AlignUSER：通过世界模型实现与人类对齐的大语言模型智能体，用于推荐系统评估

Nicolas Bougie[1], **Gian Marconi Marconi**[1],**Tony Yip**[1],**Narimasa Watanabe**[1]
{nicolas.bougie,gianmaria.marconi,tony.yip,narimasa.watanabe}@woven.toyota

[1]Woven by Toyota

arXiv:2601.00930v1 [cs.IR] 2 Jan 2026

摘要

由于离线指标与真实用户行为之间存在差距，加之交互数据稀缺，评估推荐系统仍然具有挑战性。近期研究探索将大语言模型（LLM）智能体作为合成用户，但这类方法通常依赖少样本提示，导致其对环境的理解较为浅显，也限制了其忠实复现用户行为的能力。我们提出 ALIGNUSER，这是一种从人类交互中学习世界模型驱动智能体的框架。给定行动与状态的轨迹序列，我们将世界建模形式化为下一状态预测任务，帮助智能体内化环境。为了使行动与人类角色画像保持对齐，我们围绕示范生成反事实轨迹，并提示大语言模型将自身决策与人类选择进行比较、识别次优行动并提取经验。随后，利用学习得到的策略驱动智能体与推荐系统进行交互。我们在多个数据集上评估 ALIGNUSER，结果表明，无论在微观层面还是宏观层面，它都比以往方法更接近真实人类。

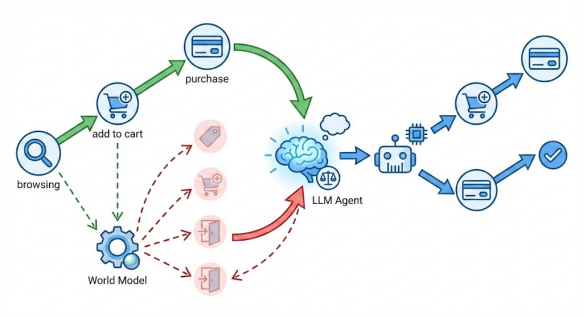

图 1：ALIGNUSER 框架通过隐式建模世界模型并探索替代情境来评估推荐系统。

1 引言

推荐系统是许多在线服务的核心，从电子商务到媒体平台皆是如此；它们能够实现内容个性化并提升用户参与度（Li 等人，2024年）。尽管用户偏好建模已取得显著进展，评估仍然是瓶颈（Yoon 等人，2024年）。当前的评估实践主要采用基于静态数据集计算的离线指标（例如，归一化折损累计增益、召回率），但模型部署后，这些指标往往与在线行为不一致（Zhang 等人，2019年；Jannach 和 Jugovac，2019年）。此外，这些指标无法反映销售额或满意度等商业价值（Jannach 和 Jugovac，2019年）。另一方面，在线 A/B 测试能够提供更真实的反馈，但成本高昂、迭代缓慢，且受到限制于伦理与隐私考量。一种颇具前景的替代方案是利用基于大语言模型的智能体作为合成用户，在模拟环境中与推荐系统互动（Bougie 和 Watanabe，2025c）。这些智能体能够以自然语言表达丰富的偏好并提供反馈，有望近似反映满意度或感知相关性等用户级指标（Hou 等人，2024年；Zhang 等人，2023年；Huang 等人，2023年；Wang 等人，2023c；Yoon 等人，2024年）。然而，现有的大多数方法都依赖少样本提示来模仿人类行为。通常，系统会根据少量示例要求智能体“像典型用户一样行动”（Wang 等，2025；Bougie 和 Watanabe，2025c），但智能体并未明确理解环境如何随自身行动而变化。因此，智能体对世界的理解仅停留在表面，难以忠实复现人类行为轨迹，尤其是在推理长期后果时（例如，何时将商品加入购物车或退出）。此外，若没有明确的对齐机制，智能体主要表现出的会是自身的内在偏见，而非由角色设定真实地引导其决策（Salecha 等人，2024年；Kaiser 等人，2025年；Bisbee 等人，2024年）。

在本文中，我们提出，智能体应当

理解世界的运作方式，以便忠实地复现人类动作。鉴于此，给定展开轨迹（例如浏览、搜索、将物品加入购物车等），我们首先在一项世界模型 任务上预训练智能体策略，根据状态–动作对预测下一个状态。这项任务有助于智能体内化环境动态：点击某件物品、前往下一页或决定离开时会发生什么。为了使智能体与对应的人类用户对齐，我们进一步围绕示范生成反事实轨迹。对于每个状态，我们都会考虑替代行动，展开其后果，并提示大语言模型将其与人类动作进行比较、识别次优决策，并提取见解以指导未来选择。这一反思过程产生了一种策略，它经过明确训练，既能与人类决策对齐，又能感知环境动态。在测试时，学习得到的策略驱动智能体与推荐系统进行交互。我们在多个数据集上评估 ALIGNUSER，结果表明，相较于以往基于大语言模型的用户智能体，它在微观层面上与人类的对齐程度更高，同时比传统离线指标更能为推荐系统选择提供可靠指导。

2 相关工作

推荐系统评估。 传统的推荐系统评估主要依赖于在历史日志上计算的离线指标，例如归一化折损累计增益、召回率或均方根误差（Zhang 等人，2019年； Jannach 和Jugovac， 2019年）。尽管这些指标有助于模型选择，但无法直接反映用户体验或商业价值，且与在线 A/B 测试的相关性通常较弱（Jannach 和 Jugovac， 2019年； Bougie 和 Watanabe，2025c）。因此，近期研究开始探索交互式和反事实评估，包括多臂老虎机模拟器、用户模型和因果推断技术（Li 等人，2024年）。

基于大语言模型的智能体。 近年来，大语言模型为在虚拟世界中模拟类人智能体开辟了新的可能性（Park 等人，2023年； Li 等人，2023年； Wei 等人，2022 年）。基于大语言模型的智能体能够通过自然语言进行推理、规划和交互（Wei 等人， 2022年； Bougie 和 Watanabe，2025b； Park 等人， 2023年； Bougie 和 Watanabe，2025a）。多项研究将大语言模型用作推荐场景中的用户模拟器或对话智能体。RecMind（Wang 等人， 2023c）和 In- t eRecAgent（Huang 等人，2023年）提出了在工具增强型智能体上实现规划和反思机制。Agent4Rec（Zhang 等人，2023年）及相关工作（Hou 等人， 2024年； Yoon 等人，2024 年）研究了与推荐模型交互并提供评分或文本反馈的生成式用户智能体。近期，（Bougie 和 Watanabe， 2025c）考虑采用图像驱动感知和高级推理模块，使智能体与人类对应者保持一致。尽管这些系统与在线 A/B 测试呈现出积极相关性（Bougie 和 Watanabe，2025c），但它们通常将智能体策略视为一个黑箱：在单步内将文本状态映射为行动，而没有明确建模行动如何塑造未来状态。此外，智能体通常会根据其角色设定、示例和人口统计属性被指示采取行动，这产生的行为反映了模型先验，而非真实的用户模式。

世界模型与自我反思。 世界模型在强化学习领域有着悠久的历史，用于预测未来状态和奖励（Ha 和 Schmidhuber，2018年）。近期研究将这些思想扩展到语言智能体，将状态和行动表示为文本，并学习下一状态预测器（Kim 等人，2022 年）。STaR （ Zelikman 等人，2022 年）等自我反思策略以及较新的方法（Wang 等人，2023b ）利用思维链解释来提升推理能力和鲁棒性。与我们的工作最为相关的是，近期的“早期经验”方法（Zhang 等人， 2025年）通过生成替代轨迹，并借助环境反馈比较专家动作与替代方案来训练大语言模型智能体。我们秉持类似理念，但目标是使用户代理与人类推荐系统行为对齐；我们引入由角色设定驱动的反思，使角色设定与行动保持一致，并将世界模型引导的反事实与来自人类轨迹的显式监督相结合。

3 问题表述

我们将环境建模为一个马尔可夫决策过程 $\mathcal{M} = (\mathcal{S}, \mathcal{A}, T)$，其中状态$s \in \mathcal{S}$ 是页面的文本表示（例如搜索结果、商品详情、购物车），而$a \in \mathcal{A}$ 是诸如 [搜索], [点击][添加_到_购物车], [购买], [评分],或 [退出]之类的行动。我们假设有一个包含 n 人类轨迹的数据集：

$$\mathcal{D}_{\text{human}} = \{(s_t^{(n)}, a_t^{(n)}, \hat{s}_{t+1}^{(n)}, p^{(n)})\}_t^n, \quad (1)$$

采集自真实用户会话，其中 $a_t^{(n)}$ 表示时间 t处的人类动作，而 $\hat{s}_{t+1}^{(n)}$ 表示后续状态。每位示范者还关联有一个角色设定 p。我们的目标是学习一个策略$\pi_\phi(a \mid s, p)$，由 ϕ参数化，使其与环境交互时所产生的轨迹在微观层面（逐步行动）和宏观层面（会话结果）上都与人类轨迹相似。我们进一步假设存在一个经验数据集 $\mathcal{D}_{\text{rollout}}$，其中经验通过随机交互或遵循好奇心驱动策略收集（Bougie 和 Ichise, 2020），$\{(s_t^{(n)}, a_t^{(n)}, \hat{s}_{t+1}^{(n)}) \ldots\}$。

4 方法

其核心在于，ALIGNUSER 通过根据状态–动作对预测下一状态来理解世界，并通过将人类动作与反事实示例进行比较，使其与人类行为保持一致。在这一步预训练之后，智能体与推荐系统进行交互。图1 展示了整体架构。

4.1 世界建模

我们首先训练基于大语言模型的策略 π_ϕ，使其近似环境转移动态。在我们的研究中，状态完全以自然语言表示，因此可以将下一状态预测建模为标准的下一词元预测目标。受先前将大语言模型训练为世界模型的研究启发，我们使用轨迹集$\mathcal{D}_{rollout}$ 中的下一状态，作为语言智能体策略π_ϕ的直接训练信号。

给定来自 $\mathcal{D}_{\text{rollout}}$ 的状态–动作对（s_t, a_t），模型将下一状态 s_{t+1} 预测为一个词元序列：$\hat{s}_{t+1} \sim \pi_\phi(\cdot \mid s_t, a_t)$，并训练 π_ϕ 以最大化人类下一状态的似然 s_{t+1}^*：

$$\mathcal{L}_{\text{wm}}(\phi) = -\sum_{(s_t, a_t, \hat{s}_{t+1}) \in \mathcal{D}_{\text{rollout}}} \log p_\phi(\hat{s}_{t+1} \mid s_t, a_t). \tag{2}$$

例如，在浏览电子商务网站时，模型可能会学会预测：点击某个产品会进入详细的商品页面，而提交空搜索查询则会显示“无结果”状态。这些自然语言页面描述充当下一状态监督，使模型能够内化不同用户行为如何改变购物会话，而无需任何手工设计的监督。

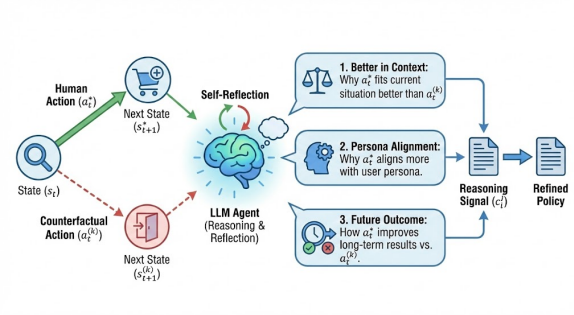

图 2：基于反事实轨迹的反事实反思。

4.2 通过反事实推理实现人类对齐

为了使策略与人类决策保持一致，我们将人类轨迹与反事实情况进行比较（图2）。对于每个人类状态转移$(s_t, a_t, \hat{s}_{t+1}, p) \in \mathcal{D}_{\text{human}}$，我们会采样备选行动$\{a_t^{(1)}, \ldots\ldots, a_t^{(K)}\}$，这些行动被当前策略$\pi_\phi$视为合理，但与示范行动有所不同。给定状态 s_t，我们首先从$\pi_\phi(\cdot \mid s_t, p)$中抽取一组 K 候选行动$\{a_t^{(1)}, \ldots\ldots, a_t^{(K)}\}$，使生成的行动与真实值不同：

$$a_t^{(k)} \sim \pi_\phi(\cdot \mid s_t, p) \quad \text{s.t.} \quad a_t^{(k)} \neq a_t \tag{3}$$

这确保了对模型当前认为合理的行动进行探索，因此这些行动若不加以纠正，最有可能导致不对齐。

然后，我们让智能体通过将反事实状态与人类状态-动作对进行比较，在这些状态上进行推理。给定下一状态后，我们提示大语言模型解释：（1）为什么在当前上下文中，人类选择更好；（2）为什么人类选择更符合其角色设定和偏好；（3）与替代方案相比，人类动作如何改善未来结果。这些解释比单独使用专家动作提供了更丰富、可迁移的监督，并利用大语言模型处理语言的优势，使其内化可泛化至不同任务的决策原则。实际操作中，我们提示模型生成思维链 c_t^j，解释为什么根据行动后状态$\hat{s}_{t+1}$ 与 s_t^j 之间的差异，人类动作 a_t 优于替代方案 a_t^j。该提示词旨在引导自然语言推理，突出a_t^j 中潜在的局限性或低效之处，并以观察到的实际状态转移为依据。

这种反思既适用于环境驱动的行动（例如点击、搜索），也适用于以物品为中心的行动（例如点赞、评分）。这些经验会存储在 $\mathcal{D}_{\text{CR}}$中。随后，我们训练智能体根据状态 s_t 联合预测思维链和专家动作，并针对目标序列（ c_t^j， a_t）使用下一词元预测损失：

$$\mathcal{L}_{\text{CR}} = -\sum_{(s_t,a_t,c_t^j,p)\in\mathcal{D}_{\text{CR}}} \log p_\phi(c_t^j, a_t \mid s_t, p), \tag{4}$$

其中，p_ϕ 表示语言模型的输出分布，并与智能体的策略 π_ϕ保持一致。
学习语言智能体时所求解的整体优化问题可表示为：

$$\mathcal{L}(\phi) = \lambda_{\text{wm}}\mathcal{L}_{\text{wm}}(\phi) + \lambda_{\text{CR}}\mathcal{L}_{\text{CR}}(\phi), \tag{5}$$

其中， λ_{wm} 和 λ_{CR} 是标量，用于平衡世界模型项和反事实项。

4.3 与推荐系统交互

预训练完成后，ALIGNUSER 使用学习得到的策略 π_ϕ 充当合成用户。给定其角色设定 p后，智能体会与推荐系统交互，直至购买物品或决定终止会话。每个智能体都配备了情景记忆，用于存储其与推荐系统的交互记录。记忆最初包含用户的浏览和评分历史。当智能体执行新的行动或为物品评分时，相应的交互记录会添加到情景记忆中。

在每个步骤中，策略 π_ϕ 会接收对当前状态 s_t 的自然语言描述（例如，推荐商品页面），我们会提示 π_ϕ 在内部推理当前情境并输出一项行动，如下方的伪提示词所示：

动作选择提示结构

[状态]s_t[角色设定]p[近期_历史]H[可能的_行动]a_1， a_2， ……， a_M指令：请逐步思考，设想符合该角色设定的谨慎用户接下来会采取什么行动，同时考虑他们的目标、偏好以及每个行动可能带来的未来后果。最后用单独一行按以下格式作答：BEST-ACTION： <行动_标记>推理依据： <推理依据>

随后在环境中执行所选行动（例如点击某个物品、前往下一页或退出），并重复此过程，直到选中终止动作。为了进一步增强智能体对物品进行推理的能力，我们将基础版 ALIGNUSER 与 ALIGNUSER+进行比较。后者整合了图记忆、路径驱动的检索和因果推理，做法与（Bougie 和 Watanabe，2025c）相同。具体而言，智能体将其偏好存储在基于图的记忆中，并检索证据，以判断自己是否喜欢某件物品。在初始动作选择之后 a_{tent}，我们引入一个因果推理步骤，智能体会生成问题（$Q = \pi_\phi$（a_{tent}， H， p， P_{causal}）），根据近期历史验证暂定行动 H 和提示词 P_{causal}。对于每个反事实情景（例如，“如果你现在退出，会发生什么？”），智能体会估计结果，并依据因果一致性调整其最终行动。

5 实验

基线方法我们将 ALIGNUSER 与 RecAgent（Wang 等人，2023a）、Agent4Rec（Zhang 等人，2023年）以及 SimUSER（Bougie 和 Watanabe，2025c）进行比较，这些方法代表了最接近的可比方法。部分实验涉及 AlignUSER 的两个版本：ALIGNUSER 和 ALIGNUSER+，以便单独分析预训练和少样本提示的影响。在可能的情况下，我们还报告 RecMind（Wang 等人，2023c）这一基于智能体的推荐系统的结果。

5.1 实现细节

我们采用 Qwen3-8B 作为框架的骨干LLM。在策略训练期间，我们为每个状态生成 $K = 3$个反事实动作，并从环境中获取其预测的下一状态。该策略使用世界模型损失和反思损失的加权组合进行训练，遵循第 4.2 节。当无法获得角色设定时，我们通过角色设定匹配来估计用户画像属性，具体做法与 SimUSER（Bougie 和 Watanabe，2025c）一致。我们研究了四个真实世界数据集：MovieLens-1M（Harper 和 Konstan，2015年）、Steam（Kang 和 McAuley，2018年）和 AmazonBook（McAuley 等人，2015年）、OPeRA（Wang 等人，2025年）。

5.2 偏好对齐

为了恰当地回应推荐，合成用户必须清楚地了解自身偏好。因此，我们

	MovieLens				AmazonBook				Steam			
方法（1:m）	准确率	精确率	召回率	F1 分数	准确率	精确率	召回率	F1 分数	准确率	精确率	召回率	F1 分数
RecAgent（1:1）	0.5807	0.6391	0.6035	0.6205	0.6035	0.6539	0.6636	0.6587	0.6267	0.6514	0.6490	0.6499
RecAgent（1:3）	0.5077	0.7396	0.3987	0.5181	0.6144	0.6676	0.4001	0.5003	0.5873	0.6674	0.3488	0.4576
RecAgent（1:9）	0.4800	0.7491	0.2168	0.3362	0.6222	0.6641	0.1652	0.2647	0.5995	0.6732	0.1744	0.2772
Agent4Rec（1:1）	0.6912	0.7460	0.6914	0.6982	0.7190	0.7276	0.7335	0.7002	0.6892	0.7059	0.7031	0.6786
Agent4Rec（1:3）	0.6675	0.7623	0.4210	0.5433	0.6707	0.6909	0.4423	0.5098	0.6505	0.7381	0.4446	0.5194
Agent4Rec（1:9）	0.6175	0.7753	0.2139	0.3232	0.6617	0.6939	0.2369	0.3183	0.6021	0.7213	0.1901	0.2822
SimUSER（1:1）	0.7912	0.7976	0.7576	0.7771	0.8221	0.7969	0.7841	0.7904	0.7905	0.8033	0.7848	0.7939
SimUSER（1:3）	0.7737	0.8173	0.5223	0.6373	0.6629	0.7547	0.5657	0.6467	0.7425	0.8048	0.5376	0.6446
SimUSER (1:9)	0.6791	0.8382	0.3534	0.4972	0.6497	0.7588	0.3229	0.4530	0.7119	0.7823	0.2675	0.3987
AlignUSER (1:1)	0.8203	0.8372	0.7969	0.8166	0.8432	0.8427	0.8179	0.8301	0.8138	0.8421	0.8263	0.8340
AlignUSER (1:3)	0.7994	0.8423	0.5987	0.6999	0.6843	0.7784	0.6380	0.7014	0.7641	0.8339	0.6118	0.7058
AlignUSER (1:9)	0.7061	0.8438	0.4273	0.5663	0.6648	0.7827	0.3892	0.5198	0.7284	0.7964	0.3379	0.4745
AlignUSER+ (1:1)	0.8317	0.8483	0.8075	0.8274	0.8546	0.8533	0.8292	0.8416	0.8269	0.8549	0.8376	0.8462
AlignUSER+ (1:3)	0.8119	0.8532	0.6113	0.7121	0.6985	0.7915	0.6511	0.7145	0.7781	0.8461	0.6254	0.7190
AlignUSER+ (1:9)	0.7195	0.8524	0.4451	0.5848	0.6787	0.7935	0.4042	0.5356	0.7413	0.8079	0.3528	0.4922

表 1：MovieLens、AmazonBook 和 Steam 数据集上的用户偏好对齐。所有改进均具有统计显著性（$p < 0.05$）。粗体：各类型的最佳结果 (1:1)、(1:3) (1:9)。

方法	MovieLens		AmazonBook		Steam	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
MF	1.2142	0.9971	1.2928	0.9879	1.3148	1.0066
AFM	1.1762	0.8723	1.3006	1.1018	1.2763	0.9724
RecAgent	1.1021	0.7632	1.2587	1.1191	1.0766	0.9598
RecMind-SI（少样本）	1.0651	0.6731	1.2139	0.9434	0.9291	0.6981
Agent4Rec	0.7612	0.7143	0.8788	0.6712	0.7577	0.6880
SimUSER	0.5020	0.4460	0.5676	0.4210	0.5866	0.5323
ALIGNUSER	0.4693	———	———	———	———	———
ALIGNUSER+	**0.4292**					

表 2：评分预测性能。**粗体**：最佳结果；下划线：次佳结果。ALIGNUSER 的改进具有统计显著性（$p < 0.05$）。

	MovieLens	AmazonBook	Steam	OPeRA
RecAgent	3.01 ± 0.14	3.14 ± 0.13	2.96 ± 0.17	3.05 ± 0.15
Agent4Rec	3.04 ± 0.12	3.21 ± 0.14	3.09 ± 0.16	3.15 ± 0.17
SimUSER（角色设定）	4.41±0.16	3.99±0.18	4.02±0.23	4.05±0.20
ALIGNUSER	4.53±0.15*	4.19±0.17*	4.17±0.21*	4.31±0.19*
ALIGNUSER+	**4.58±0.14***	**4.25±0.16***	**4.22±0.20***	**4.34±0.18***

表 3：GPT-4o 在各推荐领域评估的人类相似度得分（越高越好）。*与最佳基线（$p < 0.05$）相比有显著提升。

要求智能体根据人类对应者是否与物品发生过交互，对物品进行分类。我们为 1,000 个智能体各随机分配 20 件物品，并设置不同的交互物品与未交互物品比例（1:m 其中$m \in \{1, 3, 9\}$）。我们将此视为一项二元分类任务。表 1显示，ALIGNUSER 智能体能够准确识别符合其口味的物品，在所有干扰项等级下均显著优于基线方法（配对 t 检验，95% 置信度，$p < 0.05$）。这些提升可直接归因于反思步骤，它使大语言模型能够理解人设与智能体的行动和偏好之间的关系。与 ALIGNUSER+ 相比，知识图谱记忆也带来了进一步提升。

5.3 评分物品

推荐系统交互的一个核心组成部分，是判断用户是否会喜欢某件物品。我们通过将几种基于大语言模型的智能体与标准基线方法进行比较来评估我们的方法在这一任务上的表现，其中包括矩阵分解（MF）（<style id='1'>Koren 等人</style>，<style id='3'>2009年</style>）以及注意力因子分解机（AFM）（Xiao 等人，2017年）。结果见表2。在各个数据集上，我们的智能体始终取得低于其他基线方法的误差。其他大语言模型方法往往会产生更大的偏差，尤其是在长尾物品或观测数据稀疏的物品上，这反映出它们倾向于生成看似合理但不正确的评分。相比之下，ALIGNUSER 在世界模型预训练阶段明确使智能体的行为与其角色设定保持一致。这为偏好一致性和项目关系提供了辅助信号，使智能体能够在给出评分之前形成更加连贯的内部偏好状态。

5.4 人类相似度

为了评估智能体轨迹与真实用户行为的相似程度，我们采用 GPT-4o 作为自动评估器，这一做法基于先前的证据：大语言模型评审者的可靠性与人工标注者相当（Chiang 和 Lee，2023年）。对于每个交互序列，评估器都会按照 5 点李克特量表打分，分数越高，表示其推理方式和行为模式与人类的对齐程度越高。如表 3所示，我们的方法取得了显著更高的人类相似度

方法	思考–行动一致性 (%) ↑	角色–行为一致性 (%) ↑	页面/会话 ≈ 人类	购买率差距（绝对值，%）↓
人类（OPeRA）	–	–	5.3	–
随机	38.7	36.1	2.4	22.8
RecAgent	49.5	46.7	3.5	16.3
Agent4Rec	55.8	52.4	4.0	12.1
SimUSER	64.3	61.5	4.6	9.9
ALIGNUSER	86.7	82.4	**5.1**	2.5
ALIGNUSER+	**89.3**	**85.6**	**5.1**	**2.1**

表 4：OPeRA-测试集上的思维与角色设定一致性，以及会话级统计数据。**粗体**：最佳结果；带下划线：合成智能体中的次佳结果。

在所有数据集上的得分。我们的世界建模任务减少了不一致的行为，并促使智能体评估人类在不同情境下会如何行动。相比之下，Agent4Rec 等基线大语言模型智能体会表现出一些被评估器可靠地判定为非人类的模式，包括过早执行 [EXIT] 行动，以及对相似物品表现出异常评分行为。

5.5 推理与角色设定一致性

我们进一步衡量智能体在 OPeRA 数据集上复现类人推理和会话动态的能力（Wang 等人，2025年）。该数据集包含状态–行动对及推理依据。首先，我们报告**思考–行动一致性**，由 GPT-4o 比较大语言模型生成的推理依据与真实推理依据，并将其判定为连贯、部分连贯或矛盾。一致性得分是被标记为连贯的步骤所占比例。其次，我们衡量**角色设定–行为一致性**，即行动是否符合所述的购物风格和偏好。这旨在评估智能体能否保持稳定且个性化的行为模式，而非逐渐偏向一般化的购物者。最后，分析会话层面的指标：访问的**页面数**，以及**购买率差距**，其定义为人类与智能体购买频率之间的绝对差值（%）。

如表4所示，反事实反思显著提升了内部一致性。与基线方法相比，ALIGNUSER 提高了思考与行动的一致性。角色设定与行为一致性也呈现出类似趋势，这表明该策略不仅能复现局部决策，还能在整个会话期间保持稳定的购物风格。会话统计数据也更接近人类行为。RecAgent 和 Agent4Rec 往往对网站探索不足（浏览页面数少于人类），而且相较于人类购物者，购买量不是过多就是不足；我们的方法则生成了更贴近真实情况的浏览会话。

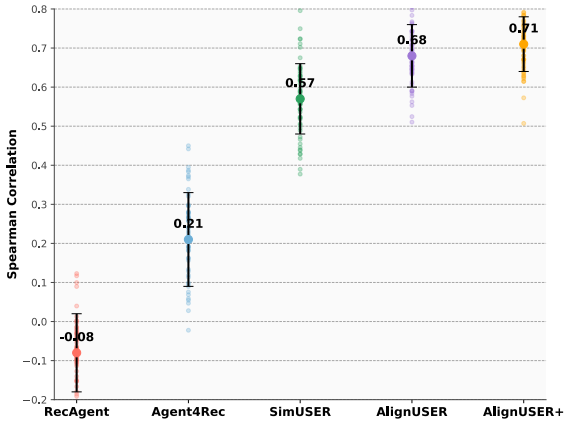

图 3：估计参与度指标与实际参与度指标之间的斯皮尔曼相关系数。数值越高，表示与真实指标的对齐程度越好。

5.6 动作对齐

接下来，我们衡量动作对齐情况。我们采用精确匹配标准：只有当预测的所有参数都与真实值匹配时，才将预测计为正确。对于点击动作，这要求点击目标匹配（例如，正确的产品或按钮）。对于输入动作，模型必须同时识别出适当的输入字段，以及用户输入的确切文本。我们还评估各方法对动作类型的分类效果，并报告点击、输入和终止这几种高层动作类别的 F1 分数。为评估细粒度行为，我们进一步计算点击子类型的加权 F1，以衡量模型能否区分不同的点击意图（例如，评论、产品_链接、购买）。最后，由于在线购物本质上是目标导向的，我们评估会话结果预测。我们使用准确率和加权 F1 衡量模型对这些终端动作的预测表现，从而反映模型在整个会话过程中捕捉用户最终决策和长期目标的能力。如表5所示，ALIGNUSER 超越了以往基于大语言模型的模拟器，而 ALIGNUSER+ 取得了最优结果，在动作生成准确率和会话结果预测方面的提升尤其显著。

5.7 离线 A/B 测试

我们进一步考察 ALIGNUSER 能否作为在线 A/B 测试的可靠代理。我们使用了一个大型美食推荐平台的专有数据集，其中包含 55 次历史 A/B 实验，每次实验都涉及数千个推荐

模型	动作生成（准确率）	动作类型（宏平均 F1）	点击类型（加权 F1）	会话结果（加权 F1）
GPT-4.1	21.51	48.78	44.47	47.54
无角色设定	22.06	45.55	43.45	58.47
无推理过程	21.28	34.93	42.63	51.17
Claude-3.7	10.75	31.58	27.27	43.52
无角色设定	10.75	25.33	22.76	43.10
无推理过程	10.08	26.06	20.29	43.10
Llama-3.3	8.31	24.29	19.99	36.64
无角色设定	8.31	23.69	18.59	33.21
无推理过程	8.76	23.60	19.22	34.19
RecAgent	22.71	49.18	45.25	54.12
Agent4Rec	23.09	50.05	46.37	56.70
SimUSER	24.21	52.44	48.68	59.63
ALIGNUSER	51.47	69.81	66.29	78.07
ALIGNUSER+	**52.92**	**71.94**	**66.88**	**80.52**

表 5：下一动作预测评估。我们报告四项指标：动作生成准确率、动作类型宏平均 *F1*、点击类型加权 *F1*和会话结果加权 *F1*。“Claude-3.7”表示 Claude-3.7-Sonnet；“Llama-3.3”表示 Llama-3.3-70B-Instruct。所有指标均以百分比（%）表示。

	$P_{\text{查看}}$	$\overline{N}_{\text{like}}$	$P_{\text{喜欢}}$	$\overline{N}_{\text{exit}}$	$\overline{S}_{\text{sat}}$
随机	0.295	3.05	0.247	2.80	2.60
Pop	0.388	4.15	0.365	2.95	3.28
MF	0.468	**5.72**	0.439	3.08	3.70
MultVAE	0.521	5.31	**0.452**	3.22	3.82
LightGCN	**0.552**	5.49	0.446	**3.26**	**3.88**

表6：在MovieLens数据集上的推荐任务中对推荐策略的评估。

物品。每次测试都会比较多种推荐策略，并将平均访问页面数作为主要业务指标。对于每种策略，我们都会运行相应的模拟器，并估算相同的参与度指标，然后计算 55 次测试中模拟结果与真实结果之间的斯皮尔曼相关系数。如图所示3，ALIGNUSER+ 与真实值的相关性最高，优于所有其他基线方法。统计检验证实，其相较于所有基线方法的提升均具有显著性（$p < 0.05$），且 ALIGNUSER 也明显优于 SimUSER。

5.8 推荐系统评估

理解各种推荐算法的效能，对于提升用户满意度至关重要。通过模拟人类代理，我们能够更好地预测用户将如何与推荐系统互动，并提供有价值的交互指标。我们使用 MovieLens 数据集比较了多种推荐策略，包括最受欢迎（Pop）、矩阵分解（MF）（Koren 等人，2009年）、LightGCN（He 等人，2020）和 MultVAE（Liang 等人，2018年）。退出时，智能体会在 1 到 10 分的范围内评定满意度。评分高于 3 分即视为表示点赞。指标包括平均观看比例（P_{view}）、平均点赞数（$\overline{N}_{\text{like}}$）、平均点赞比例（$P_{\text{点赞}}$）、平均退出页面数（$\overline{N}_{\text{exit}}$）和平均用户满意度评分（$\overline{S}_{\text{sat}}$）。在此，我们并非评估所提出的框架本身，而是使用模拟用户考察其与推荐系统的互动是否符合文献中确立的趋势。表 6显示，与随机方法和流行度方法相比，智能体对先进推荐结果的满意度更高，这与现实生活中的趋势一致。

6 讨论与局限性

我们的研究表明，加入明确的世界建模任务和反事实自我反思，能够使用户代理更忠实地再现人类交互模式，并为推荐系统提供更可靠的评估信号。尽管取得了这些进展，仍存在若干局限

有待解决。

当前实现完全依赖自然语言页面描述。尽管这种抽象能够在不同领域实现统一建模，但它忽略了真实电子商务界面中存在的细粒度视觉、布局和交互线索。将该框架扩展至多模态表示（例如屏幕截图、产品图像或结构化DOM状态），可以提升世界模型和下游策略的保真度。

训练依赖人类轨迹，而这些轨迹仅覆盖了部分动作空间。尽管世界模型引导的反事实展开通过让智能体接触替代转移来缓解这一局限，但策略在较少被人类访问的状态中，或在与典型浏览路径显著偏离的序列中，仍可能难以进行良好推断。

我们的实验聚焦于时间深度适中的交互场景（例如购物、图书），此类会话通常持续数十步。在新闻阅读、持续移动应用使用或社交媒体信息流等长时域场景中部署推荐系统智能体，可能需要额外组件，例如持久记忆、分层规划或显式的长期目标建模。

最后，智能体的行为不可避免地会继承底层大语言模型的偏见和特性。尽管反思机制限制了偏离人类示范的情况，但偏好建模、情感或感知中残留的偏见仍可能显现，并影响评估结果。

7 结论

我们提出了 ALIGNUSER，这是一种基于世界模型的框架，可从人类轨迹中学习用户代理。ALIGNUSER 通过下一状态预测对环境动态进行建模，并生成反事实，使行动与人类决策和人设相对齐，从而让基于大语言模型的智能体更贴近真实的用户行为。我们在购物和推荐系统评估场景中的实验表明，该方法在动作预测、评分和偏好对齐方面均有所提升。我们的方法还显示，模拟结果与在线 A/B测试结果之间存在正相关。我们相信，合成用户为可扩展且保护隐私的推荐系统评估提供了颇具前景的基础，并为更真实、可控且可解释的

基于智能体的仿真框架。

8 局限性

尽管在评估的基线方法中，ALIGNUSER 与人类轨迹的对齐程度最高，但仍须承认其存在若干局限。首先，可复现性受人类交互日志可用性的限制。用于评估对齐程度的少数数据集为专有数据集，限制了完全透明和复现的可能性。

其次，我们的方法继承了大语言模型中存在的文化、人口统计和社会经济偏见。由于 ALIGNUSER 依赖大语言模型生成的反思和反事实推理，基础模型中的任何潜在偏见都可能表现为对用户动机或偏好的片面解读。与此相关的是，我们偶尔会在世界模型展开中观察到幻觉，例如预测出不合理的下一状态或误解页面语义；这些问题可能会传播到反思策略中，导致决策欠佳。

第三，ALIGNUSER 的有效性与底层大语言模型的优势和局限紧密相关。推理质量不一致、在分布偏移下表现脆弱，以及偶尔出现缺乏依据的判断，都可能降低模拟用户的保真度，尤其是在状态-动作空间覆盖稀疏的区域。

最后，该框架整合了多个相互作用的组件，包括世界建模、反事实生成和反思，这可能导致难以单独评估各模块的贡献。虽然我们通过消融研究在一定程度上区分了这些影响，但仍需开展后续研究，以更好地了解不同训练信号和架构选择如何影响对齐结果。

9 伦理声明

本研究提出了一个用于训练合成用户、以支持推荐系统评估的框架。尽管此类智能体在可扩展性、成本效益和隐私保护方面具有明显优势，但这一方法也引发了若干伦理方面的考量。

在人工日志上训练的合成用户智能体，可能会无意中再现或放大底层大语言模型或用于世界模型学习的行为数据中存在的人口、文化或社会经济偏见。这些偏见可能会体现在

智能体的模拟偏好或交互模式中，从而产生误导性的评估信号。尤其是，有偏反思或幻觉式反事实可能会偏袒某些用户群体或产品类别，进而加剧下游推荐系统中的不公平现象。

一个更广泛的担忧在于，将逼真的合成用户用作真实个体的代理。当此类智能体被用于评估或优化推荐系统的行为时，系统设计者可能会过度依赖模拟结果，从而减少真实用户、领域专家或受影响利益相关者的参与。在电子商务、职位推荐或内容消费等领域，这一点尤其值得关注，因为算法决策可能影响用户自主权及其接触信息的机会。

最后，生成反事实轨迹虽然有助于对齐，但依赖的模型可能会给出听起来合理、实则不符合事实的预测。如果不谨慎解读，这类不准确的预测可能误导评估或优化。

我们强调，合成用户应当补充而非取代真实用户的反馈。负责任的部署要求公开说明模型局限性、持续监测偏见，并采取防护措施，防止模拟结果遭到滥用或误读。

参考文献

James Bisbee、Joshua D. Clinton、Cassy Dorff、Brenton Kenkel 和 Jennifer M. Larson。2024年。用合成数据取代人类调查数据？大语言模型的危险。*Political Analysis*，32(4):401–416。

Nicolas Bougie 和 Ryutaro Ichise。2020。面向内在动机强化学习的基于技能的好奇心。*MachineLearning*，109：493–512。

Nicolas Bougie 和 Narimawa Watanabe。2025a。Citysim：利用大规模大语言模型驱动的智能体模拟对城市行为与城市动态建模。载于*2025*年自然语言处理经验方法会议：产业赛道论文集，第 215–229 页。

Nicolas Bougie 和 Narimawa Watanabe。2025b。生成式评论者智能体：可扩展的同行评审模拟体。载于*2025*年自然语言处理经验方法会议：产业赛道论文集，第 98–116 页。

Nicolas Bougie 和 Narimawa Watanabe。2025c。Simuser：利用大型语言模型模拟用户行为以进行推荐系统评估。

载于 计算语言学协会第*63*届年会论文集（第*6*卷：产业赛道），第43–60页。

Cheng-Han Chiang 和 Hung-yi Lee。2023。大语言模型能否成为人工评估的替代方案？*arXiv* 预印本 *arXiv:2305.01937*。

David Ha 和 Jürgen Schmidhuber。2018年。循环世界模型促进策略演化。载于神经信息处理系统进展，第31卷。

F Maxwell Harper 和 Joseph A Konstan。2015年。MovieLens数据集：历史与上下文。*ACM*交互式智能系统汇刊（*TIIS*），5(4):1– 19。

Xiangnan He、Kuan Deng、Xiang Wang、Yan Li、Yong- dong Zhang 和 Meng Wang。2020。Lightgcn：简化并增强用于推荐的图卷积网络。载于第*43*届国际*ACM SIGIR*信息检索研究与发展大会论文集，第639–648页。Yupeng Hou、Junjie Zhang、Zihan Lin、Hongyu Lu、Ruobing Xie、Julian McAuley 和 Wayne Xin Zhao。2024。大语言模型是推荐系统的零样本排序器。载于欧洲信息检索会议，第364–381页。Springer。Xu Huang、Jianxun Lian、Yuxuan Lei、Jing Yao、Defu Lian 和 Xing Xie。2023。推荐人工智能智能体：集成大语言模型以实现交互式推荐。*arXiv*预印本 *arXiv:2308.16505*。Dietmar Jannach 和 Michael Jugovac。2019。衡量推荐系统的商业价值。*ACM*管理信息系统汇刊（*TMIS*），10(4)：1–23。Carolin Kaiser、Jakob Kaiser、Vladimir Manewitsch、Lea Rau 和 Rene Schallner。2025。利用大语言模型模拟人类意见：个性化调查数据建模的机遇与挑战。载于第*33*届*ACM*用户建模、自适应与个性化会议附属会议论文集，第82–86页。计算机协会。Wang-Cheng Kang 和 Julian McAuley。2018。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。IEEE。Minsoo Kim、YeonJoon Jung、Dohyeon Lee 和 Seung-won Hwang。2022。用于文本游戏的基于预训练语言模型的世界模型。载于*2022*年自然语言处理经验方法会议论文集，第1324–1341页。计算语言学协会。Yehuda Koren、Robert Bell 和 Chris Volinsky。2009。推荐系统的矩阵分解技术。《计算机》，42(8)：30–37。

Guohao Li、Hasan Hammoud、Hani Itani、Dmitrii Khizbullin 和 Bernard Ghanem。2023。Camel：用于探索大语言模型社会“心智”的交际智能体。《神经信息处理系统进展》，36：51991–52008。

Yang Li、Kangbo Liu、Ranjan Satapathy、Suhang Wang 和 Erik Cambria。2024年。推荐系统的最新进展：综述。*IEEE*计算智能杂志，19(2)：78–95。

Dawen Liang、Rahul G Krishnan、Matthew D Hoffman 和 Tony Jebara。2018年。用于协同过滤的变分自编码器。载于*2018*年万维网大会论文集，第689–698页。

Julian McAuley、Christopher Targett、Qinfeng Shi 和 Anton Van Den Hengel。2015年。基于图像的风格与替代品推荐。载于第*38*届国际*ACM SIGIR*信息检索研究与发展会议论文集，第43–52页。

Joon Sung Park、Joseph CO’Brien、Carrie J Cai、Meredith Ringel Morris、Percy Liang 和 Michael S Bernstein。2023年。生成式智能体：人类行为的交互式拟像。*arXiv*预印本 *arXiv:2304.03442*。

Aadesh Salecha、Molly E Ireland、Shashanka Subrah-manya、João Sedoc、Lyle H Ungar 和 Johannes C Eichstaedt。2024年。大语言模型在问卷回答中表现出类人社会赞许偏差。*arXiv* 预印本 *arXiv:2405.06058*。

Lei Wang、Jingsen Zhang、Xu Chen、Yankai Lin、Rui-hua Song、WayneXin Zhao 和 Ji-Rong Wen。2023a。Recagent：一种用于推荐系统的新型仿真范式。*arXiv* 预印本 *arXiv:2306.02552*。

Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc V. Le、Ed H. Chi、Sharan Narang、Aakanksha Chowdhery 和 Denny Zhou。2023b。自洽性能够提升语言模型中的思维链推理。载于国际学习表征会议论文集。

Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu 和 Yingzhen Yang。2023c。Rec-mind：由大语言模型驱动的推荐智能体。*arXiv* 预印本 *arXiv:2308.14296*。

Ziyi Wang, Yuxuan Lu, Wenbo Li, Amirali Amini, Bo Sun, Yakov Bart, Weimin Lyu, Jiri Gesi, Tian Wang, Jing Huang 和其他 1 位作者。2025年。OPeRA：用于评估大语言模型模拟人类在线购物行为的数据集，包含观察、角色设定、推理依据和行动。*arXiv* 预印本 *arXiv:2506.05606*。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Fei Xia、Ed Chi、Quoc V Le、Denny Zhou 及另外1位作者。2022年。思维链提示能够激发大语言模型的推理。*Advancesin neural informationprocessingsystems*，35：24824–24837。

Jun Xiao、Hao Ye、Xiangnan He、Hanwang Zhang、Fei Wu 和 Tat-Seng Chua。2017年。注意力因子分解机：通过注意力网络学习特征交互的权重。*arXiv* 预印本 *arXiv:1708.04617*。

Se-eun Yoon、Zhankui He、JessicaMaria Echterhoff 和 Julian McAuley。2024年。评估大语言模型作为对话式推荐生成式用户模拟器的效果。*arXiv* 预印本 *arXiv:2403.09738*。

Eric Zelikman、Yuhuai Wu、Jesse Mu 和 Noah D. Goodman。2022年。STaR：通过推理进行推理自举。载于 *AdvancesinNeuralInformationProcessing Systems*，第35卷。

An Zhang、Leheng Sheng、Yuxin Chen、Hao Li、Yang Deng、Xiang Wang 和 Tat-Seng Chua。2023年。关于推荐中的生成式智能体。*arXiv* 预印本 *arXiv:2310.10108*。

Kai Zhang、Xiangchao Chen、Bo Liu、Tianci Xue、Zeyi Liao、Zhihan Liu、Xiyao Wang、Yuting Ning、Zhaorun Chen、Xiaohan Fu 及另外1人。2025年。通过早期经验进行智能体学习。*arXiv*预印本 *arXiv:2510.08558*。

Shuai Zhang、Lina Yao、Aixin Sun 和 Yi Tay。2019年。基于深度学习的推荐系统：调查与新视角。*ACM*计算调查（*CSUR*），52(1):1–38。

A 实验设置

实验设置。我们采用基于时间的划分，将数据集分为训练集、验证集和测试集（80/10/10%）。这样可以反映现实世界中可能出现的时间分布偏移。为解决隐私问题，我们省略姓名和性别。此外，为了确保普适性，我们不使用这些数据集中提供的用户专属信息，而是依赖通过角色设定匹配识别出的人设（Bougie 和 Watanabe，2025c）。为了获得能够覆盖环境广阔状态空间的轨迹 $\mathcal{D}_{\text{rollout}}$，我们使用随时间衰减的 ϵ-贪心探索策略收集轨迹。在第 t 回合，行为策略以 ϵ_t 的概率选择随机行动，否则遵循当前策略的贪心行动。我们在 ϵ_t 个回合内，将 0。3 线性退火至 0。05，历时 100，000 个回合。 $\mathcal{D}_{\text{rollout}}$加入了人类转移及其反事实转移。这样可以确保世界模型覆盖人类未曾访问的区域。除非另有说明，推荐模型采用矩阵分解（MF）。在我们的模拟器中，智能体每页会看到四个物品$n = 4$，并可以根据自身偏好浏览和评分物品。当大语言模型的输出偏离所需格式并导致错误时，我们会向其重新提供以下提示词：你还有一次机会给出正确答案。

反事实。 我们为每个状态生成 $K = 3$ 个反事实行动（不包括 a），并从环境中获取其下一状态。训练期间，我们使用批量大小 16、学习率 $1e^{-5}$，训练 8 个轮次。我们将损失权重设为 $\lambda_{\text{wm}} = 1.0$ 和 $\lambda_{\text{CR}} = 0.5$。对于 MovieLens 等不包含行动的数据集（例如，[点击], ），我们只让智能体在物品层级进行反思，方法是采样备选评分行动（[1]，[2],.....），并将不同的评分值视为动作空间中的不同选择。相比之下，对于提供交互动作的数据集，我们还会在轨迹层级生成反事实，从而能够针对备选的动作与状态序列进行反思，而不只是针对孤立的物品层级决策。例如，当监督信号是评分决策时（例如 MovieLens/AmazonBook），我们还会将 {1，2，3，4，5} 中的每个离散评分值视为一个独立行动，并采样备选评分值作为反事实。行动在模拟中执行，以收集 s_{t+1}，随后将其用于反事实反思。

偏好。每个智能体的偏好都存储在记忆中，初始内容取自其对应的人类用户的历史记录。当某件物品的评价分数高于 4 时，智能体会以如下形式存储一条记忆条目：我喜欢 {item_name}，我的评分是 {score}。评分为 2 或更低时，则采用以下格式：我不喜欢 {item_name}，我的评分是 {score}。中性评分对应的条目为：我对 {item_name} 感觉一般，我的评分是 {score}。在所有实验中，用户将评分为 ≥ 4 的物品视为喜欢，而将 ≤ 2 的物品视为不喜欢。这些交互既以纯文本形式存储在情景记忆中，也作为关系存储在知识图谱记忆中。知识图谱记忆采用与 SimUSER 相同的检索实现和参数（Bougie 和 Watanabe，2025c）。返回排名靠前的k_2 件物品及其属性和路径，用于辅助决策（见页面提示词）。具体而言，检索到的物品标题和评分会串接到提示词中。

角色设定。为了为生成式智能体后续的互动和评估奠定可靠基础，每个智能体都有其自身的角色设定 p。角色设定p 包含一组用于刻画用户的特征：年龄、性格和职业。性格特质由大五人格特质定义：开放性、尽责性、外向性、宜人性和神经质，每项均采用 1 至 3 的量表进行衡量。此外，还包括从其历史数据中提取的属性： $p \cup$ {pickiness, habits}。用户的挑剔程度根据其平均评分，在{not picky, moderately picky, extremely picky}中进行采样。习惯反映用户在参与度、从众性和多样性方面的倾向（Zhang 等人，¯2023年）。也就是说，对于用户的平均评分 R：$\bar{R} = \frac{1}{N}\sum_{i=1}^{N} r_{ui}$，根据以下阈值确定用户的挑剔程度 $P(\bar{R})$：

$$P(\bar{R}) = \begin{cases} P_1 & \text{if } \bar{R} \geq 4.5 \\ P_2 & \text{if } 3.5 \leq \bar{R} < 4.5 \\ P_3 & \text{if } \bar{R} < 3.5 \end{cases}$$

其中， P_1 对应不挑剔， P_2 对应比较挑剔，而 P_3 对应极其挑剔。参与度衡量用户与推荐的物品互动的频率和广度，

用于区分高度活跃用户与只和少数物品互动的用户。参与度可用数学公式表示为：$T^u_{act} = \sum_{i\in\mathcal{I}} y_{ui}$，其中，对于用户 $u \in \mathcal{U}$ 和物品 $i \in \mathcal{I}$，物品质量记为 $R_i = \frac{1}{\sum_{u\in\mathcal{U}} y_{ui}} \sum_{u\in\mathcal{U}} y_{ui} \cdot r_{ui}$。 $y_{ui} = 0$ 表示用户 u 尚未为物品 i评分；反之， $y_{ui} = 1$ 表示用户已为该物品评为 $r_{ui} \in \{1, 2, 3, 4, 5\}$。从众性反映用户评分与物品平均评分的接近程度，用以区分具有独特品味的用户与观点接近大众倾向的用户。对于用户 u，从众性特征定义为： $T^u_{conf} = \frac{1}{\sum_{i\in\mathcal{I}} y_{ui}} \sum_{i\in\mathcal{I}} y_{ui} \cdot |r_{ui} - R_i|^2$。多样性反映用户偏好多种物品类别的倾向，或其偏好特定类型的倾向。用户 u 的多样性特征表示为：$T^u_{div} = |U_{i\in\{y_{ui}=1\}} g_i|$。

与推荐系统的交互。预训练完成后，ALIGNUSER 使用学习得到的策略π_ϕ 作为合成用户。给定一个角色设定p，智能体会逐页与推荐模拟器交互，直到选择终止动作。每一步都包括对当前页面上的物品进行内部 [观看]/[跳过] 筛选，以识别与角色设定和记忆一致的候选物品，并选择一个环境动作（例如导航、点击查看详情或退出）。 [观看]/[跳过] 筛选并非环境动作，而是一种内部决策流程，用于减轻用户的心理负荷。在动作选择过程中，我们通过提示词π_ϕ引导智能体在内部对情境进行推理并输出一个动作。所选动作会在环境中执行（例如点击物品以查看更详细的描述、转到下一页或退出），然后循环重复，直至终止。在推荐领域（MovieLens、Steam、AmazonBook）中，会话通过 [退出]终止；而在网络购物领域（OPeRA）中，终止]之前可能还包含与购买相关的决策 [。ALIGNUSER+ 整合了图记忆、路径驱动检索和因果验证，做法与 SimUSER（Bougie 和 Watanabe，2025c）相同。智能体将偏好证据存储在基于图的记忆中，并检索支持路径，以判断自己是否喜欢某件物品。在动作选择之后，我们引入一个因果推理步骤，智能体会生成问题（$Q = \pi_\phi(a_{\text{tent}}, H, p, P_{\text{causal}})$）来验证暂定动作。对于每个反事实情景（例如，“如果你现在退出，会发生什么？”），智能体会估计结果，并根据因果一致性修正最终动作。

B 数据集

MovieLens-1M. MovieLens-1M 数据集是推荐系统研究中广泛使用的数据集。该数据集包含约 100 万条电影评分，评分范围为 1–5 星，由 6,040 名用户针对 3,706 部电影给出。除用户-物品评分交互外，该数据集还包含电影元数据，例如标题和类型标签，以及基本的用户人口统计属性，包括年龄、性别和职业。

Steam. Steam 数据集由从 Steam平台收集的用户-游戏交互数据构成。该数据集包含用户标识符、游戏标识符以及相关的英文评论。数据集还提供游戏级元数据，例如标题。

AmazonBook. AmazonBook 数据集是亚马逊商品评论语料库中限定于图书类别的一个子集。它包含以评分和文本评论形式呈现的用户-物品交互，以及图书级元数据，例如标题和类别信息。

OPeRA. OPeRA 是一个旨在研究和评估用于模拟人类在线购物行为的大语言模型的数据集。它包含真实世界的购物会话日志，其中整合了通过调查问卷收集的用户画像信息、对网页内容的观察、细粒度的用户行动（例如点击和导航事件），以及用户对其决策原因的自我报告。

C 模拟环境

我们的模拟器模拟 Netflix、Steam 等真实世界的推荐平台，采用逐页交互的方式。用户最初会在每个页面看到一系列推荐结果：（i）MovieLens、Steam 和 AmazonBook 的推荐结果；（ii）OPeRA 的网上购物页面。推荐算法被设计为独立模块，因此可以纳入任意算法。该设计预先实现了基于协同过滤的策略，包括随机、最受欢迎、矩阵分解、LightGCN 和 MultVAE。

在**推荐领域**中，环境会将 M 个推荐物品显示在页面上，作为一个文本状态 s_t。对于每个物品，状态都包含

其标题和物品描述。简短描述取自现有领域元数据（若有），否则根据标题生成。如果智能体点击某个物品，模拟器会在下一个状态中显示该物品更详细的描述。

我们将每个页面格式化为：

页面格式（推荐领域）

页面 {page_编号}<– {item_标题} –> <– 历史评分：{item_评分} –> <– 摘要：{item_描述} –> <– 相似项目：{similar_items} –><– {item_标题} –> <– 历史评分：{item_评分} –> <– 摘要：{item_描述} –> <– 相似项目：{similar_items} –>.....

这里的 {item_评分} 是智能体在有记录时自身的历史评分，否则为根据数据集得出的统计值（即全局平均评分）。{similar_items} 会以 **标题（评分/5）** 的形式列出从智能体记忆图中检索到的邻居，仅对 SimUSER 和我们的 ALIGNUSER+ 显示。环境支持以下显式行动：[下一_页]：前往第 (page_编号 + 1) 页。[上一_页]：当页码大于 1 时，返回第 (page_编号 − 1) 页。_ 点击 > 1[物品 _：< 物品 _ 编号 >]：在下一个状态中显示所选物品的详细描述。[退出]：终止会话。

在 OPeRA 等**网页购物领域**中，每个状态包括：(i) 页面上下文，(ii) 产品列表及其属性，这些属性会出现在观察中；(iii) 由语义 ID 标识的交互元素列表。

状态／页面格式（OPeRA）。

页面格式（OPeRA）

页面 {page_number} 上下文：{page_context} 产品：<– {product_title} –> <– 价格：{price}–> <– 可用性：{availability} –> <–详情：{short_description} –><– {product_title} –> <– 价格：{price}–> <– 可用性：{availability} –> <–详情：{short_description} –>...

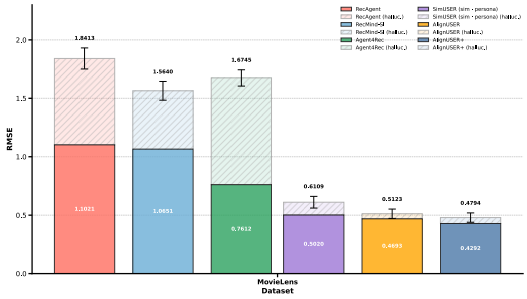

图 4：MovieLens 评分任务中，原始模型（深色）与受幻觉影响的模型（浅色）的均方根误差对比。

交互元素（语义 ID）：{semantic_id_1}、{semantic_id_2}、...、{se- mantic_id_L}

行动采用 OPeRA 数据集中所述的相同动作空间（Wang 等人，2025年），并扩展加入了上述导航行动。

D 附加实验

D.1 幻觉下的物品评分

在本实验中，我们专门针对大语言模型不熟悉的物品，旨在评估我们训练的智能体通过记忆和对齐模块缓解幻觉的能力。与第 5.3 节类似，我们要求用户为电影（MovieLens）评分，但仅纳入大语言模型识别为未知的物品。我们通过向大语言模型查询、让其将每部电影划分为 18 种类型之一，从而确定这些物品 i。如果大语言模型的类型分类与实际类别 g_i一致，则表明大语言模型熟悉该物品，因此会将这类电影排除在实验之外。从图4可以看出，尽管所有方法在幻觉情况下的均方根误差均有所上升，但 ALIGNUSER 和 ALIGNUSER+整体上最具鲁棒性。这种相对较强的鲁棒性可归因于反思与知识图谱记忆的结合：通过利用先前交互中用户、电影和评分之间的关系，智能体可以将陌生电影与相似且熟悉的电影进行比较，并以熟悉的上下文为依据作出预测。

D.2 评分分布

除了个体评分对齐之外，人的代理还必须在宏观层面复现现实世界的行为

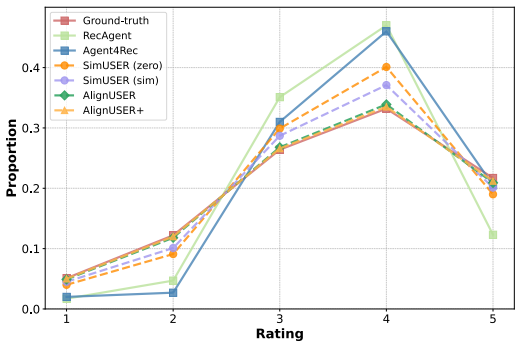

图 5：真实值与人类代理之间评分分布的比较。

模型变体	MovieLens	AmazonBook	Steam	OPeRA
RecAgent	3.01 ± 0.14	3.14 ± 0.13	2.96 ± 0.17	3.05 ± 0.15
Agent4Rec	3.04 ± 0.12	3.21 ± 0.14	3.09 ± 0.16	3.15 ± 0.17
SimUSER（角色设定）	4.41±0.16	3.99±0.18	4.02±0.23	4.05±0.20
ALIGNUSER−WM	4.27±0.17	4.03±0.16	4.06±0.22	4.11±0.18
ALIGNUSER−CR	4.12±0.18	3.91±0.17	3.87±0.21	3.98±0.19
ALIGNUSER−角色设定	4.21±0.15	3.95±0.16	3.92±0.20	4.02±0.18
ALIGNUSER	4.53±0.15*	4.19±0.17*	4.17±0.21*	
ALIGNUSER+	**4.58±0.14***	**4.25±0.16***	**4.22±0.20***	**4.31±0.18***

表7：类人性任务的消融研究。* 相较于最佳基线（$p < 0.05$）有显著提升。

。这意味着要确保智能体生成的评分分布与原始数据集中观察到的分布高度一致。图5 展示了 MovieLens-1M 数据集中的评分分布，以及不同模拟器生成的评分。这些结果表明，模拟评分分布与实际评分分布之间具有高度一致性，其中评分为 4 的数量占主导，而低评分（1–2）较少。尽管 RecAgent 和 Agent4Rec 分配的低评分少于真实用户，SimUSER 缩小了这一差距，而 ALIGNUSER 和 ALIGNUSER+ 最接近真实分布。

D.3 消融研究

为理解我们方法中各组成部分的贡献，我们在四个数据集上对类人性评估进行消融实验。具体而言，我们移除：（1）世界模型目标（–WM），使智能体无法学习环境动态；（2）反事实推理（–CR），禁用与人类决策的对比对齐；以及（3）角色设定基础（–Persona），从策略输入中移除角色信息。结果报告于表7 中。

移除世界模型目标会导致

骨干网络	MovieLens	AmazonBook	Steam	OPeRA
ALIGNUSER+ (Llama-3.2-3B)	4.32 ± 0.18	4.05 ± 0.19	4.07 ± 0.21	4.01 ± 0.20
ALIGNUSER+ (Qwen-2.5-7B)	4.44 ± 0.16	4.13 ± 0.18	4.14 ± 0.20	4.08 ± 0.19
ALIGNUSER+ (Llama-3.1-8B)	4.52 ± 0.15	4.21 ± 0.17	4.20 ± 0.20	4.15 ± 0.18
ALIGNUSER+ (Qwen3-8B)	**4.58 ± 0.14**	**4.25 ± 0.16**	**4.22 ± 0.20**	**4.31 ± 0.18**

表8：ALIGNUSER+ 采用不同骨干LLM时的人类相似度得分，由GPT-4o在四个推荐领域中评估。

人类相似度持续下降，这表明理解环境动态对于生成连贯的交互模式至关重要。消除反事实推理会导致最显著的下降，证实反思人类与反事实之间的差距对于行为对齐至关重要。最后，消融角色设定基础会降低模拟行为的多样性和表现力，尤其是在 MovieLens 和 Steam 上，个人偏好会强烈影响物品选择。完整模型 ALIGNUSER 和 ALIGNUSER+ 的表现优于所有消融版本，凸显了联合训练世界模型动态、角色设定基础和反事实反思的重要性。

D.4 LLM骨干模型选择

我们进一步通过替换基础模型、同时保持 ALIGNUSER+ 的其余部分不变，研究骨干 LLM的影响。如表8所示，所有骨干模型的人类相似度得分均较高，这表明所提出的世界模型学习和反事实对齐对底层模型的选择具有鲁棒性。总体而言，Qwen3-8B的表现最佳，但与Llama-3.1-8B和Qwen-2.5-7B的性能差距仍然不大，这表明大部分对齐收益源自 ALIGNUSER架构，而非骨干模型规模本身。

D.5 运行时间分析

我们比较了 AlignUSER、SimUSER 和 Agent4Rec 在 1,000 次用户交互中的运行时间。Agent4Rec 和 SimUSER 会调用 GPT-4o 的 API，而 AlignUSER 主要使用本地部署的 Qwen2.5 策略（vLLM）进行推理。在不进行并行化的情况下，Agent4Rec 和 SimUSER 分别需要 9.3 小时和 10.1 小时，而 AlignUSER 使用 4 个 GPU 需要 ≈0.6 小时。此外，按 GPT-4o 的定价计算，1,000 次交互的费用约为 16–21 美元，而 AlignUSER 的 GPU 运行时间成本约为 6–8 美元（不包括一次性训练成本）。

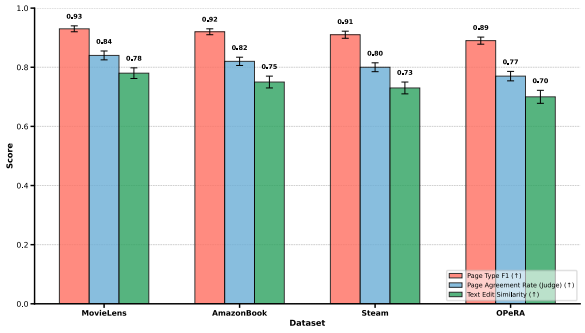

图 6：**ALIGNUSER 的下一状态预测性能**。误差线表示 10 次运行的变化。

D.6 下一状态预测

这项消融实验评估了下一状态预测任务能否捕捉环境的基本动态。对于每个留出转移 (s_t， a_t，s_{t+1})，我们将预测的下一状态与真实状态进行比较。由于下一状态在页面结构和文本内容两方面都存在差异，因此我们采用了三种互补的指标。(i) 页面类型 F1（↑）：针对粗粒度页面类别（例如 浏览、购物车）计算的 F1 分数；(ii) 页面一致率（↑）：由评判模型判断预测状态和真实状态是否描述同一页面；以及 (iii) 文本编辑相似度。根据归一化编辑距离，计算两个页面状态的规范化文本表示之间的相似度。如图 6所示，AlignUSER 能够可靠地预测下一页面状态。评判一致率略低，这是意料之中的，因为即使某些可选细节有所不同（例如展示的物品略有差异），两个状态仍可能是“连贯的”。此外，OPeRA 的小幅下降可归因于网络购物网站中固有的状态多样性增加和内容噪声增多。

D.7 对反事实数量的敏感性。

现在，我们研究 ALIGNUSER 对反思期间每个人工步骤所采样的反事实动作数量的敏感性。除第 4.2 节中的设置外，其他设置均相同，唯一变化的是 K $\in\{0、1、2、3、5\}$，其中 K=0 表示禁用反事实反思（仅进行世界模型预训练）。我们按照表 5中的相同协议，在 OPeRA-test 上报告下一动作预测指标。结果表明，增加 K 能够持续提升下一动作预测性能，其中最大的提升出现在从 K=0 到 K=2的过程中。性能在超过K=3后趋于饱和，这表明额外的反事实监督收益递减，可能是因为它探索了不合理的行动。

D.8 人设匹配准确率

由于人设是模拟多样且异质用户的核心组成部分，我们评估了本研究所采用的自洽人设匹配技术的有效性。我们利用 MovieLens1 数据集，根据用户的交互历史预测用户的智能体和职业。实验结果汇总于表10。总体而言，人设匹配在为模拟智能体补充详细背景信息方面表现出较强的稳健性，包括那些未明确提供人口统计数据的领域。

D.9 对损失权重的敏感性

我们通过改变方程5中的损失权重，分析平衡世界建模与反事实反思的影响。当 λ_{CR}=0时，反事实反思被禁用，训练目标仅简化为世界模型预训练。我们按照与表5相同的协议，报告 OPeRA-test 上的下一动作预测指标。如表11所示，引入反事实反思（λ_{CR}>0）相比仅进行世界模型训练可带来显著提升。围绕默认设置（1。0，0。5）时，性能相对稳定。

E 提示词

E.1 访谈后提示

向每个智能体呈现的访谈后提示词如下：

访谈后提示

您对最近使用过的推荐系统满意吗?

说明：
1. 评分：请给出 1 到 10 分的评分。2. 说明：解释评分理由。

回复格式：
- 评分： [1 到 10]之间的整数- 原因：[详细解释]

模型	动作生成（准确率）	动作类型（宏平均 F1）	点击类型（加权 F1）	会话结果（加权 F1）
AlignUSER（K=0）	37.26	58.41	54.66	69.92
AlignUSER（K=1）	44.53	63.72	60.11	74.35
AlignUSER（K=2）	48.98	67.10	63.89	76.84
AlignUSER (K=3)	51.47	69.81	66.29	78.07
AlignUSER (K=5)	52.08	70.12	66.10	78.66
AlignUSER+ (K=0)	39.41	60.07	56.20	71.58
AlignUSER+ (K=1)	46.00	65.31	61.35	75.32
AlignUSER+ (K=2)	50.41	69.02	64.92	78.11
AlignUSER+ (K=3)	52.92	71.94	66.88	80.52
AlignUSER+ (K=5)	**53.31**	**72.10**	**67.05**	**80.87**

表 9：用于 OPeRA 反思训练的反事实动作数量 K 的敏感性。所有指标均为百分比（%）。 K=0 禁用反思。

指标	Age	职业
准确率	0.7184	0.6691
精确率	0.7512	0.6875
召回率	0.7863	0.7386
F1 分数	0.7683	0.7120

表10：利用 MovieLens-1M 预测年龄和职业时人物匹配的表现。

人类：
{interaction logs}

请按照 1 到 5 的等级进行评分，其中 1 表示最像人工智能，5 表示最像人类。

E.2 合成用户提示的可信度

在第5.2节中，评分提示根据以下说明进行了修改：

合成用户提示的可信度

说明
1. 查看 ## 推荐列表 ## 中的每个{item_类型}。2. 对每个{item_类型}，判断你是否已经与其互动（“已互动”），或尚未与其互动（“未互动”）。

E.3 大语言模型评估器提示

以下提示词用于区分人类与人工智能生成的交互：

大语言模型评估器提示

请评估以下智能体与推荐系统的互动，并判断其是由大语言模型（LLM）人工智能生成的，还是由真人生成的

模型	动作生成（准确率）	动作类型（宏平均 F1）	点击类型 (加权 F1)	会话结果 (加权 F1)
AlignUSER（λ_{wm}=1.0, λ_{CR}=0）	37.26	58.41	54.66	69.92
AlignUSER（λ_{wm}=1.0, λ_{CR}=0.25）	49.92	67.92	64.98	77.41
AlignUSER（λ_{wm}=1.0, λ_{CR}=0.5）	51.47	69.81	66.29	78.07
AlignUSER (λ_{wm}=1.0, λ_{CR}=1.0)	51.88	70.03	66.14	78.29
AlignUSER (λ_{wm}=2.0, λ_{CR}=0.5)	51.21	69.55	66.33	78.01
ALIGNUSER+(λ_{wm}=1.0, λ_{CR}=0)	39.41	60.07	56.20	71.58
ALIGNUSER+(λ_{wm}=1.0, λ_{CR}=0.25)	51.02	69.64	65.37	79.04
ALIGNUSER+(λ_{wm}=1.0, λ_{CR}=0.5)	52.92	71.94	66.88	80.52
ALIGNUSER+(λ_{wm}=1.0, λ_{CR}=1.0)	53.12	72.01	66.74	80.61
ALIGNUSER+(λ_{wm}=2.0, λ_{CR}=0.5)	52.81	71.76	66.92	80.50

表 11：OPeRA 上对损失权重的敏感性（ λ_{wm}， λ_{CR}）。所有指标均为百分比（%）。