大量实验验证了我们方法的有效性。除了在建模观察到的用户行为方面表现出色外，我们还表明，生成的报告不仅具有出色的准确性和覆盖度，还能显著减少用户进行商品比较和评估所需付出的努力。具体而言，我们的报告能够提供超越表层偏好匹配的新颖商品推荐，彰显深度研究范式在推荐系统中的独特价值。

2 方法

本节将正式定义深度研究任务，并介绍我们为解决该任务而提出的多智能体方法。

2.1 方法概述

任务定义。 推荐系统的深度研究旨在降低用户在电子商务平台上的决策成本。具体而言，我们希望最大限度地减少用户进入平台后，直至购买到满意项目这一过程中产生的认知负担和操作负担。为实现这一目标，我们聚焦于两个子任务：用户探索模拟和探索报告生成。形式化地，我们将用户的历史行为表示为 $X = [(a_1、\ v_1).....、(a_t、\ v_t)]$,，其中每个元组 $(a_i、\ v_i)$ 表示一个商品-行为对。平台上的探索过程被建模为连续的行为会话 $Y = [(a_{t+1}、\ v_{t+1})$ $.....、(a_{t+T}、\ v_{t+T})]$,，其中 T 为探索过程中发生的行为数量。用户探索模拟旨在实现对商品池的高效遍历，将其形式化为条件生成任务 $g_{\text{exp}}(Y \mid X)$[3], ，其中g_{exp} 是探索函数，用于生成用户可能会发现高度相关的项目。基于模拟的探索轨迹，我们生成研究报告 $\mathcal{R}$，其形式化表示为 $\mathcal{R} \sim f_{\text{rep}}(\cdot \mid X、\ Y)$，其中 $f_{\text{rep}}(\cdot)$ 表示报告生成函数，可通过各种架构设计来实现。最终生成的报告以全面的解释和依据补充推荐商品集 Y，所有内容均基于用户的历史行为上下文 X，从而促进更加明智且自信的决策。

总体架构。 本文提出了**RecPilot**，一种代表用户探索商品池并生成探索报告、协助用户做出最终购买决策的多智能体框架。RecPilot由两个核心智能体组成：用户轨迹模拟智能体（第2.2节）和自进化报告生成智能体（第2.3节），分别实现探索函数$g_{\text{exp}}(\cdot)$和报告生成函数 $f_{\text{rep}}(\cdot)$。如图1所示，用户轨迹模拟智能体以上下文信息和历史用户行为作为输入，主动在商品池中模拟探索行为。通过将初始阶段低价值且重复的交互过程从用户转移出去，该智能体有效减轻了繁琐探索带来的负担。为对此模拟任务进行建模，我们采用生成式方法，并使用基于无模型过程奖励的强化学习，使模型在复杂上下文条件下具备更稳健的推荐性能。随后，模拟轨迹被传递给报告生成智能体，由其将轨迹转换为旨在支持决策的可解释报告。在此过程中，该智能体将用户兴趣分解为用户在评估过程中可能考虑的多个方面，为每个方面生成一个排序列表，随后将其整合为一个综合排序列表，以支持快速且知情的决策。为确保个性化，我们使用结构化评价标准和基于经验的记忆来刻画用户偏好，并引入一种自进化优化机制，随着时间推移从这两个维度不断改进报告生成。

[3]这里我们稍微滥用一下 条件“|”的概率记号，以表示 所涉及的上下文。

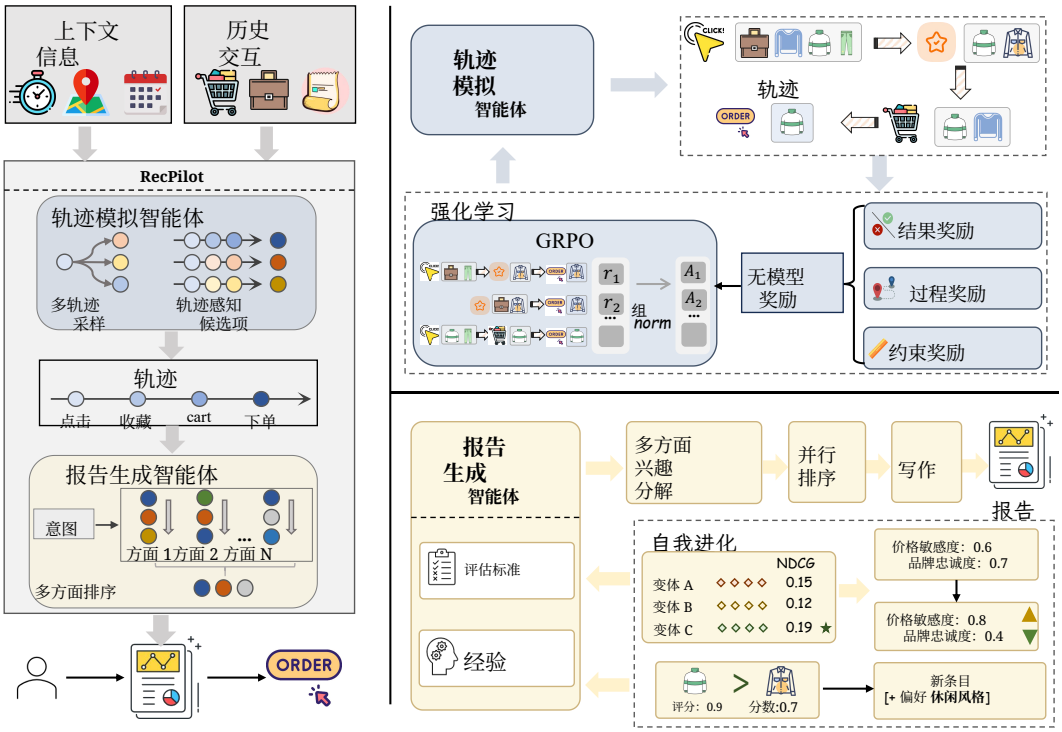


图 1：我们的方法 RecPilot 概览。左侧展示了整体流程。轨迹模拟智能体接收上下文信息和历史交互作为输入，并生成模拟轨迹，供报告生成智能体生成最终报告。右上部分介绍了轨迹智能体，该智能体交替进行项目生成和行为生成，并通过无模型奖励强化学习进行优化。右下部分展示了报告生成智能体，该智能体首先将用户兴趣分解为多个方面，以便进行排序，然后撰写报告。它通过评价标准和经验刻画用户偏好，并沿这两个维度实现自我进化以进行优化。

2.2 用户轨迹模拟智能体

在大规模电子商务场景中，用户通常面临较高的时间成本和探索疲劳，这阻碍了他们高效发现潜在感兴趣项目的能力。为应对这一挑战，我们提出了用户轨迹模拟智能体。通过主动在候选空间中模拟用户探索行为，该智能体充当代理，执行深入、定向且细粒度的交互——推断用户可能偏好的项目——从而减轻用户手动进行重复探索的负担。具体而言，该智能体首先从上下文信息和历史行为序列中学习并提取用户的动态兴趣与偏好。随后，它采用自回归生成范式，合成高质量、多样化且长度自适应的从探索到决策 轨迹。通过模拟这一行为路径，该智能体有效免除了用户手动筛选候选项目的需要。更重要的是，这些模拟轨迹充当明确的推理路径，促进智能体内部更深入的认知处理，从而更准确地检索相关候选项。因此，该智能体能够提供高质量的候选集，以支持下游决策模块。

2.2.1 生成式用户轨迹学习

本节提出一种用于学习用户轨迹的生成式框架。首先，我们将交互数据重构为具备会话感知能力的序列格式。在此表述的基础上，我们采用监督学习（SL），引导模型掌握从探索到决策的范式。

会话感知轨迹标记化。 在电子商务场景中，用户决策通常沿着从探索到决策的渐进轨迹展开。最终购买决策

对先前的探索行为表现出强烈的因果依赖，且不同行为类型天然呈现出参与层级（例如，购买 $>$ 收藏 $>$ 点击）。与将轨迹中的动作和项目展平为独立词元的传统方法不同 [11], 我们采用动作引导聚合方案，通过将动作-物品交互组织为前缀结构化片段，以更好地结构化用户轨迹。为此，我们首先将异构动作类型和离散物品 ID 映射到统一词汇空间。然后，对于具有相同动作类型的连续行为，我们以相应的动作词元作为前缀，其后按时间顺序排列所有物品词元。回顾可知，轨迹指的是连续的行为会话（第 2.1 节），因此，为进一步增强对会话边界的感知，我们在每条轨迹的前后分别添加特殊词元 <bos> 和 <eos> 。形式上，词元化轨迹表示为:

$$\tilde{S} = h(S) = \langle \texttt{<bos>}, a_{[1,2]}, v_1, v_2, a_{[3,4]}, v_3, v_4, \cdots, a_T, v_T, \texttt{<eos>} \rangle, \tag{1}$$

其中， $S = [(a_1, v_1).....(a_T, v_T)]$ 是原始用户会话，由序列动作-项目对组成（定义见第 2.1 节）； $h(\cdot)$表示所提出的词元化函数； $a_{[i,j]} =$ {<点击>, <购买>, <收藏>, <购物车>} 表示覆盖同类型第 i 次至第 j 次连续交互的动作词元； a_T 固定为 <购买> 动作，表示用户决策过程完成，并且当前会话终止。通过合并同类型的连续交互，这种格式大幅压缩了高频<点击> 词元。这种压缩减轻了长尾动作分布造成的干扰，促使模型关注内在的状态转移，而非仅仅复现频繁出现的动作词元。

多行为轨迹建模。 我们将用户轨迹建模表述为序列到序列的生成任务。给定词元化的用户历史序列 $X = h(X)$，模型经过训练后会自回归地生成后续的"探索到决策"轨迹（即一系列探索行为，例如点击和收藏，最终以购买告终），~记作 $Y = h(Y)$。为解决历史交互数据中固有的类别不平衡问题（频繁的点击行为远多于稀疏的购买行为），我们采用编码器-解码器架构（例如T5 [12]）。通过将历史上下文的编码与未来动作的生成解耦，我们避免解码器过度模仿历史模式，或使预测偏向高频行为。在此阶段，我们采用监督学习来优化模型，从而建立模拟用户探索轨迹的基础能力。优化目标如下:

$$\mathcal{L} = -\sum_{\langle \tilde{X}, \tilde{Y} \rangle \in \mathcal{B}} \sum_{t=1}^{|\tilde{Y}|} \log P(\tilde{Y}_t | \tilde{X}, \tilde{Y}_{<t}), \tag{2}$$

~ ~ ~其中， $\mathcal{B}$ 表示训练集。 Y_t 是目标会话 Y 的第 t 个词元， $Y_{<t}$ 表示 Y_t 之前的词元~。

2.2.2 基于无模型过程奖励的强化学习

监督学习（SL）使模型能够初步掌握用户从探索到决策的行为演变模式。然而，作为一种教师强制范式，它严格复现训练数据中观察到的模式，这往往会导致泛化能力有限、探索能力不足。这通常会造成泛化能力有限，且随机探索能力下降。为解决这些局限，我们引入强化学习（RL）阶段，进一步提升模型的泛化能力和探索表现。具体而言，我们通过随机采样生成多条探索轨迹，并引入无模型过程奖励来区分和评估这些轨迹的质量。最后，我们采用 GRPO算法 [13] ，以提升长期推理过程中生成轨迹的稳定性和质量。

无模型过程奖励。奖励函数是强化学习中的主要训练目标。近期研究凸显了基于规则的奖励的优势 [14, 15], : 这类奖励易于开发，也难以被利用。鉴于这些优势，我们采用了类似的无模型、基于规则的奖励方法。为了联合优化探索性模拟过程和

最终推荐质量，我们提出了一种复合奖励函数，由三个不同的组成部分构成：结果奖励、过程奖励和约束奖励。

• 结果奖励：为提升推荐性能，我们引入结果奖励，用于评估生成的轨迹是否会导向正确的决策。具体而言，当最终预测的项目 $\hat{v}_T$ 与实际购买的项目 v_T 匹配时，模型会获得正奖励，公式如下：

$$R_O = \mathbb{I}(\hat{v}_T = v_T), \tag{3}$$

其中，$\mathbb{I}(\cdot)$表示指示函数。该奖励直接与推荐系统的最终目标保持一致，为强化学习提供了明确的优化目标。然而，该奖励本质上较为稀疏，难以为中间阶段的探索行为提供有效指导。

• 过程奖励：仅依赖结果奖励不足以捕捉生成过程中不同探索轨迹之间的质量差异。因此，我们进一步引入过程奖励，用于衡量生成的探索轨迹与真实用户轨迹之间的语义一致性。这一设计基于电子商务场景中的一个关键观察：有效的用户探索模拟并不要求逐词元复现历史交互，而应保留所浏览项目在语义和协同关系上的一致性。严格的 ID 级匹配可能会抑制用户行为的自然多样性，并阻碍模型泛化到合理但未见过的轨迹。基于这一洞察，我们采用一种以协同一致性为基础的过程奖励，超越硬匹配。具体而言，我们利用在 SL阶段学习到的 ID嵌入来定义协同语义空间。对于生成轨迹中的每个有效物品词元，我们将其与真实轨迹中的每个有效物品词元进行配对，计算每一对的余弦相似度，并选取最大相似度分数作为该生成物品的个体奖励，随后对所有生成物品词元执行均值池化。因此，整体过程奖励通过最大相似度池化 [16]计算如下：

$$R_P = \frac{1}{|\mathcal{V}_{\hat{Y}}|} \sum_{\hat{v} \in \mathcal{V}_{\hat{Y}}} \max_{v \in \mathcal{V}_Y} \frac{\boldsymbol{e}_{\hat{v}} \cdot \boldsymbol{e}_v}{\|\boldsymbol{e}_{\hat{v}}\| \, \|\boldsymbol{e}_v\|}, \tag{4}$$

其中 $\mathcal{V}_{\hat{Y}}$ 表示生成轨迹 $\hat{Y}$ 中的项目，$\mathcal{V}_Y$ 表示真实轨迹 Y 中的项目。该设计可从两个角度加以论证。首先，同一交互阶段内项目之间的严格排序（例如，连续点击行为）往往较弱，且容易受到页面布局或推荐位置等外部因素的影响。其次，通过鼓励语义对齐而非精确的物品ID匹配，模型能够学习生成合理且多样化的探索轨迹，从而提升其对未见用户行为模式的泛化能力。

• 约束奖励：为进一步优化轨迹生成，我们引入轻量级约束奖励，以同时规范轨迹长度和格式。长度约束引导模型生成与用户实际探索深度相符的轨迹。它通过指数衰减函数惩罚生成轨迹长度 $L_{\hat{Y}}$ 与真实长度 L_Y 之间的偏差：

$$R_L = e^{-\mu |L_{\hat{Y}} - L_Y|}, \tag{5}$$

其中 μ 是此处设定为 0.2 的缩放因子。如第 2.2.1 节所述，有效轨迹必须遵循预定义的“探索—决策”顺序。为严格满足这一结构要求，我们引入格式惩罚 R_F，通过以下两条硬性规则约束生成过程。若轨迹出现以下情况，则视为格式错误：(1) 直接以 <购买> 动作开头（即探索步数为零）；或 (2) 连续出现相同的动作标记，中间没有项目，因此该步骤对实际物品探索毫无意义。相应的惩罚定义如下：

$$R_F = \begin{cases} 0 & \text{if the format is correct,} \\ -1 & \text{otherwise.} \end{cases} \tag{6}$$

最终的复合约束奖励为这两个组成部分之和:

$$R_C = R_L + R_F. \tag{7}$$

通过GRPO进行策略更新。 对于每个用户，我们通过从当前策略 π_θ 中采样，生成一组 G 轨迹，记为 $o_1, \ldots, o_G$。与轨迹 o_i 相关联的总奖励综合了前述三种奖励，具体如下:

$$r_i = R_O + R_P + R_C, \tag{8}$$

其中， R_O （公式（3））是用于评估最终预测的结果奖励， R_P （公式（4））是用于评估中间推理的过程奖励，而 R_C （公式（7））是用于执行格式规则的约束奖励。为在不引入单独价值网络的情况下减少方差，GRPO通过在组内对奖励进行归一化，为每条轨迹 A_i 估计优势:

$$A_i = \frac{r_i - \text{mean}(r_1, \ldots, r_G)}{\text{std}(r_1, \ldots, r_G)}. \tag{9}$$

随后，通过最大化截断目标函数来更新策略 π_θ :

$$L_{\text{GRPO}} = \frac{1}{G} \sum_{i=1}^{G} \Big(\min(r_i A_i,\, \text{clip}(r_i, 1-\epsilon, 1+\epsilon) A_i) - \beta\, \mathbb{D}_{\text{KL}}(\pi_\theta \,||\, \pi_{\text{ref}}) \Big), \tag{10}$$

其中 $r_i = \pi_\theta(o_i)/\pi_{\text{ref}}(o_i)$， π_θ 和 π_{ref} 分别表示当前策略模型和参考模型。 $\mathbb{D}_{\text{KL}}$ 项是KL正则化，用于防止策略过度偏离初始模型分布。

2.2.3 探索轨迹生成

在推理和生成过程中，模型充当用户代理，在大规模物品空间中执行模拟探索。这种模拟会产生一组多样化的从探索到决策的轨迹，涵盖从广泛、发散式浏览到逐步聚焦的意图收敛过程。具体而言，推理过程包含两个核心组成部分。

用于兴趣探索的多样化轨迹采样。 由于用户交互行为具有随机性和内在不确定性，贪心解码策略往往无法捕捉潜在用户意图的完整分布。尽管束搜索可以生成多个候选路径，但所得轨迹通常只有轻微差异，尤其是在末尾部分。这一局限阻碍了模型对用户兴趣多样化分支的表示能力，并导致对物品空间的探索不足。为了更好地模拟用户意图的演化过程，我们在解码过程中采用多轨迹采样方法。具体而言，我们采用 top-p采样策略，在每一步从累积概率质量超过阈值 p 的词元中进行采样。这样，我们便能并行生成 N 条彼此不同的探索轨迹，从而捕捉更广泛的合理用户行为路径。

轨迹感知候选生成。 基于上述生成的 N 探索轨迹，我们推断用户的意图收敛点，并预测最有可能触发决策行为（即 购买）的目标物品。对于每条生成的探索轨迹，我们利用其最后一步的隐藏状态来预测下一个项目，并检索排名前K 的相关候选项。该过程会生成一个大小为 $N \times K$的初始候选池。随后，我们将探索轨迹与最终预测的项目拼接起来，形成一条完整的模拟轨迹。该轨迹的对数似然将作为候选项目的最终分数。最后，在从候选池中去除重复项目后，我们选择分数最高的前K 个项目作为最终检索结果，并保留其关联的探索轨迹——即“从探索到决策”的轨迹。这些保留的轨迹可作为用户行为的高置信度模拟，随后将其传递给下游报告生成智能体，以进行进一步的意图分析和利用。

2.3 自演化报告生成智能体

本节介绍报告生成智能体的开发，该智能体旨在将模拟轨迹转化为人类可读报告，以支持决策。接下来，我们将详细说明如何生成报告，并通过自我进化提升报告质量。

2.3.1 智能体报告生成

如第 2.2.3 节所述，用户轨迹模拟智能体会生成多个用户可能感兴趣的候选项目。然而，若在报告中直接呈现所有候选项目，仍会给用户带来较高的认知负荷。因此，有必要先从候选集中选出最有希望的项目。这一筛选过程类似于传统推荐系统中的精排阶段：系统为每个候选项目分配分数，以确定最终的排序列表。受这一范式启发，我们对轨迹模拟智能体生成的候选项目进行最终排序。至关重要的是，排序时应综合考虑用户偏好的多个维度。为此，我们首先将用户兴趣分解为多个方面（即属性子集）。然后，针对每个方面分别进行排序，再整合各方面的排序结果，生成统一的最终列表。得到排序列表后，我们将其转化为人类可读报告。该报告不仅展示各个方面内及各方面之间排名靠前的候选项目，还以叙述性格式提供详细的解释性信息，以支持用户做出明智的决策。

量表—经验双通道偏好。 为实现有效排序并在报告中提供可解释的决策指南，我们提出从两个角度刻画用户偏好，即 结构化评价标准和文本经验。结构化评价标准构成用户偏好的框架。为支持项目间的数值比较，我们将项目属性用作所有用户通用的评价标准。对于每位用户 u，其偏好表示为针对每项属性 $a \in \mathcal{A}$ 的一组优先级分数，可记为 $\mathcal{W}_u = \{w_u^{(a)} \mid a \in \mathcal{A}\}$。作为结构化评价标准的补充，文本经验能够捕捉用户偏好的语义细微差别。由于基于属性的评价标准难以编码依赖上下文的偏好，文本经验便成为灵活且语义丰富的补充，用于捕捉隐含的、特定于场景的用户倾向。对于每位用户 u，我们维护一个经验记忆 $\mathcal{E}_u$，其中每条经验条目都以键值对的形式组织，上下文条件作为键，相关内容作为值。

用于并行排序的多方面兴趣分解。 为支持报告中的方面感知决策指南部分，我们首先基于上文介绍的评价标准和经验，将用户兴趣分解为多个方面，然后针对每个具体方面执行并行排序。在此，我们将用户偏好的一个方面与一个属性子集相关联，可记为 $\mathcal{D} = \{a \mid a \in \mathcal{A}\}$。每个属性 a 对用户 u 的重要性权重对应于上文定义的其评估标准分数 $w_u^{(a)}$。为将用户兴趣分解为多个方面，我们提出一种基于大语言模型的方法。具体而言，对于目标用户 u，我们首先提示大语言模型将模拟轨迹压缩为意图摘要。然后，我们将意图摘要作为查询 q，从经验记忆 $\mathcal{E}_u$中检索一定数量的经验。该检索过程通过计算意图摘要与 $\mathcal{E}_u$中每个条目的上下文条件之间的语义相似度来完成，随后选择最相似的条目。以意图摘要和检索到的经验作为上下文，我们利用大语言模型识别多个属性子集，这些子集正式构成用户偏好的分解方面。对于每个方面，我们取关联属性对应的评估标准分数，并将每个条目增加 δ，以强调其相关性。随后，我们并行地对每个方面执行排序。这里，我们以一个方面 $\mathcal{D}$ 为例，详细说明排序过程。对于每个候选项 i，我们首先提示大语言模型评估其与该方面关联的每个属性 a 的匹配度，并生成匹配分数 $s_i^{(a)}$。接下来，我们对每个属性应用评估标准分数，计算所有属性匹配分数的加权和。由此得到方面 $\mathcal{D}$的最终项目分数，表示如下:

$$s_i^{(\mathcal{D})} = \frac{1}{|\mathcal{D}|} \sum_{a \in \mathcal{D}} w_u^{(a)} \cdot s_i^{(a)}. \tag{11}$$

通过这种方式，我们完成了一个方面的排序。为了得出涵盖多个方面的排序列表，我们直接将所有方面的分数相加作为最终分数，并据此对项目进行排序。

结构化报告生成。 在获得每个方面内及跨方面的排序列表后，我们将其扩展为一份包含详细解释性信息、用于决策支持的人类可读报告。我们以结构化方式组织报告 $\mathcal{R}$，由四个部分组成：模拟轨迹，

用户意图摘要、整体项目建议，以及针对每个方面的项目建议。其表示如下：

$$\mathcal{R} = \left\{ h(\hat{Y}), h(q), h(L_0), h(L_1, \ldots, L_n) \right\}, \tag{12}$$

其中，$h(\cdot)$表示使用大语言模型将输入内容转换为报告组件的函数，$\hat{Y}$表示模拟轨迹，q表示从轨迹中总结出的用户意图，L_0表示涵盖所有方面的整体排序列表，而L_1、..... L_n则是对应各个方面的排序列表。对于最后一个组件，我们利用各个方面的属性生成相应的理由，为决策提供指导。

2.3.2 个性化自我演化

在上一节中，我们讨论了生成人类可读报告的方法。然而，现实推荐场景中的一项关键挑战在于，用户兴趣并非一成不变，而是会在多次交互过程中随时间动态演变和迁移。因此，在初始探索阶段捕捉到的偏好可能会逐渐过时。为持续提供准确且个性化的决策支持，推荐智能体必须能够不断学习并适应这些动态变化，这就需要一种演化机制。此外，为应对用户兴趣的变化，传统推荐系统通常需要人工收集新数据，并定期对底层模型进行重新训练或微调。为加快演化进程，我们提出了一种无需训练的自我演化机制。该系统不依赖人工更新模型，而是使智能体能够以实时闭环的方式自主优化用户的结构化评价标准和经验记忆（如第2.3.1节所定义）。通过在服务过程中持续收集并吸收用户反馈，该自我演化机制无需额外训练开销，即可确保个性化决策支持的及时性和准确性。

评估标准优化。为使评价标准与不断变化的用户兴趣保持一致，并减轻噪声的影响，我们提出一种 best-of-n 优化方法。回顾第2.3.1节，我们将用户兴趣分解为多个方面，并生成相应的排序列表。为了进行优化，我们首先使用 NDCG指标 [17]，根据高层用户行为[4]（即购买行为）评估每个排序列表。随后，我们找出 NDCG 分数最高的列表，并使用与其相关的评估标准分数（增加了 δ）更新原始值。通过仅选择表现最佳的变体，我们能够有效过滤掉次优或含噪的评估标准调整，确保更新后的分数严格符合高层行为偏好。

经验整合。为使经验记忆与不断变化的用户兴趣保持一致，我们提出一种对比策略，用于提取用户偏好线索并将其作为新的经验条目纳入其中。具体而言，我们将评估标准优化过程中取得最高 NDCG 分数的排序列表作为正样本，同时将根据优化前的评估标准分数生成的排序列表作为负样本。当与高层行为相关的项目在正样本中排名更高时，我们利用这两个排序列表构建对比样本对，并提示大语言模型生成纠正性经验条目，明确说明在特定场景下需要调整的决策逻辑。这种对比方法放大了次优的原始排序列表与优化后高性能排序列表之间的差异，从而能够有针对性且可解释地改进经验记忆。

扩展经验挖掘。上述优化依赖于包含高层行为（例如，购买行为）的用户会话。然而，在真实场景中，此类会话十分稀少。为解决数据稀缺问题，我们进一步纳入包含低层行为（例如，点击行为）的会话，并采用基于大语言模型的挖掘方法来丰富偏好学习。具体而言，我们提示大语言模型对每个低层会话进行推理，推断潜在的负面偏好或未满足的需求，以补充现有经验库。值得注意的是，我们不会使用这些会话来更新评价标准，因为评价标准以精确的数值形式编码用户偏好，而低层行为本身具有较强噪声。

[4]高层行为和低层行为根据发生频率加以区分，其中高层行为通常是用户意图的更强信号。

3 实验

3.1 实验设置

数据集。 我们在天猫数据集[5], 上评估模型，该数据集包含四种用户行为：点击、收藏、加入购物车和购买。我们将“购买”作为待预测的目标行为，并将其他行为视为决策前探索过程的一部分。我们选取两周的交互记录，并按天划分用户会话。为缓解数据稀疏性，我们过滤掉在训练集中出现次数少于五次的项目。处理后的数据包含 288,777 名用户、556,233 个项目和 1,294,696 个会话。数据集划分方面，我们遵循已有工作 [18]，采用按会话留一策略。由于我们的目标是预测购买行为，因此仅保留包含购买行为的会话用于验证。最终划分得到 306,088 个训练会话、1,941 个验证会话和 1,272 个测试会话。

基线方法。 本文考虑两个子任务：用户探索模拟和探索报告生成，如第 2.1 节所述。为了进行比较，我们选取了几种针对各子任务设计的代表性方法。

(1) 序列推荐方法:

•**SASRec** [19] 采用自注意力机制来捕捉用户交互序列中的长期依赖关系，配置为 2 层和 2 个注意力头。

• **BERT4Rec** [20] 采用双向 Transformer架构，通过完形填空任务学习深层行为表示，配置为 2 层、2 个注意力头，掩码比例为 0.2。

• **GRU4Rec** [21] 使用门控循环单元（GRU）来增强推荐任务中的长期记忆，采用2个循环层，dropout率为0.1。

（2）多行为序列推荐方法:

• **PBAT** [22]通过整合用户特定行为特征来识别个性化交互特征，参数设置为2层和2个注意力头。

• **MBHT** [23] 通过在超图结构中建模项目之间的交互，揭示项目之间隐藏的多行为关联，配置为2层、2个注意力头、掩码比例0.2，以及 [5, 8, 40]的多尺度超边。

• **MB-STR** [24] 通过采用稀疏的基于混合专家的选择机制来解决异构行为之间的不一致性，实现为2层、2个注意力头、3个共享专家和32个路由桶。

（3）推理 ning增强的序列推荐meth ods:

• **ReaRec** [25] 在最终预测之前引入潜在推理阶段，以分析用户历史记录背后的潜在动机，并配置为2个推理步骤。

（4）基于大语言模型的报告生成方法:

我们将这些基线方法分为两组。第一组是**直接推理**，其中包括领先的大语言模型，如 DeepSeek-V3.2 [26], GLM-5 [27], Qwen3-Max [28], 和GPT-5.2 [29]。这些模型接收与我们的方法相同的输入，并被直接指示生成决策报告。第二组包括**基于智能体的方法**，它们利用智能体工作流来处理复杂的生成任务。该组包括:

• **ReAct** [30]将推理与特定任务的行动交织进行，使模型能够根据观察结果动态调整其思路。

• **计划与求解** [31] 将复杂任务分解为可管理的子任务，并按照静态预定义计划依次求解。

评估指标。 对于用户探索模拟子任务，由于目标行为是与报告中的最终排序列表直接相关的最关键因素，我们重点关注对其的预测准确性。具体而言，我们采用召回率@k 和归一化折损累积增益@k （$k \in \{5, 10\}$）进行性能评估。

[5]https://tianchi.aliyun.com/dataset/140281

表 1：天猫 数据集上轨迹模拟任务的性能比较。最佳结果以粗体标出。

指标	序列			多行为			推理	Ours
	GRU4Rec	BERT4Rec	SAS4Rec	PBAT	MBHT	MBSTR	ReaRec	
Recall@5	0.0618	0.0527	0.0636	0.0551	0.1015	0.1025	0.0794	**0.1557**
召回率@10	0.0793	0.0708	0.0808	0.0726	0.1139	0.1175	0.0904	**0.1706**
NDCG@5	0.0470	0.0398	0.0526	0.0418	0.0793	0.0833	0.0645	**0.1241**
NDCG@10	0.0526	0.0456	0.0581	0.0476	0.0833	0.0880	0.0681	**0.1292**

对于探索报告生成子任务，由于报告质量的评估具有主观性，我们设计了六个重要维度，使其可衡量。每项指标的评分范围为1至5分，具体说明如下：

- **准确性**衡量审阅报告后所选项目是否与标签一致。
- **覆盖度**衡量报告是否涵盖并呈现了对决策过程具有重要影响的所有核心属性。
- **信息丰富度**量化报告中可验证的事实内容相对于缺乏依据或推测性内容的比例。
- **清晰度**评估报告是否提供了明确无歧义的比较和清晰的结论，从而有效地支持用户做出明智的决策。
- **一致性**衡量用户意图和项目排序列表各部分是否严格遵循模拟轨迹。
- **新颖性**评估报告是否提出了意图之外的额外兴趣，以及这些提议是否有可追溯的证据支持。

为基于这些主观维度进行评估，我们同时采用人类与大语言模型。对于人工评估，我们邀请十名真实用户对随机选取的样例进行盲测。为减轻主观偏差并确保可复现性，我们严格遵循人工评估协议，而非依赖整体印象。评估器须依据细粒度的五点评分量表，并针对六个维度中的每一个使用明确的行为锚点，独立为生成的报告评分。对于基于大语言模型的评估，我们采用 Gemini-3-flash [32] 作为模拟用户，并使用相同的评分指令对其进行提示。详细的评估指南、任务设置以及完整的评分标准见附录 A.1。

实现细节。 对于我们的用户轨迹模拟智能体，我们采用 T5 [12] 作为骨干模型。该模型配置为 2 个编码器层和 2 个解码器层。隐藏维度设为 64，前馈网络的中间维度为 256，并使用 ReLU 激活函数。对于因果自注意力，我们使用 2 个注意力头，每个头的维度为 64。我们使用 AdamW 优化器优化模型，批量大小为 32。初始学习率设为 1×10^{-3}，并采用带线性预热阶段的余弦退火调度。最终的模型检查点根据验证集上的 NDCG@10 指标选取。对于我们的自进化报告生成智能体，我们采用 DeepSeek-V3.2 [26] 作为骨干模型。为确保公平比较，所有方法均使用相同的解码设置进行评估：温度设为 0.2，最大生成词元数为 16384。我们采用 Qwen3-Embedding-8B [33] 作为经验检索的文本嵌入模型。为减轻格式差异对评估的影响，我们统一了所有方法的输出格式，要求每份生成的报告包含公式（12）中的四个组成部分。我们的代码可在 https://github.com/RUCAIBox/RecPilot 获取。

3.2 轨迹模拟任务评估

主要结果。 表 1 报告了我们提出的模拟器模型（第 2.2 节）在 TMALL 数据集上相对于各种竞争模型的性能，从中可以得出以下观察结果：

表 2：所提出的轨迹模拟智能体在 TMALL 数据集上的消融研究。“CR”和“PR”分别表示约束奖励和过程奖励。

模型	Recall@5	召回率@10	NDCG@5	NDCG@10
Ours	**0.1557**	**0.1706**	**0.1241**	**0.1292**
不含 CR	0.1439	0.1651	0.1142	0.1211
不含 PR	0.1454	0.1643	0.1165	0.1227
不使用 RL	0.1187	0.1376	0.0922	0.0982

首先，传统序列模型（例如，SASRec、GRU4Rec 和 BERT4Rec）表现出相对有限的性能。仅依赖统一的行为建模，限制了这些模型捕捉驱动复杂电子商务场景中多样化用户行为的潜在用户动机的能力。其次，多行为模型（例如，PBAT、MBHT 和 MBSTR）通过整合异构交互建模，持续优于传统基线方法。其中，MBSTR 取得了最强性能，表明对不同行为类型之间的相关性进行建模对于意图识别的重要性。对于基于推理的模型，ReaRec 通过引入潜在推理阶段，在最终预测之前分析用户状态，从而在多项指标上超越了 SASRec。然而，该方法仍然遵循判别式范式，即直接将历史行为映射到预测结果。

相比之下，我们提出的模拟器模型在所有评估指标上均取得了最佳表现。与现有方法不同，我们的框架明确模拟用户的决策轨迹，刻画从“发散探索”到“意图收敛”的动态演变。多行为模型虽然能够捕捉多样的交互信号，却无法对用户在做出最终决策前持续进行的兴趣探索过程建模。通过生成和分析多步用户轨迹，我们的模型能够解码购买决策前复杂的行为逻辑。在实证评估中，我们的方法显著优于最强的基线模型。这一结果凸显了明确建模“从探索到决策”过程对于提升推荐准确率的关键价值。

消融研究。为评估用户轨迹模拟器中各核心组件的作用，我们通过比较完整模型设计与若干变体开展消融研究。结果见表 2。

• **不使用 CR：**此变体移除了公式 (7) 中定义的约束奖励。约束奖励通过惩罚轨迹长度偏差和结构异常来确保逻辑连贯性。性能显著下降表明，生成轨迹的结构完整性对下游预测至关重要。这一发现说明，高质量的轨迹建模是捕捉复杂用户决策过程的必要前提。

• **不使用过程奖励：**该变体去除了公式 (4) 中的过程奖励，仅利用稀疏的结果奖励。性能下降表明，单靠结果奖励不足以区分发散探索轨迹的质量。过程奖励通过在潜在空间中优化协同一致性，而非强制进行僵化的令牌级匹配，提供了必要的中间引导，同时使模型能够泛化到多样且合理的探索行为。

• **不使用强化学习：**该变体跳过强化学习阶段，仅采用监督学习进行模型优化。结果表明，监督学习难以提升模型泛化能力并实现多样化探索。强化学习阶段通过多轨迹采样和策略优化，显著提升了轨迹生成的稳定性和逻辑一致性。

总体而言，消融结果表明，每个组件都发挥着关键且互补的作用。CR 确保结构连贯性，PR 引导语义意图收敛，而 RL 阶段则为长时域生成提供了必要的鲁棒性。这些组件协同作用，使我们的模型能够有效模拟“探索到决策”的轨迹，从而获得更优的推荐性能。

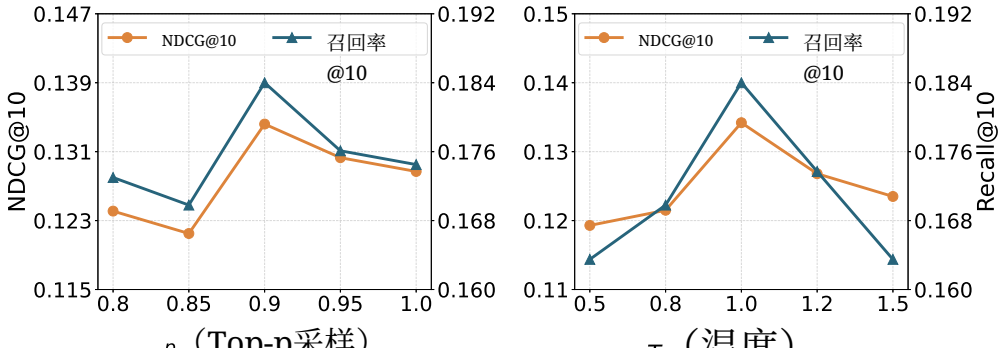


图3：TMALL 数据集上两个采样参数的性能比较：阈值p 和温度 τ。

轨迹模拟深度的影响。我们进一步研究了最大生成长度对模型性能的影响。如图2所示，轨迹长度与推荐性能之间存在明显的正相关关系。模拟轨迹的长度决定了推理过程的深度。

更长的轨迹使模型能够通过从广泛的类别级探索逐步过渡到细粒度的兴趣收敛，来模拟完整的决策过程。这种上下文信息的逐步积累能够实现更精准的意图推断。相反，缩短最大长度会严重限制可用的推理上下文，迫使模型在信息不足的情况下过早做出决策。这导致对用户真实意图的刻画不充分，进而造成性能下降。这些发现表明，显式的探索轨迹模拟能够帮助模型逐步捕捉复杂的用户偏好，从而增强其最终决策预测能力。

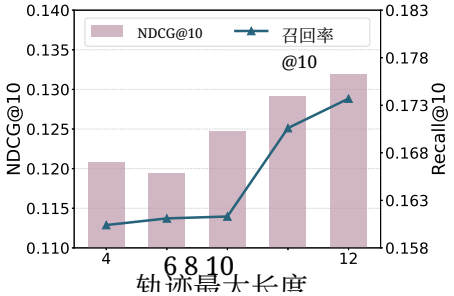


图2：TMALL 数据集上不同最大轨迹长度的性能比较。

超参数分析。 为研究轨迹多样性如何影响推荐性能，我们评估了两个关键采样参数的影响：令牌采样阈值 p 和温度 τ。结果如图3所示。

• **Top- 的影响p： </style}较小的 （例如，** $p = 0.8$）会过度限制采样空间，导致生成过程趋于保守，无法生成多样且信息丰富的候选轨迹。随着 p 增加，模型获得更大的探索自由，从而显著提升召回率和排序质量。然而，过大的 p会引入低概率噪声词元，损害模拟轨迹的语义一致性和稳定性，最终导致性能下降。

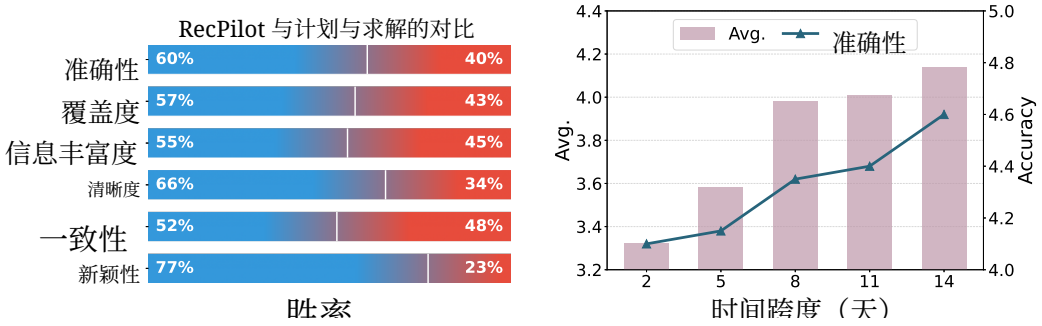
• **温度的影响** τ： </style}模型性能随 τ 变化呈现类似趋势，并在 $T = 1.0$时达到峰值。较低的取值会产生过于尖锐的概率分布，导致模型偏向高概率但可能并非最优的模式。适度提高 τ 有助于模型摆脱局部最优，并生成信息更丰富的探索轨迹。相反，过高的 τ会使分布过于平滑，由此产生的随机性会对推荐准确率造成不利影响。

3.3 报告生成任务评估

主要结果。 表3展示了我们提出的报告生成智能体（第2.3节）的自动评估与人工评估结果。通过将基线方法划分为直接推理模型

表3：TMALL 数据集上的报告生成任务性能对比。最佳和次佳结果分别以粗体和下划线标出。

评估器	方法	准确性	覆盖度	信息丰富度	清晰度	一致性	新颖性	Avg.
模拟用户	DeepSeek-V3.2	4.46	3.44	3.67	4.50	3.88	2.86	3.80
	GLM-5	4.38	3.30	3.47	4.24	3.70	2.73	3.64
	Qwen3-Max	4.05	3.24	3.66	4.48	3.78	2.91	3.68
	GPT-5.2	4.35	3.40	<u>3.77</u>	<u>4.67</u>	<u>3.91</u>	2.89	3.83
	ReAct	4.45	3.35	3.62	4.41	3.82	2.92	3.76
	计划与求解	<u>4.48</u>	<u>3.52</u>	3.76	4.55	3.90	<u>3.07</u>	<u>3.88</u>
	RecPilot	**4.60**	**3.62**	**3.80**	**4.80**	**3.94**	**4.09**	**4.14**
真实用户	DeepSeek-V3.2	<u>4.25</u>	3.42	3.79	4.35	<u>4.04</u>	2.95	3.80
	GLM-5	4.13	3.19	3.57	4.10	3.85	2.76	3.60
	Qwen3-Max	3.76	3.29	3.77	4.35	3.94	2.94	3.67
	GPT-5.2	4.07	3.41	**3.83**	<u>4.54</u>	**4.05**	2.97	3.81
	ReAct	4.17	3.26	3.65	4.20	3.98	2.95	3.70
	计划与求解	<u>4.25</u>	<u>3.45</u>	**3.83**	4.41	4.00	<u>3.13</u>	<u>3.84</u>
	RecPilot	**4.36**	**3.60**	<u>3.82</u>	**4.69**	4.00	**4.06**	**4.09**



（a）计划与求解与.....之间的成对比较我们的方法。

(b) 使用交互记录在不同时间跨度上进行训练。

图 4：关于 TMALL 数据集上报告生成任务的详细分析。

（例如，GPT-5.2、DeepSeek-V3.2）和基于智能体的框架（即，ReAct与计划与求解），评估揭示了若干关键见解。首先，基于智能体的方法在大多数指标上普遍优于直接推理基线方法。这验证了如下前提：综合生成一份全面的推荐报告是一项复杂的认知任务，其本质上更受益于迭代式的智能体工作流，而非单步的零样本提示。其次，尽管通用智能体带来了改进，我们提出的RecPilot仍始终取得最佳的整体性能。

具体而言，RecPilot 在新颖性维度上建立了显著优势。这种优越性可以正式归因于我们提出的多方面兴趣分解策略，该策略将用户兴趣显式分解为多个细粒度的属性子集。通过将一般偏好与小众兴趣和新兴兴趣区分开来，这一方案使智能体能够捕捉往往被通用任务规划基线方法所忽视的细粒度用户意图。因此，RecPilot 能够生成更加多样且出人意料的推荐结果，同时保持严格的事实相关性。

模拟用户与真实用户之间的科恩卡帕系数为 0.7064，表明一致性良好。为了在有限预算内开展大规模评估，除非另有说明，我们在后续部分使用模拟用户对每个指标进行评分。

成对比较。 为进一步验证我们方法的优越性，我们将其与最强的基于智能体的基线方法——计划与求解——进行正面对比。对于每个测试案例，我们使用两种方法生成报告，并指导模拟用户选出“更好”的报告，随后计算平均胜率。对比结果如图 4a 所示。如水平条形图所示，我们的方法在全部六项评估指标上均持续取得更高的胜率。值得注意的是，在新颖性方面高达 77% 的胜率，为我们此前关于多方面兴趣分解有效性的观察提供了有力的量化

表4：所提出的天猫数据集报告生成智能体的消融研究。

方法	准确性	覆盖度	信息丰富度	清晰度	一致性	新颖性	Avg.
RecPilot	**4.60**	**3.62**	**3.80**	**4.80**	**3.94**	**4.09**	**4.14**
不含评价标准	4.53	3.57	3.77	4.76	3.90	4.08	4.10
不含经验	4.49	3.55	3.76	4.74	3.87	4.06	4.08
不含兴趣分解	4.46	3.56	3.79	4.79	3.87	4.05	4.09

依据。除新颖性外，我们的方法在清晰度方面也取得了显著优势，胜率达到 66%。这一优势源于我们的方面感知并行排序设计：RecPilot 不会生成一份庞杂的摘要，而是呈现不同方面的专属排序列表，让商品比较一目了然，帮助用户直接做出决策。此外，在准确性（60%）和覆盖度（57%）方面取得的稳健胜利，验证了我们的标准-经验双通道机制能够以可量化属性和隐含的历史上下文为生成内容提供依据，相比通用推理路径，可提供更精准、个性化的决策支持。

消融研究。 为评估各核心组件对我们报告生成智能体的贡献，我们通过比较完整的模型设计与若干变体开展了消融研究。结果见表4。

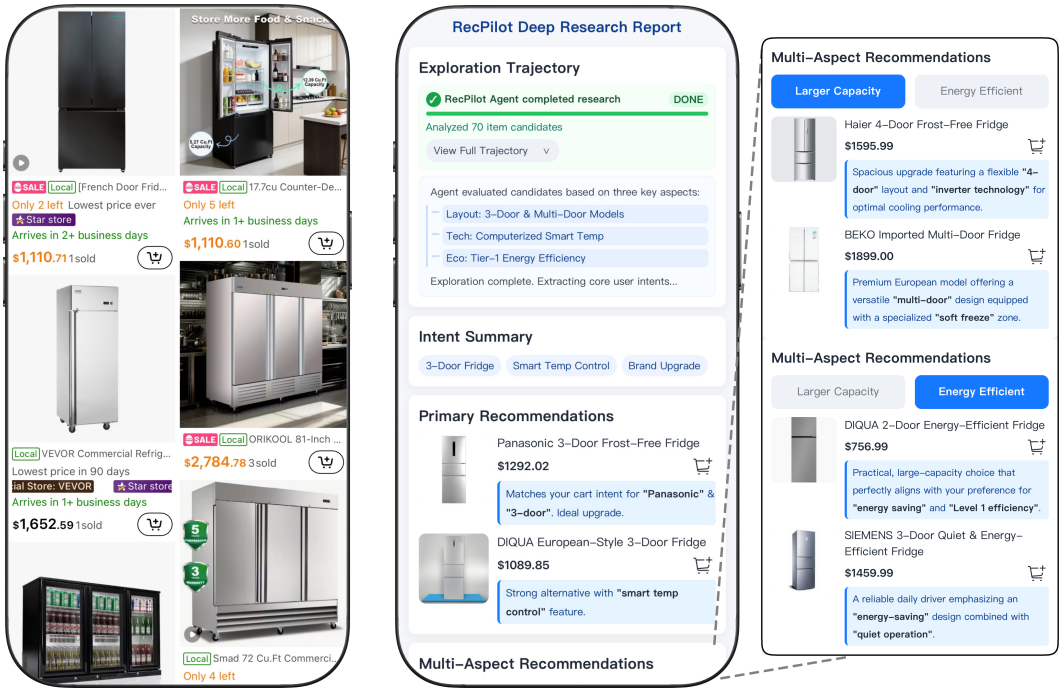
• **无评价标准：** 该变体移除了结构化评价标准，仅利用经验进行排序。所有指标均持续下降，这表明缺乏统一属性维度会损害生成内容的逻辑结构。结构化评价标准提供了可量化的比较依据，从而显著提升了报告的组织性和一致性。

• **无经验：** 该变体排除了文本经验记忆，仅利用评价标准进行排序。性能下降表明，仅依赖显式属性分数不足以捕捉依赖上下文的隐式用户偏好。经验模块从历史交互中提取语义丰富的细微信息，作为结构化评分的灵活补充，这对于提升报告的整体信息丰富度和鲁棒性至关重要。

• **无兴趣分解：** 该变体移除了多方面兴趣分解策略，不再依据评价标准子集将用户兴趣解耦为多个细粒度方面。准确性显著下降，表明将宏观层面的用户偏好分解为具体的兴趣片段，对于捕捉复杂的用户意图至关重要。

自我进化的效果。在这一部分，我们评估报告生成智能体的自我进化能力。该智能体能够持续优化评价标准并积累经验。具体而言，我们利用不同时间跨度（即2、5、8、11、14天）的交互记录进行优化，然后比较其在测试集上的表现。结果如图4b所示。可以看到，随着时间跨度增加，我们的智能体表现持续提升，直接证明了所提出的自我演化机制行之有效。得益于评价标准的设计，自我进化模块能够利用用户行为数据持续优化其权重。数据越多，这些权重值就越稳定，从而能够更精准地进行长期偏好建模。尤其是在早期阶段（即从2天到8天），智能体在几轮内便实现了快速的性能提升。这得益于经验记忆模块，它能够从有限的反馈中提取有意义的文本偏好信号，从而帮助智能体快速适应类似场景。

案例研究。为了直观地展示我们提出的 RecPilot 如何将推荐体验从被动的商品展示转变为主动的综合报告，我们在图5中给出了一个具体示例。在这一场景中，用户打算购买一台冰箱。如图5a所示，现有推荐系统通常以商品列表的形式展示可能引起用户兴趣的冰箱。这类列表仅包含有限的信息，包括商品图片、标题和价格。因此，用户需要逐一点击每个项目以查看详细规格，这无疑是一个费时费力的过程，也会对用户体验产生负面影响。



（a）传统商品列表。　　（b）由我们的 RecPilot 生成的报告。

图5：现有推荐系统与我们提出的 RecPilot 中的示例交互界面。

图5b展示了 RecPilot 针对同一场景生成的报告。与传统商品列表不同，智能体会自主地将其发现综合整理成连贯且可解释的报告。首先，报告提供“探索轨迹”部分，展示物品探索过程，旨在增强报告的可信度。接下来，“意图摘要”部分总结用户可能感兴趣的核心属性（例如，三门冰箱）。随后，智能体通过多方面的兴趣分解，提供量身定制的推荐结果。报告提供“主要推荐”以便用户快速决策，同时提供“多方面推荐”，明确对比不同的决策维度——例如，升级到更大容量，还是优先考虑能源效率。这种结构化呈现显著减轻了信息过载，成功地将推荐从由用户主导的被动筛选转变为由智能体驱动的主动服务。

4 相关工作

4.1 基于大语言模型的推荐

大语言模型（LLM）凭借先进的语义理解和推理能力，已广泛应用于推荐系统 [6, 34, 35]。早期研究主要采用零样本或少样本范式，将序列推荐等任务 [36, 37], 、用户数据增强 [38, 39], 和对话式推荐 [40]改写为文本提示。然而，通用语言建模与特定领域推荐任务之间的差异，往往难以通过提示工程弥合。因此，后续研究着重探索如何更好地使大语言模型适应推荐场景，包括使用推荐数据进行指令微调以提升性能 [41, 42, 43], 、引入偏好优化以更好地与用户偏好保持一致 [44, 45], ，以及采用强化学习来提升模型在推荐情境中的深度推理能力 [46, 47]。此外，一些研究还尝试开发基于大语言模型的推荐智能体 [48, 49], ，运用多智能体框架 [50, 51], 、记忆建模 [52, 53] 和强化学习 [54, 55] 等技术。

4.2 深度研究

近年来，大语言模型已发展成为强大的问题求解器，但复杂的开放式任务往往超出单个提示或标准检索增强生成（RAG）的能力范围 [56]。这一局限促使深度研究（DR）范式应运而生：它通过端到端工作流 [57, 10]，将大语言模型构建为自主研究智能体。典型的深度研究系统包含四个核心组件：查询规划通过并行、序列或树状策略，将复杂查询拆解为可执行的子任务 [58, 59, 60]。自适应信息获取包括调用检索工具、自适应搜索时间规划和信息过滤 [61, 62]。动态记忆管理通过整合、索引、更新和遗忘机制来管理记忆生命周期 [63, 64, 65]。答案生成则综合上游信息、解决相互冲突的证据，并进行逻辑推理 [30]。近期研究还通过智能体提示 [66, 10], 、监督微调 [67, 68], 和端到端强化学习 [60, 69] 等技术进一步优化深度研究。

5 结论

在这项工作中，我们探讨了推荐系统设计中一个长期存在的局限：系统最终向用户返回一系列项目供其评估，这给用户带来了不必要的认知负担，迫使他们自行充当信息综合者。为解决这一问题，我们提出了交互方式上的根本转变——从被动的商品展示转向主动的综合报告生成。我们通过 RecPilot 实现了这一构想：这是一个多智能体框架，能够自主探索物品空间，并生成可解释、以用户为中心的报告，直接为决策提供支持。实验验证了这种方法不仅能够有效建模用户偏好，还能生成连贯且有说服力的内容，显著减少用户的付出。

在考虑我们的方法未来的部署时，我们认识到有两个关键方面值得谨慎关注。首先，并非所有产品类别都同样适合这种新的交互范式；我们预计，对于高价商品，这种方式最有价值，因为用户通常会投入大量时间做出最终决策。其次，一些用户可能更偏好快速交互，并且可能认为，与传统的基于列表的推荐相比，报告生成过程更耗时。鉴于这些因素，我们认为，整合快速传统推荐与较慢的基于推理的报告生成的双模式系统将具有显著价值。在这种设计中，用户可以在需要更深入的决策支持时选择启用更审慎的模式。

展望未来，这项工作开辟了若干颇具前景的方向。新兴的深度研究范式指向一种潜在的未来：推荐系统将从简单的过滤器演进为智能助手——能够对用户需求进行推理，并论证其推荐结果。关键挑战依然存在，包括提升大规模商品池的轨迹模拟效率、确保生成的报告的事实一致性，以及开发超越传统准确性度量的稳健评估指标，以刻画用户认知负荷与满意度。我们相信，这项工作为新一代推荐系统奠定了基础，使其不仅更加有效，也更加有帮助——最终将用户从信息寻求的负担中解放出来。

参考文献

[1] Yehuda Koren、Steffen Rendle和Robert Bell。协同过滤的进展。推荐系统手册，第91–142页，2021年。[2] Kun Zhou、Hui Yu、Wayne Xin Zhao和Ji-Rong Wen。滤波增强型多层感知机足以实现序列推荐。载于*2022*年 *ACM Web* 会议论文集，第2388–2399页，2022年。[3] Balázs Hidasi、Alexandros Karatzoglou、Linas Baltrunas和Domonkos Tikk。使用循环神经网络的基于会话的推荐。载于Yoshua Bengio和Yann LeCun主编，第四届国际学习表征会议，*ICLR 2016*，波多黎各圣胡安，*2016*年*5*月*2*日至*4*日，会议论文集，2016年。

[4] Puneet Kumar Sharma、Chetan Bhardwaj、Dolly Sharma和Rishabh Mishra。基于卷积神经网络的推荐系统综合综述。载于*2023*年智能设备国际会议（*ICSD*），第1–6页。电气电子工程师学会，2024年。[5] Wang-Cheng Kang和Julian McAuley。自注意力序列推荐。载于*2018*年*IEEE*国际数据挖掘会议（*ICDM*），第197–206页。电气电子工程师学会，2018年。[6] Wayne Xin Zhao、Kun Zhou、Junyi Li、Tianyi Tang、Xiaolei Wang、Yupeng Hou、Yingqian Min、Beichen Zhang、Junjie Zhang、Zican Dong、Yifan Du、Chen Yang、Yushuo Chen、Zhipeng Chen、Jinhao Jiang、Ruiyang Ren、Yifan Li、Xinyu Tang、Zikang Liu、Peiyu Liu、Jian-Yun Nie和Ji-Rong Wen。大语言模型综述。*CoRR*，abs/2303.18223，2023年。[7] Shashank Rajput、Nikhil Mehta、Anima Singh、Raghunandan Hulikal Keshavan、Trung Vu、Lukasz Heldt、Lichan Hong、Yi Tay、Vinh Q. Tran、Jonah Samost、Maciej Kula、Ed H. Chi和Mahesh Sathiamoorthy。基于生成式检索的推荐系统。载于神经信息处理系统进展第*36*卷：*2023*年神经信息处理系统年度会议，神经信息处理系统大会*2023*，美国路易斯安那州新奥尔良，*2023*年*12*月*10*日至*16*日，2023年。[8] Likang Wu、Zhi Zheng、Zhaopeng Qiu、Hao Wang、Hongchao Gu、Tingjia Shen、Chuan Qin、Chen Zhu、Hengshu Zhu、Qi Liu等。用于推荐的大语言模型综述。万维网，27(5):60，2024年。[9] Xi Zhu、Yu Wang、Hang Gao、Wujiang Xu、Chen Wang、Zhiwei Liu、Kun Wang、Mingyu Jin、Linsey Pang、Qingsong Weng等。推荐系统与大语言模型智能体的交汇：综述。隐私与安全基础与趋势®，7(4):247–396，2025年。[10] OpenAI。Deep Research系统卡。技术报告，OpenAI，2025年2月。发布于2025年2月25日。[11] Zihan Liu、Yupeng Hou和Julian J. McAuley。多行为生成式推荐。载于第*33*届*ACM*信息与知识管理国际会议论文集，信息与知识管理国际会议*2024*，美国爱达荷州博伊西，*2024*年*10*月*21*日至*25*日，第1575–1585页。美国计算机协会，2024年。[12] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li和Peter J. Liu。利用统一的文本到文本转换器探索迁移学习的极限。机器学习研究期刊，21:140:1–140:67，2020年。[13] Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Mingchuan Zhang、Y. K. Li、Y. Wu和Daya Guo。DeepSeekMath：推动开放语言模型中数学推理的极限。*CoRR*，abs/2402.03300，2024年。[14] DeepSeek-AI。DeepSeek-R1：通过强化学习激励大语言模型的推理能力。*CoRR*，abs/2501.12948，2025年。[15] Kimi Team。Kimi K1.5：利用大语言模型扩展强化学习。*CoRR*，abs/2501.12599，2025年。[16] Omar Khattab和Matei Zaharia。ColBERT：通过基于BERT的上下文化后期交互实现高效且有效的段落检索。载于第*43*届国际*ACM SIGIR*信息检索研究与发展会议论文集，第39–48页，2020年。[17] Kalervo Järvelin和Jaana Kekäläinen。信息检索技术的累积增益评估。*ACM*信息系统汇刊（*TOIS*），20(4):422–446，2002年。[18] Zhefan Wang、Guokai Yan、Jinbei Yu、Siyu Gu、Jingyan Chen、Peng Jiang、Zhiqiang Guo和Min Zhang。通过分层行为建模实现生成式序列推荐。*CoRR*，abs/2511.03155，2025年。[19] Wang-Cheng Kang和Julian J. McAuley。自注意力序列推荐。载于电气电子工程师学会国际数据挖掘会议，*ICDM 2018*，新加坡，*2018*年*11*月*17*日至*20*日，第197–206页。电气电子工程师学会计算机学会，2018年。[20] Fei Sun、Jun Liu、Jian Wu、Changhua Pei、Xiao Lin、Wenwu Ou和Peng Jiang。BERT4Rec：使用来自Transformer的双向编码器表示的序列推荐。载于第*28*届*ACM*信息与知识管理国际会议论文集，信息与知识管理国际会议*2019*，中国北京，*2019*年*11*月*3*日至*7*日，第1441–1450页。美国计算机协会，2019年。[21] Balázs Hidasi、Alexandros Karatzoglou、Linas Baltrunas和Domonkos Tikk。使用循环神经网络的基于会话的推荐。载于Yoshua Bengio和Yann LeCun主编，

第*4*届国际学习表征会议，*ICLR 2016*，波多黎各圣胡安，*2016*年*5*月*2*日至*4*日，会议论文集，2016年。[22] Jiajie Su、Chaochao Chen、Zibin Lin、Xi Li、Weiming Liu 和 Xiaolin Zheng。用于多行为序列推荐的个性化行为感知Transformer。载于第*31*届美国计算机协会国际多媒体会议论文集，第6321–6331页，2023年。[23] Yuhao Yang、Chao Huang、Lianghao Xia、Yuxuan Liang、Yanwei Yu 和 Chenliang Li。用于序列推荐的多行为超图增强Transformer。载于第*28*届美国计算机协会*SIGKDD*知识发现与数据挖掘会议论文集，第2263–2274页，2022年。[24] Enming Yuan、Wei Guo、Zhicheng He、Huifeng Guo、Chengkai Liu 和 Ruiming Tang。多行为序列Transformer推荐器。载于第*45*届国际美国计算机协会*SIGIR*信息检索研究与发展会议论文集，第1642–1652页，2022年。[25] Jiakai Tang、Sunhao Dai、Teng Shi、Jun Xu、Xu Chen、Wen Chen、Jian Wu 和 Yuning Jiang。推荐前先思考：释放序列推荐的潜在推理能力。*arXiv*预印本 *arXiv:2503.22675*，2025年。[26] Aixin Liu、Aoxue Mei、Bangcai Lin、Bing Xue、Bingxuan Wang、Bingzheng Xu、Bochao Wu、Bowei Zhang、Chaofan Lin、Chen Dong 等。Deepseek-v3. 2：推动开放式大语言模型的前沿。*arXiv*预印本 *arXiv:2512.02556*，2025年。[27] Aohan Zeng、Xin Lv、Qinkai Zheng、Zhenyu Hou、Bin Chen、Chengxing Xie、Cunxiang Wang、Da Yin、Hao Zeng、Jiajie Zhang 等。Glm-4.5：智能体、推理与编程（ARC）基础模型。*arXiv*预印本 *arXiv:2508.06471*，2025年。[28] An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv 等。Qwen3技术报告。*arXiv*预印本 *arXiv:2505.09388*，2025年。[29] Aaditya Singh、Adam Fry、Adam Perelman、Adam Tart、Adi Ganesh、Ahmed El-Kishky、Aidan McLaughlin、Aiden Low、AJ Ostrow、Akhila Ananthram 等。OpenAI GPT-5系统卡。*arXiv*预印本 *arXiv:2601.03267*，2025年。[30] Shunyu Yao、Jeffrey Zhao、Dian Yu、Nan Du、Izhak Shafran、Karthik R. Narasimhan 和 Yuan Cao。ReAct：协同语言模型中的推理与行动。载于第十一届国际学习表征会议，*ICLR 2023*，卢旺达基加利，*2023*年*5*月*1*日至*5*日。OpenReview.net，2023年。[31] Lei Wang、Wanyu Xu、Yihuai Lan、Zhiqiang Hu、Yunshi Lan、Roy Ka-Wei Lee 和 Ee-Peng Lim。计划并求解提示：通过大语言模型改进零样本思维链推理。载于第*61*届计算语言学协会年会论文集（第*1*卷：长篇论文），第2609–2634页。计算语言学协会，2023年。[32] Gemini团队、Rohan Anil、Sebastian Borgeaud、Jean-Baptiste Alayrac、Jiahui Yu、Radu Soricut、Johan Schalkwyk、Andrew M Dai、Anja Hauth、Katie Millican 等。Gemini：一系列高能力多模态模型。*arXiv*预印本 *arXiv:2312.11805*，2023年。[33] Yanzhao Zhang、Mingxin Li、Dingkun Long、Xin Zhang、Huan Lin、Baosong Yang、Pengjun Xie、An Yang、Dayiheng Liu、Junyang Lin 等。Qwen3嵌入：通过基础模型推进文本嵌入与重排序。*arXiv*预印本 *arXiv:2506.05176*，2025年。[34] Likang Wu、Zhi Zheng、Zhaopeng Qiu、Hao Wang、Hongchao Gu、Tingjia Shen、Chuan Qin、Chen Zhu、Hengshu Zhu、Qi Liu、Hui Xiong 和 Enhong Chen。用于推荐的大语言模型综述。*万维网（WWW）*，27(5):60，2024年。[35] Keqin Bao、Jizhi Zhang、Yang Zhang、Wenjie Wang、Fuli Feng 和 Xiangnan He。TallRec：一种用于使大语言模型与推荐对齐的有效高效微调框架。载于 Jie Zhang、Li Chen、Shlomo Berkovsky、Min Zhang、Tommaso Di Noia、Justin Basilico、Luiz Pizzato 和 Yang Song（编），第*17*届美国计算机协会推荐系统会议，*RecSys 2023*，新加坡新加坡，*2023*年*9*月*18*日至*22*日论文集，第1007–1014页。美国计算机协会，2023年。[36] Yupeng Hou、Junjie Zhang、Zihan Lin、Hongyu Lu、Ruobing Xie、Julian J. McAuley 和 Wayne Xin Zhao。大语言模型是推荐系统的零样本排序器。载于

信息检索进展——第*46*届欧洲信息检索会议，*ECIR 2024*，英国格拉斯哥，*2024*年*3*月*24*日至*28*日，论文集，第二部分，计算机科学讲义第14609卷，第364–381页。Springer，2024年。[37] Sunhao Dai、Ninglu Shao、Haiyuan Zhao、Weijie Yu、Zihua Si、Chen Xu、Zhongxiang Sun、Xiao Zhang 和 Jun Xu。揭示ChatGPT在推荐系统中的能力。载于第*17*届美国计算机协会推荐系统会议，*RecSys 2023*，新加坡新加坡，*2023*年*9*月*18*日至*22*日论文集，第1126–1132页。美国计算机协会，2023年。[38] Yunjia Xi、Weiwen Liu、Jianghao Lin、Xiaoling Cai、Hong Zhu、Jieming Zhu、Bo Chen、Ruiming Tang、Weinan Zhang 和 Yong Yu。通过大语言模型的知识增强迈向开放世界推荐。载于第*18*届美国计算机协会推荐系统会议，*RecSys 2024*，意大利巴里，*2024*年*10*月*14*日至*18*日论文集，第12–22页。美国计算机协会，2024年。[39] Sichun Luo、Yuxuan Yao、Bowei He、Yinya Huang、Aojun Zhou、Xinyi Zhang、Yuanzhang Xiao、Mingjie Zhan 和 Linqi Song。通过相互增强与自适应聚合将大语言模型融入推荐。*CoRR*，abs/2401.13870，2024年。[40] Yunfan Gao、Tao Sheng、Youlin Xiang、Yun Xiong、Haofen Wang 和 Jiawei 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。[41] Shijie Geng、Shuchang Liu、Zuohui Fu、Yingqiang Ge 和 Yongfeng Zhang。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*RecSys '22*：第*16*届美国计算机协会推荐系统会议，美国华盛顿州西雅图，*2022*年*9*月*18*日至*23*日，第299–315页。美国计算机协会，2022年。[42] Bowen Zheng、Yupeng Hou、Hongyu Lu、Yu Chen、Wayne Xin Zhao、Ming Chen 和 Ji-Rong Wen。通过融合协同语义使大语言模型适应推荐任务。载于第*40*届电气电子工程师学会国际数据工程会议，*ICDE 2024*，荷兰乌得勒支，*2024*年*5*月*13*日至*16*日，第1435–1448页。电气电子工程师学会，2024年。[43] Junjie Zhang、Ruobing Xie、Yupeng Hou、Wayne Xin Zhao、Leyu Lin 和 Ji-Rong Wen。将推荐作为指令遵循：一种由大语言模型赋能的推荐方法。*CoRR*，abs/2305.07001，2023年。[44] Yuxin Chen、Junfei Tan、An Zhang、Zhengyi Yang、Leheng Sheng、Enzhi Zhang、Xiang Wang 和 Tat-Seng Chua。关于用于推荐的softmax直接偏好优化。载于神经信息处理系统进展*38*：神经信息处理系统年会*2024*，*NeurIPS 2024*，加拿大不列颠哥伦比亚省温哥华，*2024*年*12*月*10*日至*15*日，2024年。[45] Chongming Gao、Ruijun Chen、Shuai Yuan、Kexin Huang、Yuanqing Yu 和 Xiangnan He。Sprec：利用自博弈，消除基于大语言模型的推荐中的偏好对齐偏差。*CoRR*，abs/2412.09243，2024年。[46] Bowen Zheng、Xiaolei Wang、Enze Liu、Xi Wang、Hongyu Lu、Yu Chen、Wayne Xin Zhao 和 Ji-Rong Wen。用于基于大语言模型的推荐的自主推理检索。载于第*34*届美国计算机协会国际信息与知识管理会议，*CIKM 2025*，大韩民国首尔，*2025*年*11*月*10*日至*14*日论文集，第4423–4433页。美国计算机协会，2025年。[47] Yang Zhang、Wenxin Xu、Xiaoyan Zhao、Wenjie Wang、Fuli Feng、Xiangnan He 和 Tat-Seng Chua。面向基于大语言模型的推荐的强化潜在推理。*CoRR*，abs/2505.19092，2025年。[48] Junjie Zhang、Yupeng Hou、Ruobing Xie、Wenqi Sun、Julian J. McAuley、Wayne Xin Zhao、Leyu Lin 和 Ji-Rong Wen。AgentCF：利用自主语言智能体进行推荐系统协同学习。载于美国计算机协会*2024*年万维网会议论文集，*WWW 2024*，新加坡，*2024*年*5*月*13*日至*17*日，第3679–3689页。美国计算机协会，2024年。[49] Xu Huang、Jianxun Lian、Yuxuan Lei、Jing Yao、Defu Lian 和 Xing Xie。推荐器人工智能智能体：集成大语言模型以实现交互式推荐。美国计算机协会信息系统汇刊，43(4):96:1–96:33，2025年。[50] Wujiang Xu、Yunxiao Shi、Zujie Liang、Xuying Ning、Kai Mei、Kun Wang、Xi Zhu、Min Xu 和 Yongfeng Zhang。InstructAgent：通过大语言模型智能体构建用户可控推荐器。*CoRR*，abs/2502.14662，2025年。

[51] Yu Xia、Sungchul Kim、Tong Yu、Ryan A. Rossi 和 Julian J. McAuley。多智能体协同过滤：协调用户与项目，实现智能体推荐。*CoRR*，abs/2511.18413，2025年。[52] Jiahao Liu、Shengkang Gu、Dongsheng Li、Guangping Zhang、Mingzhe Han、Hansu Gu、Peng Zhang、Tun Lu、Li Shang 和 Ning Gu。AgentCF：用于流行度感知跨域推荐的记忆增强型大语言模型智能体。载于第*48*届国际*ACM SIGIR*信息检索研究与发展会议（*SIGIR 2025*），*2025*年*7*月*13*日至*18*日，意大利帕多瓦，第2566–2571页。美国计算机协会。[53] Jiani Huang、Xingchen Zou、Lianghao Xia 和 Qing Li。Mr.Rec：结合记忆与推理，打造基于大语言模型的个性化推荐助手。*CoRR*，abs/2510.14629，2025年。[54] Shihao Cai、Chongming Gao、Haoyan Liu、Wentao Shi、Jianshan Sun、Ruiming Tang 和 Fuli Feng。MGFRec：探索结合多重依据与反馈的强化推理推荐。*CoRR*，abs/2510.22888，2025年。[55] Chenghao Wu、Ruiyang Ren、Junjie Zhang、Ruirui Wang、Zhongrui Ma、Qi Ye 和 Wayne Xin Zhao。StaRec：通过自主审慎推理构建高效的推荐系统智能体框架。载于第*34*届*ACM*国际信息与知识管理会议（*CIKM 2025*），*2025*年*11*月*10*日至*14*日，大韩民国首尔，第3355–3365页。美国计算机协会，2025年。[56] Agada Joseph Oche、Ademola Glory Folashade、Tirthankar Ghosal 和 Arpan Biswas。检索增强生成（RAG）关键系统的系统综述：进展、缺口与未来方向。*CoRR*，abs/2507.18910，2025年。[57] Zhengliang Shi、Yiqun Chen、Haitao Li、Weiwei Sun、Shiyu Ni、Yougang Lyu、Run-Ze Fan、Bowen Jin、Yixuan Weng、Minjun Zhu、Qiujie Xie、Xinyu Guo、Qu Yang、Jiayi Wu、Jujia Zhao、Xiaqiang Tang、Xinbei Ma、Cunxiang Wang、Jiaxin Mao、Qingyao Ai、Jen-tse Huang、Wenxuan Wang、Yue Zhang、Yiming Yang、Zhaopeng Tu 和 Zhaochun Ren。深度研究：系统综述。*CoRR*，abs/2512.02038，2025年。[58] Denny Zhou、Nathanael Schärli、Le Hou、Jason Wei、Nathan Scales、Xuezhi Wang、Dale Schuurmans、Claire Cui、Olivier Bousquet、Quoc V. Le 和 Ed H. Chi。由少到多提示使大语言模型能够进行复杂推理。载于第十一届国际学习表征会议（*ICLR 2023*），*2023*年*5*月*1*日至*5*日，卢旺达基加利。OpenReview.net，2023年。[59] Bowen Jin、Hansi Zeng、Zhenrui Yue、Dong Wang、Hamed Zamani 和 Jiawei Han。Search-r1：通过强化学习训练大语言模型进行推理并利用搜索引擎。*CoRR*，abs/2503.09516，2025年。[60] Huatong Song、Jinhao Jiang、Yingqian Min、Jie Chen、Zhipeng Chen、Wayne Xin Zhao、Lei Fang 和 Ji-Rong Wen。R1-Searcher：通过强化学习激发大语言模型的搜索能力。*CoRR*，abs/2503.05592，2025年。[61] Xiaoxi Li、Guanting Dong、Jiajie Jin、Yuyao Zhang、Yujia Zhou、Yutao Zhu、Peitian Zhang 和 Zhicheng Dou。Search-o1：增强了智能体搜索能力的大型推理模型。载于*2025*年自然语言处理实证方法会议论文集（*EMNLP 2025*），*2025*年*11*月*4*日至*9*日，中国苏州，第5420–5438页。计算语言学协会，2025年。[62] Zhengliang Shi、Lingyong Yan、Dawei Yin、Suzan Verberne、Maarten de Rijke 和 Zhaochun Ren。迭代式自激励赋予大语言模型智能体搜索器能力。*CoRR*，abs/2505.20128，2025年。[63] Wanjun Zhong、Lianghong Guo、Qiqi Gao、He Ye 和 Yanlin Wang。MemoryBank：通过长期记忆增强大语言模型。载于第三十八届*AAAI*人工智能会议（*AAAI 2024*）、第三十六届创新人工智能应用会议（*IAAI 2024*）、第十四届教育人工智能进展研讨会（*EAAI 2014*），*2024*年*2*月*20*日至*27*日，加拿大温哥华，第19724–19731页。AAAI出版社，2024年。[64] Wujiang Xu、Zujie Liang、Kai Mei、Hang Gao、Juntao Tan 和 Yongfeng Zhang。A-MEM：大语言模型智能体的智能体记忆。*CoRR*，abs/2502.12110，2025年。[65] Prateek Chhikara、Dev Khant、Saket Aryan、Taranjeet Singh 和 Deshraj Yadav。Mem0：借助可扩展长期记忆构建生产就绪型人工智能智能体。*CoRR*，abs/2504.19413，2025年。

[66] 谷歌有限责任公司。Gemini 深度研究。网页，2024年。在线资源；访问日期：2026-02-08。[67] Xiaoxi Li、Jiajie Jin、Guanting Dong、Hongjin Qian、Yutao Zhu、Yongkang Wu、Ji-Rong Wen 和 Zhicheng Dou。Webthinker：赋予大型推理模型深度研究能力。*CoRR*，abs/2504.21776，2025年。[68] Dingchu Zhang、Yida Zhao、Jialong Wu、Liwen Zhang、Baixuan Li、Wenbiao Yin、Yong Jiang、Yu-Feng Li、Kewei Tu、Pengjun Xie 和 Fei Huang。Evolvesearch：一种迭代式自我进化搜索智能体。载于*2025*年自然语言处理实证方法会议论文集（*EMNLP 2025*），*2025*年*11*月*4*日至*9*日，中国苏州，第13123–13136页。计算语言学协会，2025年。[69] Yuxiang Zheng、Dayuan Fu、Xiangkun Hu、Xiaojie Cai、Lyumanshan Ye、Pengrui Lu 和 Pengfei Liu。Deepresearcher：通过在真实世界环境中进行强化学习扩展深度研究。载于*2025*年自然语言处理实证方法会议论文集（*EMNLP 2025*），*2025*年*11*月*4*日至*9*日，中国苏州，第414–431页。计算语言学协会，2025年。

A 附录

A.1 人工评估详情

为减轻人工评估固有的主观性并确保结果可复现，我们制定了严格的人工评估协议。评估者必须依据明确的行为锚点评估生成的报告，而非凭借整体印象。详细指南如下。

A.1.1 任务设置

我们设计了一项盲测评估任务，用于评估生成的报告能否有效辅助决策并减轻探索负担。对于每个测试案例，人工评估者都会获得历史交互、模拟探索轨迹以及候选项目集，作为评估背景。不同基线模型和所提出方法生成的报告均经过严格匿名处理，并以随机顺序呈现。随后，评估者必须依据提供的证据和真实购买项目这一准确性的绝对锚点，针对六个预先设定的维度，分别对每份报告独立评分。

A.1.2 评估者说明

评估器被要求严格遵循以下核心规则，以尽量减少主观偏差：

- 基于证据评分：评估器必须仅依据提供的历史轨迹和候选集作为证据。报告中提及的任何具体项目属性或行为细节，若无法追溯至输入内容，都必须作为幻觉内容予以扣分。

- 真实情况锚定："最佳选择"严格定义为实际购买的项目。推荐候选集之外的项目，或过早提及已移除的真实购买操作，都必须受到严厉扣分。

- 对模拟语气的容忍度：由于目标会话是通过轨迹模拟生成的，不应因报告使用推测性或模拟性的语气而扣分。不过，若将模拟轨迹断言为过去发生的事实，则应酌情扣分。

- 形式与功能：评估必须完全聚焦于报告是否真正有助于决策。评估器不得因特定术语、文体模板或单纯的语法流畅度而受影响，尤其是在底层逻辑存在缺陷时。

A.1.3 评分标准

为确保较高的标注者间一致性，我们针对每项评估指标制定了细致的五点评分量表，并明确列出相应的行为锚点。评估者必须根据表5中列出的具体描述，对报告的质量进行匹配。

表5：用于人工评估并设有明确行为锚点的五点评分量表 s.

指标	1分（差）	3分（可接受）	5分（优秀）
准确性	完全没有命中该标签或高度相似的项目，既不在主要推荐中，也不在次要推荐中tions .	主要建议未命中该标签，但该标签或高度相似的项目包含在次要建议中·	主要建议准确命中标签或完全对齐的替代。
覆盖度	缺少关键决策因素，或完全依赖于捏造的特征，且没有证据。	提及了一些相关线索，但缺乏重点，或者证据链较薄弱。	明确突出2-3决定性线索并坚定地将其绑定到候选对象比较。
信息丰富度	充斥着无意义的填充词、通用模板或严重幻觉。	包含有用信息，但其中夹杂通用冗余和缺乏深刻洞见。	高信息密度；每个主张 is 都严格基于文中提供证据。
清晰度	含义高度模糊，自相矛盾，或未能得出明确的结论，供用户参考。	给出了结论，但对比或替代选择的条件含糊不清。	主要建议明确，且条件用于选择替代方案是不同的。
一致性	严重的逻辑断裂，或直接矛盾于历史轨迹证-用户的.....证据。	总体上与输入一致，但存在轻微的——小逻辑跳跃，或略显强制性的推断推断·	完美保真度；所有推论和因果联系都严格遵循轨迹证据。
新颖性	缺乏 any 前瞻-展望见解，或依据“见解”完全基于幻觉细节 .	提供了一些见解或警告，但它们过于笼统，或对决策并不关键 .	主动指出高价值、有证据支持的风险或机遇机遇·