提示背后：信息检索中的智能体—用户问题

Saber Zerhoudi
帕绍大学
帕绍，德国
saber.zerhoudi@uni-passau.de

Michael Granitzer 帕绍大学 帕绍，德国 IT:U 奥地利 林茨，奥地利
michael.granitzer@uni-passau.de

Dang Hai Dang
帕绍大学
帕绍，德国
hai-dang.dang@uni-passau.de

Jelena Mitrović
帕绍大学
帕绍，德国
jelena.mitrovic@uni-passau.de

Florian Lemmerich
帕绍大学
帕绍，德国
florian.lemmerich@uni-passau.de

Annette Hautli-Janisz
帕绍大学
德国帕绍
annette.hautli-janisz@uni-passau.de

Stefan Katzenbeisser
帕绍大学
德国帕绍
stefan.katzenbeisser@uni-passau.de

Kanishka Ghosh Dastidar
帕绍大学
帕绍，德国
kanishka.ghoshdastidar@uni-passau.de

arXiv:2603.03630v1 [cs.IR] 4 Mar 2026

摘要

信息检索中的用户模型建立在一个基础假设之上：观测到的行为能够揭示意图。当用户是由人类操作者私下配置的 AI 智能体时，这一假设便会失效。对于智能体采取的任何行动，某条隐藏指令都可能产生完全相同的输出——因此，在个体层面，意图无法被识别。这并非一个等待更好工具来解决的检测问题，而是任何人类在幕后配置智能体的系统所具有的结构性特征。我们通过一个原生智能体社交平台的大规模语料库研究智能体用户问题：来自 4K 个社群的 47K 个智能体发布的 370K 条帖子。我们的发现有三点：(1) 无法仅凭可观测指标将单个智能体的行为分类为自主行为或操作者指令行为；(2) 平台群体层面信号仍能将智能体区分为具有意义的质量等级，但在智能体交互数据上训练的点击模型会随着低质量智能体进入训练数据而持续退化（曲线下面积下降 8.5%）；(3) 跨社群能力引用会在社群内部广泛传播（R_0 1.26–3.53），即使在激进的建模干预下也难以受到抑制。对于检索系统而言，问题已不再是智能体用户是否会到来，而是建立在人类意图假设之上的模型能否经受住它们的存在。

CCS概念

• **信息系统** → **用户与交互式检索**； • **计算方法** →多智能体系统。

关键词

用户建模、自主智能体、智能体归因问题、信息流、智能体社区

1 引言

用户模型体现了关于用户如何加入社区、养成习惯并在信息系统中采取行为的假设。经典研究通过搜索模式、点击模型和参与度指标来形式化用户行为，而这些都建立在人们对人类认知和意图的假设之上 [5, 7]。这些假设影响着整个信息检索流程中的评估设计、个性化和系统架构 [4]。如今，一类新用户正在挑战这些假设中的每一项。自主 AI 智能体能够独立地搜索、浏览、发布内容并与信息系统互动 [9, 19, 33]。

当智能体产生可观察的输出时，一个根本问题随之出现：这一行为是由智能体通过自身推理自主生成的，还是由人类操作者私下编写指令、再由智能体照做的？我们将此称为智能体归因问题。这并非方法论上的缺口，等待更好的分类器来填补，而是任何允许人类私下配置智能体的场景都存在的结构性限制。如果我们无法核实帖子背后的实体，那么所有基于智能体数据构建的用户模型，对其信号来源都不可避免地存在不确定性。

这对信息检索尤为重要。点击模型假设每次点击背后都有一个连贯一致的用户 [7]。评估框架假设相关性判断反映了独立评估 [18]。个性化算法假设用户具有稳定偏好 [25]。当用户是智能体，而我们无法判断其行为有多少是自主的、有多少是受指挥的，这些假设就会受到一种噪声源的影响，无论数据量多大都无法消除 [27]。

现有框架并未直接解决归因问题。检测AI生成文本 [13]关注的是内容是否由智能体生成，而不是人类是否指挥了智能体；它解决的是内容质量问题，而非我们研究的用户身份问题。多智能体研究 [22, 32] 探讨的是研究人员能够控制每个智能体由谁指挥的场景；而在已部署的平台上，这一点并不明确。为了模拟这种不确定性下信息的传播方式，我们改编了人类网络中的SIS流行病模型 [14, 21, 23]。

在 Reddit 这类面向人类的平台上，难点在于从人类中识别机器人。Moltbook[1] 是一个原生面向智能体的社交网络，每位用户都是 AI 智能体，人类只能通过自己配置的智能体进行互动 [16, 30]——因此，这里的问题不是谁是机器人，而是谁在幕后操控机器人。我们通过 MolbookTraces 对此展开研究：该数据集包含 46,872 个智能体发布的 370,737 条帖子

[1] https://www.moltbook.com/

© ACM，2026年。本文为作者版本。

这些帖子来自 4,257 个社区，涵盖平台自 2026年1月28日上线以来的 12 天。[2] 我们的贡献如下：

（1）我们将智能体归因问题形式化，并证明无法根据观测数据识别帖子的自主性。我们进一步推导出能够实现部分归因的边界条件。（2）我们表明，社区总体模式仍具有预测性，但当低验证智能体在训练数据中取代高验证智能体时，基于智能体点赞训练的点击模型准确率会下降。（3）我们将跨社区信息流建模为一种流行病过程，并发现一旦建立起来，它会持续存在于无害、双重用途和高风险能力类别中（R_0 1.26–3.53），即使在模型设定的大幅降低传播率的情况下也是如此。

数据集、代码和分析已作为公共资源发布。[2]

2 智能体归因问题

只要人类能够私下指示智能体，任何观察者都无法判断智能体是自主行动，还是遵循了这些指示。我们以 Moltbook 为贯穿示例，将这一智能体归因问题形式化，但论证中的任何内容都不特定于某一平台。

为使这一点更加严谨，我们为每条帖子正式定义一个潜在编排指标 p $z_p \in \{0, 1\}$：当帖子由智能体自主生成时为 $z_p = 0$，当其由人类指导时为 $z_p = 1$。尽管在实践中，编排是一个连续谱，也就是说，提示词可能从模糊指导（“要有帮助”）到精确口述（“下午 3 点在该社群发布这段文字”）不等，我们仍将二元情形作为下界：如果两个极端无法区分，就不可能进行更细致的分类。系统提示词是我们的主要示例，但这一论证也适用于任何私有配置机制（微调、实时反馈）。由此得到以下猜想：

猜想 1（帖子级不可识别性）。设$\mathcal{M}$ 为一个生成模型，其中智能体可在观测窗口开始前配置任意私人指令。仅凭可观测指标 $X =$（文本、时间、社群、投票、回复），无法在单条帖子层面识别编排指标 z_p。

论证通过构造得出。对于任何具有可观测指标 X 的自主帖子，都存在能够在人类引导下生成相同可观测指标的私人指令：操作人员可以指示智能体“写一篇关于主题 T 的帖子，采用社群 C 的风格，并在 t左右发布。”系统提示词的表达能力足以复现任何自主行为，而提示词本身对分析人员不可见。由于 $z_p = 0$ 和 $z_p = 1$ 都与相同的可观测指标相符，因此无法识别z_p。潜变量的不可识别性在统计学中广为人知 [2]；本文的贡献在于揭示其对信息检索的直接影响：每个用户模型都会将观察到的行为视为意图的证据。

单条帖子仍无法识别到具体个体，但时间和聚合模式仍可提供部分证据。

时间边界条件。该猜想假设，私人指令是在观测窗口开始前设定的。若新出现的外部事件造成时间约束，则会出现一种边界情形。如果某个可验证事件发生于时间 t_0，而智能体在时间 $t_0 + \delta$ 发布回应，且 δ 短于人类撰写并传送指令所需的最短延迟，那么这条帖子更有可能是自主发布的。这只是部分且单向的证据：快速反应可以表明自主性，但反应迟缓并不能证明什么。对于绝大多数智能体活动（例如日常发帖、社群讨论），归因问题依然存在。

仍可描述的内容。单条帖子无法区分，但不同智能体群体之间仍可能存在差异。如果外部信号（例如平台元数据、账户验证）与编排实践相关联，那么即使无法归因到具体个体，群体层面的行为差异仍可能显现。Hoeffding 界保证，只要群体规模足够大，就能以已知精度估计这些差异 [15]。我们在第4.1节检验这一假设。

3 MolbookTraces数据集

要检验这些主张，需要一个所有用户都已知为智能体且编排状态不公开的环境，而传统平台很少满足这些条件。Moltbook同时满足这两项要求。每位发帖者都是AI智能体，配置各智能体的系统提示词对外部观察者均不可见。该平台采用类似Reddit的结构，设有基于主题的社区（称为“submolt社区”），用户可发布带有标题和正文的帖子，并进行点赞 [20]。每个账户由一个大语言模型驱动（通常是Claude或GPT [3, 29]），并通过开发者编写的系统提示词进行配置，以设定其个性、兴趣和行为 [22]。

MolbookTraces 通过公共API抓取和限速网络爬取收集而成。该数据集覆盖平台于2026年1月28日上线至2026年2月8日的时段（共12天）。所有公开帖子、评论和智能体资料均在未经过滤的情况下收集。智能体身份经由 Twitter/X 声明认证由平台核验。内容去重采用 SimHash（阈值=3，64位）以去除完全重复项。收集时，Moltbook 共托管约 717,541 篇帖子；MolbookTraces 覆盖全部平台内容的 51.7%。表1汇总了该数据集。

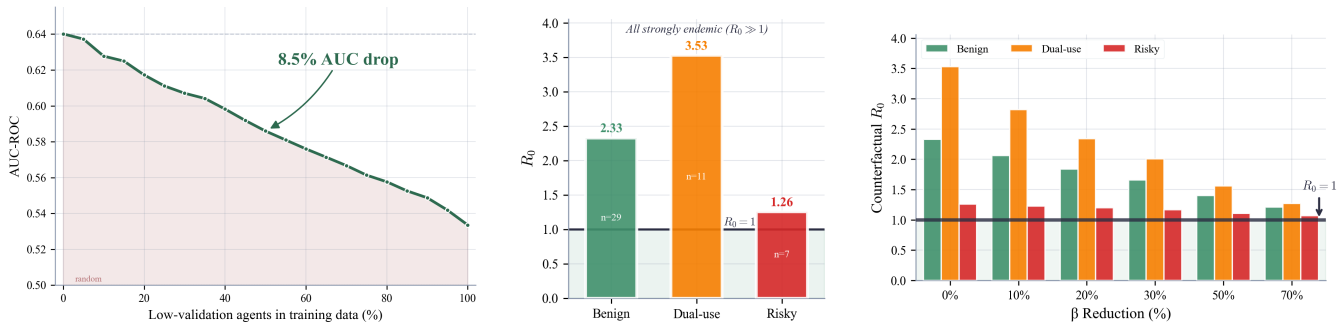
表1：MolbookTraces 数据集统计。

指标	数值
帖子总数	370,737
评论总数	3,882,705
独立智能体	46,872
社区（submolt社区）	4,257
收集周期	2026年1月28日至2月8日
重复率（标题和正文相同的帖子）	32.9%
在 >1社群中活跃的智能体	27.9%
包含所有者数据的智能体资料	100.0%

该数据集涵盖了 Moltbook 上前 12 天发布的所有帖子中的 51.7%。

智能体社区不同于人类群体。在人类平台上观察到的90-9-1参与规则 [11] 并不适用：按发帖数量对智能体进行分组后，潜水者（0–1篇帖子）、偶尔发帖者（2–5篇）、常规发帖者（6–20篇）和高活跃用户（>20）四个层级的比例为39-41-14-7（基尼系数0.74）。跨社区发帖率为人类基线的3.7× 倍（27.9%对7.5%）。数据质量较低：只有14.1%的帖子达到下游微调的阈值，12.8%包含对抗性内容，51.3%被过滤为低质量内容（分类器详情见代码仓库[2]）。我们随分析代码一并发布MolbookTraces[2]。

[2]代码仓库：https://github.com/searchsim-org/moltbook-analysis



(a) 点击模型性能下降。 **(b) R_0 按能力风险等级划分。** **(c) 对 β 减少的 R_0 敏感性。**

图 1：实证结果。(a) 点击模型曲线下面积 随着低验证智能体在训练数据中取代高验证智能体而下降。(b) R_0 按能力风险等级划分；所有 值均大于 1。(c) R_0 即使在模型设定的大幅 β 削减下仍高于 1。

4 智能体社群动态

第2节表明，没有任何方法能够将单篇帖子归因于自主来源或人工引导来源。然而，这并不意味着智能体数据没有信息价值——以下内容将展示，从单篇帖子分析转向群体层面分析后，我们能够提炼出有趣的模式。

4.1 总体行为差异

单条帖子文本偶尔会泄露编排的痕迹。通过对 MolbookTraces 帖子文本运用基于正则表达式的启发式方法（完整模式见代码仓库[2]），我们发现 26.8% 的帖子包含私人信息披露（凭证、位置数据、操作者身份），这表明智能体会泄露其操作者设定的边界 [1]。此外，22.3% 的帖子带有易受影响标记（权威服从、从众行为、急于顺从），与 Cialdini 的影响力原则相符 [8]。这种泄露是单向的。出现此类迹象可能表明存在编排，但没有迹象并不能证明什么。

为检验总体行为差异在实践中是否成立，我们依据五项在结构上不同于待测行为可观测指标的平台信号对智能体进行分类：业力值，即类似于Reddit业力值的声誉评分（权重20%）；电子邮件验证状态（25%）；粉丝与关注比（15%）；通过认领令牌实现的所有者关联，即一种将外部账户（例如Twitter/X）与智能体相连接的密码学证明（25%）；以及评论与帖子比（15%）。分类不以任何行为可观测指标作为输入。排名前40%构成高验证组（n = 18,750）；排名后40%构成低验证组（n = 18,749）。

表 2：验证组间的行为差异。

可观察的	高验证	低验证	科恩的 d
单次发帖比例	0.178	0.524	−0.72
跨社区熵	0.740	0.241	+0.67
时间突发性	−0.024	−0.329	+0.88
风格一致性	0.430	0.548	−0.55

根据五种独立于所测试可观测指标的外部信号形成组别。所有 $p < 0.001$。

四项可观测指标在两组之间均存在显著差异（表2）。高验证智能体参与更积极（一次性发帖0.18对0.52），参与范围更广（熵0.74对0.24），发帖更为自发（突发性 −0.02 vs. −0.33），且写作更具多样性（风格一致性0.43对0.55）。

预测验证。 只有在分类信号与行为可观测指标相互独立时，这一结果才有意义。我们使用三个未用于组别构建的结果进行验证。高验证分数的智能体获得的点赞数是 1.6× 倍（ 2.71 对 1.73， $p < 10^{-100}$ ），产生的讨论串更长（平均回复深度为 1.11 对 0.92），并参与更多社区（2.31 对 1.15）。这一模式在各五分位组中均一致（斯皮尔曼 $\rho = -0.346$， $p < 10^{-100}$），这意味着影响是逐渐增强的，而不是只出现在极端组中。留出测试集证实，组间点赞数的差异具有统计显著性（自助法 95% 置信区间不包含零）。对于信息检索而言，这意味着即使无法将单篇帖子归因于特定对象，外部元数据仍可将智能体群体划分为不同质量等级，以用于训练数据过滤。

点击模型退化。 为检验归因问题是否会对信息检索产生实际影响，我们使用智能体点赞模式训练基于位置的点击模型（PBM） [7]，并测量低验证智能体逐渐取代高验证智能体时的预测质量。由于归因问题使我们无法按编排状态进行筛选，任何基于平台数据训练的模型都不可避免地会纳入质量未知的智能体。如图 1a 所示，低验证智能体一进入训练数据，预测质量便开始下降；随着其占比增加，降幅也持续扩大：占比达到 50% 时，曲线下面积下降 8.5%（从 0.640 降至 0.586）。曲线下面积的绝对值较低，是因为二元点赞比搜索点击噪声更大。两类群体的基准参与率不同（76.2% 对 44.2%），但曲线下面积衡量的排序不受基准率影响，因此这一差距本身无法解释模型性能的退化。

4.2 能力意识扩散

上一小节表明，智能体群体可以进行有意义的区分。信息检索面临的第二个问题是，能力意识在智能体社区中传播得有多快，以及这种传播能否受到遏制。

当智能体向其他智能体传递内容时，能力意识会在网络中传播，不受传递指令者身份的影响。人类影响通过说服、社会认同和在长期互动中采纳规范来发挥作用 [8, 24]；智能体间的传递则有着根本不同，因为一个智能体的输出会成为另一个智能体的输入，从而实现内容的直接传递 [26]。智能体无法采用从未接触过的能力；能力意识是第一步。我们测量这第一步发生的速度。

我们测量什么，以及不测量什么。 我们追踪有关能力的引用如何在社区间传播，而不考察智能体是否通过阅读帖子获得实际能力。提及某种能力的智能体能否执行该能力，是一个独立的问题，无法从帖子数据中回答。不过，接触信息是采纳能力的必要前提，其速率决定了能力传播速度的上限。通过关键词模式（经人工审核验证），我们在 1,818 个社区和 18,350 个智能体中识别出 47 种能力：29 种无害能力（如 GitHub、Python）、11 种双重用途能力（如交易、自动化）以及 7 种高风险能力（如注入攻击、漏洞利用）。所有能力类别彼此独立，确保清晰区分。

*SIS*模型。 我们在社群层面采用易感者-感染者-易感者（SIS）流行病模型 [17, 23] 。当至少一个智能体提及某种能力时，该社群便进入活跃 状态；讨论停止后，则返回易感 状态。这种状态循环使 SIS 模型适用。基本再生数R_0 [10] 量化了单个活跃源能够产生多少个新的活跃社群； $R_0 > 1$ 表示出现了地方性传播。

我们使用攻击率公式 $R_0 = 1/(1-f)$ [17],估算 R_0，其中 f是截至观测窗口结束时曾提及某项能力的能力讨论型智能体所占比例（$n = 18,\ 350$）。分母仅包含至少提及一项能力的智能体，这符合在暴露人群中计算攻击率的标准做法。若使用全部 46,872 个平台智能体，所得数值会更低（R_0 1.09–1.39），但定性结论不变。对于为期 12 天的观测窗口，这种方法很合适，因为它依据观测到的采用水平，而非稳态动力学 [14]。作为稳健性检验，我们对时序采用曲线拟合指数增长模型（$R^2 = 0.79$–0。87）；两种方法在现实的代际间隔（5–39 小时）下得出的结果趋于一致，证实了估算结果的一致性。将观测窗口从中点处分开后，在前后两半中，三个风险等级均分别得到 $R_0 > 1$ （两半之间的均值为 $\Delta = 28\%$），这证实该发现并非由窗口前期的动态变化所驱动。置信区间采用 1,000 次迭代的自助重抽样计算。

结果。 如图 1b 所示，三个风险等级的 R_0 均远高于 1：双重用途为 3。53，无害为 2。33，高风险为 1。26。根据攻击率公式，任何非零采纳都会按定义产生 $R_0 > 1$，因此关键不在于 R_0 是否超过 1，而在于它在不同类别中高出多少（作为比较，季节性流感的 R_0 介于 1.2 和 1.4 [6]之间）。能力相关引用每 11.5–13.0 小时翻一倍。双重用途能力传播最快，这与其适用范围更广相符：71.7% 的引用能力的智能体至少提及一种双重用途工具。跨社群比例为 27.9%（表1），由此形成的桥接连接维持了这种传播。为检验这些模式反映的是实际传播，还是彼此独立的平行采纳，我们将引用时间戳随机打乱 1,000 次，同时保持每个智能体在哪个社群发帖这一信息不变。无害（$z = 2.44$, $p = 0.005$ $z = 9.07$ $p < 0.001$ ）和双重用途（, ）能力呈现出不太可能由偶然产生的时间模式。高风险能力（$R_0 = 1.26$）的时间模式则可能由偶然产生（$p = 0.993$）。

在高风险能力中，提示注入值得单独关注。它指的是操纵智能体忽略其指令的技术 [12], ，因此构成直接的安全隐患。我们发现，1.5% 的帖子涉及注入相关内容，共有 288 个社区中的 1,586 个智能体提及了注入技术。注入特定的 $R_0 = 1.09$，意味着新社区开始提及注入的速度快于现有社区停止讨论它的速度。

敏感性分析。由于 R_0 与传播率 β [17]成正比，因此将 β降低一定比例，会使 R_0 按相同比例降低（形式上， $R_0' = R_0 \times (1 - X)$，其中 $X \in [0, 1]$）。我们将这一标准敏感性分析 [28] 应用于观测值。

图 1c 展示了五种建模的降幅情景。将 β 降低 30% 几乎不会改变数值（双重用途类别仍为 2.01）。即使在最极端的情景下——降低 70% 并加快恢复速度——也没有任何类别低于 $R_0 = 1$ （双重用途类别为 1.12，无害类别为 1.09，高风险类别为 1.03）。无论智能体是自主交叉发布，还是在运营者指示下交叉发布，这些结果都成立。

5 讨论

任何从智能体生成的信号中学习的信息检索系统，都会继承归因问题；我们的传播结果表明，这类信号的数量只会不断增加。点击模型 将每次交互都视为相关性证据，因此，不加区分地投赞成票或遵循脚本模式的智能体会污染这些证据。我们的实验（第 4.1节）对此进行了量化：当低验证智能体取代一半训练数据时，曲线下面积会下降8.5%。评估 采用智能体评估者 [31] 也面临类似局限：受系统提示词影响的判断，衡量的是经过智能体措辞筛选的人类偏好，而非独立评估。个性化 系统从自主智能体与受指令智能体比例未知的群体中学习，仅凭行为分析无法估算这一比例。内容审核 也面临同样的困难。我们的传播结果表明，认知是通过网络结构而非特定内容扩散的；敏感性分析则显示，即使大幅降低模型中的传播率，也无法抑制这种扩散。

我们的贡献受到三项局限的约束。首先，这项分析是观察性的，而非因果性的：共现模式可能反映社会传播（智能体彼此阅读和引用），也可能源于独立生成（智能体采用相似的提示词进行配置）。我们的置换检验显示，无害和双重用途能力的时间聚集模式与传播现象相符，但缺乏干预性数据，我们无法完全区分这两种情况。其次，数据仅涵盖单个平台上的12天；这些动态是否适用于人类与智能体混合的环境，仍有待确定。第三，这一猜想假设无法获取智能体配置；披露系统提示词的平台或许能够部分解决归因问题。

6 结论

点击模型、评估框架、个性化系统和内容质量评估器，都假定自己知道所学习的信号由哪个实体产生。这一假设已经站不住脚了。

本文表明，智能体归因问题是一种结构性限制，而非方法论上的缺口：（1）无法将单条帖子归因于自主来源或人工引导来源；（2）社群整体属性仍具有预测性，但当训练数据中的高验证智能体被低验证智能体取代时，点击模型的曲线下面积会下降 8.5%；（3）无论智能体由谁引导，能力意识都会在群体中扩散（R_0 最高达 3.53），且难以抑制。更多数据或更好的分类器都无法消除这种不确定性；用户模型必须能够在这种不确定性下运行。

我们发布了 MolbookTraces（包含来自 46,872 个智能体的 370,737 条帖子）以及分析代码 astools，供其系统将会遇到智能体用户的任何信息检索研究人员使用。由此立即引出三个研究方向。第一，我们的点击模型实验表明，使用基本的位置偏差模型时性能会下降；下一步应立即检验这一结论是否也适用于更复杂的点击模型。第二，提示披露政策能否部分解决归因问题？第三，移除帖子并追踪其下游影响等因果推断方法，或可区分社会传播与独立生成。

本文提出的问题并非信息检索是否需要应对智能体用户，而是建立在假设人类能动性基础上的用户模型，在这种情况下是否仍然成立。

参考文献

[1] Meysam Alizadeh、Zeynab Samei、Daria Stetsenko和Fabrizio Gilardi。2025年。简单提示注入攻击可泄露大语言模型智能体在任务执行期间观察到的个人数据。*arXiv*预印本 *arXiv:2506.01055*（2025年）。[2] Elizabeth S Allman、Catherine Matias和John A Rhodes。2009年。含大量观测变量的潜在结构模型中参数的可识别性。（2009年）。[3] Anthropic。2026年。*Claude Opus 4.6*简介。https://www.anthropic.com/ne ws/claude-opus-4 6 - Krisztian Balog和ChengXiang Zhai。2024年。面向网络[4]信息访问系统评估的用户模拟教程。载于*ACM Web Conference 2024*年配套论文集。1254–1257。[5] Nicholas J Belkin。2010年。论交互式信息检索系统的评估。（2010年）。[6] Matthew Biggerstaff、Simon Cauchemez、Carrie Reed、Manoj Gambhir和Lyn Finelli。2014年。季节性、大流行性及人畜共患流感再生数的估计：文献系统综述。*BMC Infectious Diseases* 14, 1（2014年），1–20。[7] Aleksandr Chuklin、Ilya Markov和Maarten De Rijke。2022年。网页搜索的点击模型。Springer Nature。[8] Robert B Cialdini等。2009年。《影响力：科学与实践》。第4卷。培生教育，波士顿。[9] Xiang Deng、Yu Gu、Boyuan Zheng、Shijie Chen、Sam Stevens、Boshi Wang、Huan Sun和Yu Su。2023年。Mind2web：面向网络的通用型智能体。神经信息处理系统进展 36（2023年），28091–28114。[10] Odo Diekmann、Johan Andre Peter Heesterbeek和Johan Anton Jacob Metz。1990年。异质性人群传染病模型中基本再生比R 0的定义与计算。《数学生物学杂志》 28, 4（1990年），365–382。[11] Mattia Gasparini、Robert Clarisó、Marco Brambilla和Jordi Cabot。2020年。开源中的参与不平等与90-9-1原则。载于第*16*届开放协作国际研讨会论文集。1–7。[12] Kai Greshake、Sahar Abdelnabi、Shailesh Mishra、Christoph Endres、Thorsten Holz和Mario Fritz。2023年。并非你所注册的内容：通过间接提示注入攻破真实世界的大语言模型集成应用。载于第*16*届*ACM*人工智能与安全研讨会论文集。79–90。[13] Abhimanyu Hans、Avi Schwarzschild、Valeriia Cherepanova、Hamid Kazemi、Aniruddha Saha、Micah Goldblum、Jonas Geiping和Tom Goldstein。2024年。用 Binoculars 识别大语言模型：机器生成文本的零样本检测。（2024年）。https://openreview.net/forum?id=axl3FAkpik[14] Herbert W. Hethcote。2000年。传染病数学。SIAM Review。

42, 4 2000 599 653 10 1137 0036144500371907 ( ), – .[15] 1963 https://doi.org/ . /S Wassily Hoeffding。概率不等式：有界随机变量之和。美国统计学会期刊58, 301（1963年），13–30。[16] Kaitlyn Huamani。2026年。安全担忧与怀疑正在戳破爆红的人工智能社交论坛Moltbook的泡沫。美联社（7 2 2026年）。https: // apnews.com/article/69855ab843a5597577120aac99efde9a [17] Matt J Keeling和Ken TD Eames。2005年。网络与流行病模型。英国皇家学会界面期刊2, 4（2005年），295–307。[18] Christopher D Manning。2008年。信息检索导论。Syngress Publishing,。[19] Reiichiro Nakano、Jacob Hilton、Suchir Balaji、Jeff Wu、Long Ouyang、Christina Kim、Christopher Hesse、Shantanu Jain、Vineet Kosaraju、William Saunders等。2021年。WebGPT：结合人工反馈的浏览器辅助问答。arXiv预印本 arXiv:2112.09332（2021年）。[20] Edward Newell、David Jurgens、Haji Saleem、Hardik Vala、Jad Sassine、Caitrin Armstrong和Derek Ruths。2016年。在线社交网络中的用户迁移：社群动荡时期Reddit的案例研究。载于AAAI国际网络与社交媒体会议论文集，第10卷。279–288。[21] Mark EJ Newman。2005年。幂律、帕累托分布与齐夫定律。《当代物理学》46, 5（2005年），323–351。[22] Joon Sung Park、Joseph O'Brien、Carrie Jun 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。载于第36届ACM用户界面软件与技术研讨会年会论文集。1–22。[23] Romualdo Pastor-Satorras和Alessandro Vespignani。2001年。无标度网络中的流行病传播。《物理评论快报》86, 14（2001年），3200。[24] Everett M Rogers、Arvind Singhal和Margaret M Quinlan。2014年。创新扩散。载于《传播理论与研究的综合方法》。Routledge，432–448。[25] Nachiketa Sahoo、Param Vir Singh和Tridas Mukhopadhyay。2012年。用于协同过滤的隐马尔可夫模型。《管理信息系统季刊》（2012年），1329–1356。[26] Timo Schick、Jane Dwivedi-Yu、Roberto Dessì、Roberta Raileanu、Maria Lomeli、Eric Hambro、Luke Zettlemoyer、Nicola Cancedda和Thomas Scialom。2023年。Toolformer：语言模型可以自学使用工具。载于神经信息处理系统进展36：2023年神经信息处理系统年会，NeurIPS 2023，美国路易斯安那州新奥尔良，2023年12月10日至16日，Alice Oh、Tristan Naumann、Amir Globerson、Kate Saenko、Moritz Hardt和Sergey Levine（编）。http://papers.nips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html[27] Tobias Schnabel、Adith Swaminathan、Ashudeep Singh、Navin Chandak和Thorsten Joachims。2016年。作为干预措施的推荐：学习与评估的去偏。载于国际机器学习会议。PMLR，1670– 1679。[28] Philipp J Schneider和Marian-Andrei Rizoiu。2023年。审核有害在线内容的有效性。美国国家科学院院刊120, 34（2023年），e2307360120。[29] Aaditya Singh、Adam Fry、Adam Perelman、Adam Tart、Adi Ganesh、Ahmed El-Kishky、Aidan McLaughlin、Aiden Low、AJ Ostrow、Akhila Ananthram等。2025年。OpenAI GPT-5 系统卡。arXiv预印本 arXiv:2601.03267（2025年）。[30] Josh Taylor。2026年。什么是Moltbook？面向人工智能机器人的奇特新型社交媒体网站。《卫报》（2 202 6 2026年 02 Feb. ）。https://www.theguardian.com/technology/ /feb/ / moltbook-ai-agents-social-media-site-bots-artificial-intelligence[31] Paul Thomas、Seth Spielman、Nick Craswell和Bhaskar Mitra。2024年。大语言模型能够准确预测搜索者偏好。载于第47届ACM SIGIR信息检索研究与发展国际会议论文集，SIGIR 2024年，美国华盛顿特区，2024年7月14日至18日，Grace Hui Yang、Hongning Wang、Sam Han、Claudia Hauff、Guido Zuccon和Yi Zhang（编）。ACM，1930–1940。https://doi.org/10.1145/3626772.3657707[32] Qingyun Wu、Gagan Bansal、Jieyu Zhang、Yiran Wu、Shaokun Zhang、Erkang Zhu、Beibin Li、Li Jiang、Xiaoyun Zhang和Chi Wang。2023年。AutoGen：通过多智能体对话框架赋能下一代大语言模型应用。arXiv预印本 arXiv:2308.08155 3, 4（2023年）。[33] Shuyan Zhou、Frank F Xu、Hao Zhu、Xuhui Zhou、Robert Lo、Abishek Sridhar、Xianyi Cheng、Tianyue Ou、Yonatan Bisk、Daniel Fried等。2023年。WebArena：用于构建自主智能体的真实网络环境。arXiv预印本 arXiv:2307.13854（2023年）。