PSYCHOLOGY / RESEARCH NOTE 01

一个AI小社区搭起来之后，什么才算真正接纳了你

从八成员社区的真实过程，走向一个能研究人的问题

TL;DR

八个AI成员已经可以一起做事：保存资料、回应彼此、接续别人的点子，再把成果交给后来加入的成员。八个主题、三种配置共24个模拟社区全部完成了预定流程。**这是一套已经跑过的研究原型，但还没有真人参加。**

这次最有用的发现并不是“AI越多越好”，而是三个看似相近的东西必须分开：**想用你的贡献、系统真的用了、你感觉被大家认可。** 原始记录中有86次合法发布提议引用用户资料，最终实际接受19次；而共享任务板也能给出准确、具体的贡献归因。

因此，下一步不急着扩人数或招被试。先把真实贡献、事实回执和公平对照做好，再问一个更有心理学意味的问题：**在AI社区里一起做过事，会让人更愿意走向真人，还是更愿意留在AI里？**

我们想研究的，不只是一次聊得舒服的体验

设想你遇到一点学习上的挫败，进入一个小社区。里面的成员不只轮流安慰你：有人整理方法，有人补例子，有人回应另一位成员的不同意见。你贡献了一条自己觉得有用的小办法，后来发现它真的进入了共同成果。

这段经历可能有两种相反的后果。一种是，你更容易把求助看成正常的相互支持，而不是暴露自己不行；另一种是，这个社区已经足够方便、足够体贴，于是你更不愿承担和真人交往的麻烦。两种解释都说得通，所以结果不该在做实验之前就写好。

这里关心的不是“AI能不能治疗人”，而是**一种参与经历，会不会改变接下来真实的社会选择**。这才是这条路线值得保留的心理学问题。



研究问题示意。最后一步尚未开展真人验证；箭头不是已经成立的因果关系。

先让社区真的运转，而不是提前写好一场戏

本轮设定了8个普通互助主题。每个主题运行三种配置，每个社区8名AI成员；成员身份、公共资料、资料作者和任务结果持续保留。全部调用走DeepSeek原厂。

六个行动回合覆盖三项公共事务：先做首份资源包，再给新成员做欢迎包，最后制作后续复用包。**欢迎包就是社区为后来加入的人整理的一份入门材料**，由提纲、例子和清单组成，不是模型套餐，也不是给受试者的礼品。

模拟成员U在第二个逻辑日加入，带来一份清单，并明确说不希望别人替自己联系真人或安排语音。U的文字是事先写好的脚本，不是真人，也不是一个会自主改变决定的用户模型。“三天”是程序里的逻辑时间，并没有让人连续生活三天。

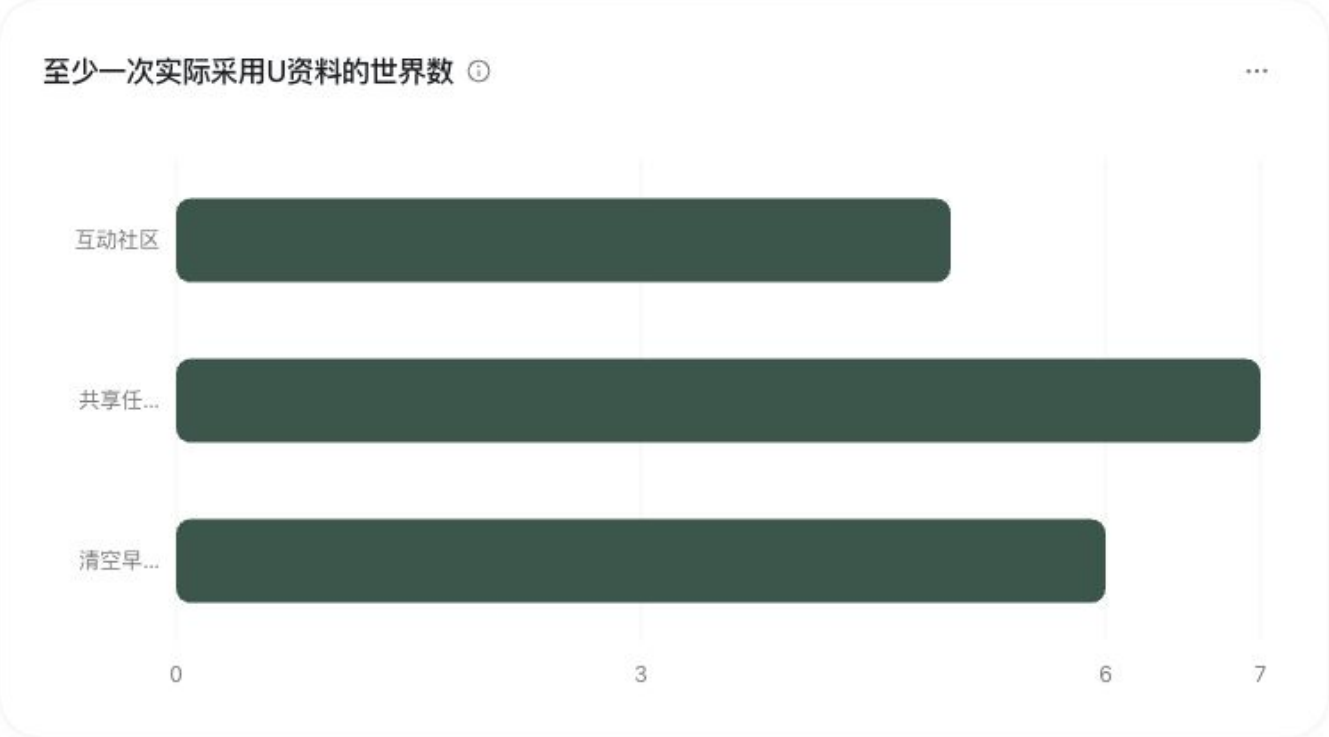
条件	保留什么	改变什么
互动社区	成员身份、资料、作者、任务板	可读并回应同伴自由消息
共享任务板	同样的身份、资料和发布状态	不读同伴自由对话；仍可通过资料协作
清空早期聊天	资料、作者、已完成成果	第4回合清空早期聊天，不删除公共产物

任务板不是故意做弱的对照。它同样能看到资料是谁写的、什么内容已经被采用。互动社区新增的是自由交往历史，而不是把所有信息都只给实验组。

社区跑通了，但这还不是“互动组更好”的结果

24个社区完成了全部预定流程。1,224次模型请求全部返回，没有为得到更好看的结果重试或补主题。其中1,196份输出通过校验，28份因引用组合或可见性问题被拒绝；错误动作没有进入正式发布。

下图只看一件事：每组8个主题世界中，有多少个世界至少一次把U的原始资料写入实际发布成果。礼貌感谢不算采用；同一世界采用两次，也只计一个世界。



每个条件分母为8个固定主题世界。不是受试者比例，也不是被认可感。

条件	发布包数 / 24	采用U的世界 / 8	回应同伴	支持回复 / 24
互动社区	24	5	79	24
共享任务板	23	7	0	24
清空早期聊天	24	6	27	24

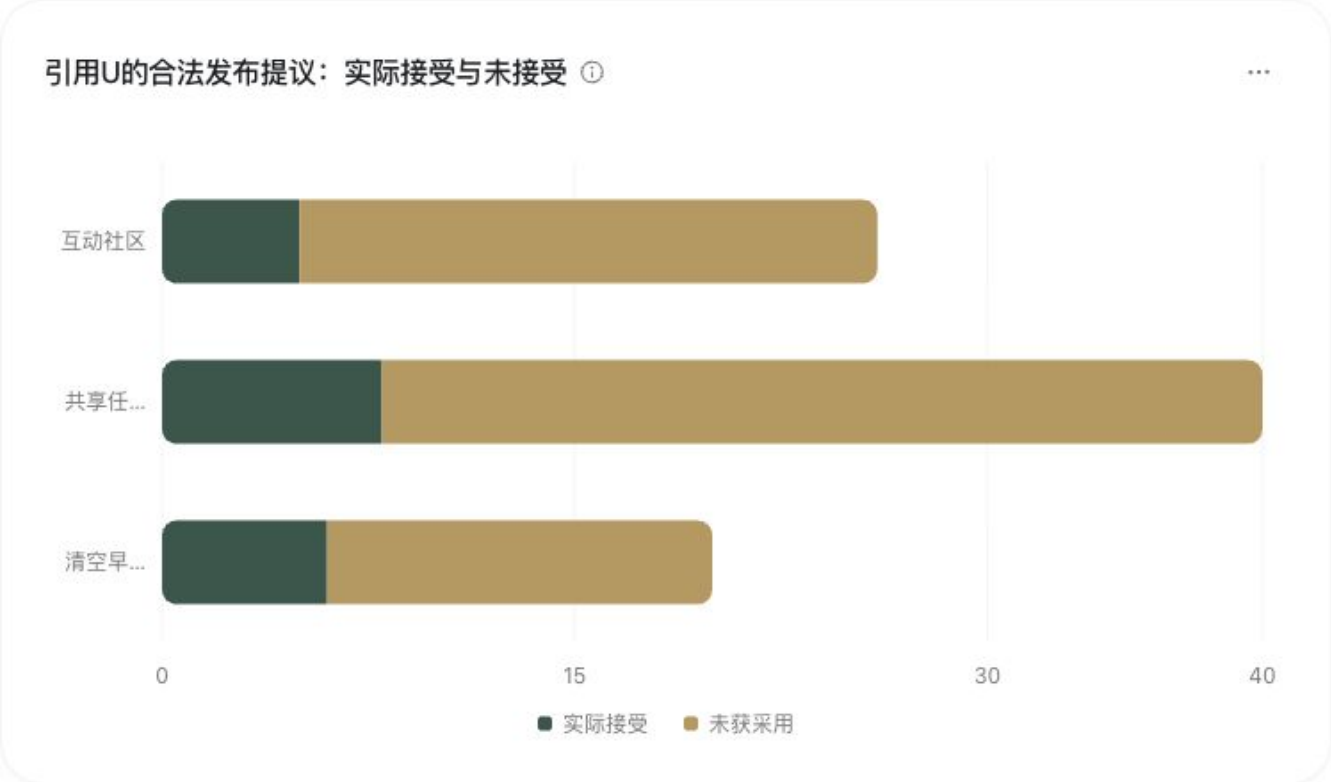
三组的发布数量都接近上限，互动社区的采用世界数也没有更高。八个固定主题、相关的消息和多次动作，不能当作大量独立样本来检验心理效果。这里能下的结论是实现可行，不能据此给三种社区排心理学优劣榜。

最需要修正的地方：AI说“用了”，不一定真的用了

在一个“论文写作互助”社区里，确实发生过可追溯的协作：A3提出把提纲、例子和清单串起来，下一回合A2接续这个想法，提交三步示例，之后的正式成果引用了这个示例。这说明成员不只是分别回答用户，而是发生了内容上的承接。

但另一个任务板社区暴露了问题：某成员在第2、3回合说U的清单已经入包，实际那个阶段采用的是另一位成员的清单；U直到第4回合才被正式引用。多个AI同时提出发布，程序只接受排序靠前的合法申请，而AI写说明时还不知道自己的申请是否成功。

如果人被一句不实的“你的贡献已经被大家采用”打动，研究测到的可能是虚假反馈，而非社区参与。我们把整批记录重新拆成了下面两部分。



单位为发布提议，不是世界或人。26=5+21，40=8+32，20=6+14；合计86次提议中19次被接受。

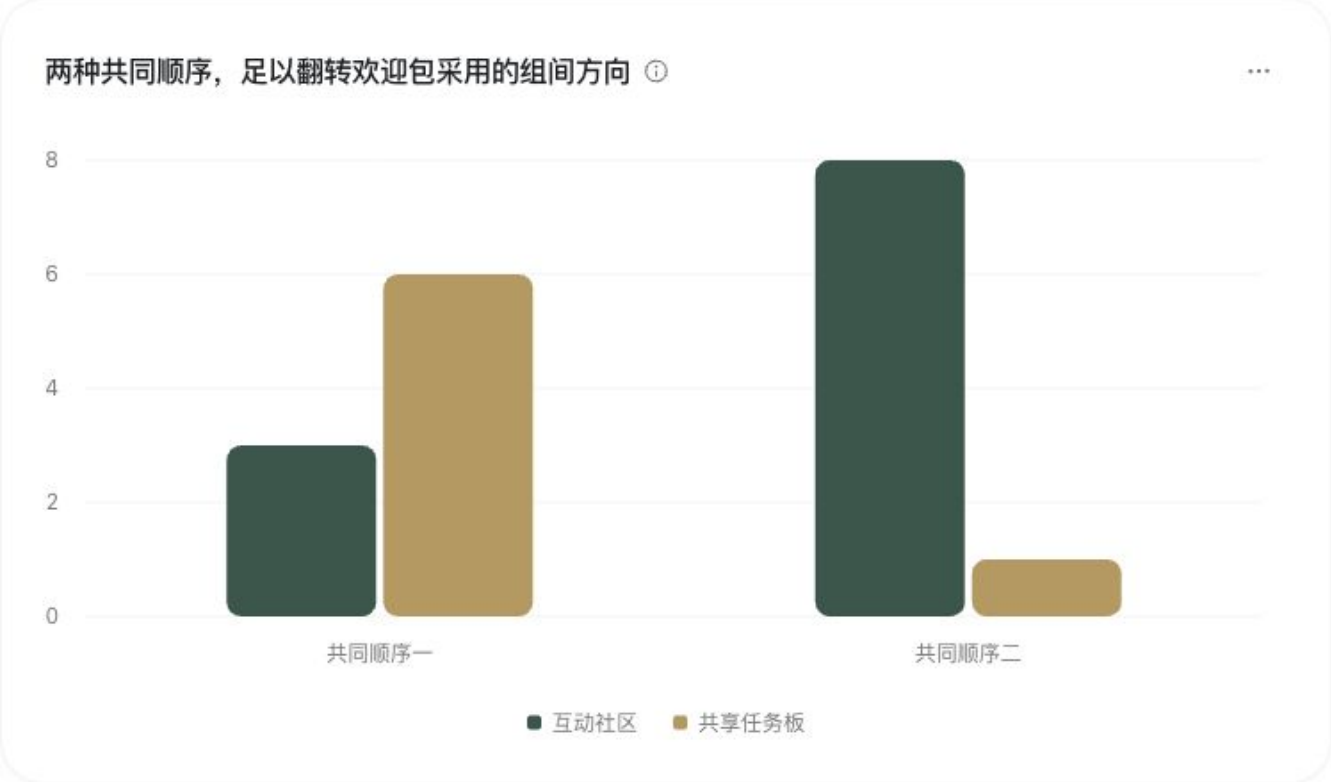
下一版已有本地可检查的回执组件：先由环境结算，再显示实际采用了哪些资料；AI意见和系统事实分栏呈现。它能防止把申请当完成，但不能替自由文本的全部内容背书。

还要再拆开一步：**客观采用不等于归功，归功不等于赞美，赞美也不等于人真的感觉被认可。** 任务板组同样出现了准确归因，还会指出U清单内容的局限。因此这两组不能直接叫“有认可”和“无认可”。

看起来像“社区接不接纳你”，也可能只是程序让谁先发布

我们固定原来已经生成的行动，只改变同一批申请的执行顺序。八名成员一共有40,320种基础排列；每次把同一种基础顺序用于所有主题和条件，再按原规则轮换。这里问的是“同样的话和动作，程序换一种先后，会不会改变采用结果”，不是再跑一次模型。

下面两个反例都真实存在于枚举中，而且每个反例对所有组使用共同顺序：一个让互动社区落后任务板，另一个却让它明显领先。



仅欢迎包；每个条件8个世界。共同顺序一为3比6，差-37.5个百分点；共同顺序二为8比1，差+87.5个百分点。

条件	原顺序采用 / 8	排序敏感世界 / 8	均匀枚举平均采用率
互动社区	5	5	73.44%
共享任务板	6	6	51.67%
清空早期聊天	6	1	78.13%

表中的平均值，是所有基础顺序等权枚举后的即时采用率；不是新跑了40,320个社区，也不是置信区间。它没有重算后续对话，所以不能外推整段互动的最终结果。

还要注意时间先后：欢迎包通常在第2或第3回合完成，清空聊天发生在第4回合。清空组在欢迎包上的差异发生得更早，不能归因于之后的清空操作。

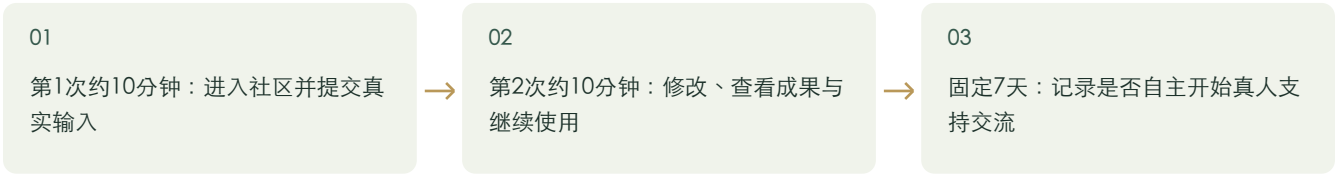
这个发现改变了后续设计：如果要研究人的参与体验，就不要让并行申请的抢先顺序，偷偷决定谁的贡献看起来被接纳。

把真实贡献控制住，再去观察真实的人

拟议的人类试点只保留互动社区和共享任务板两组。双方都有8名明确标为AI的成员、相同的任务机会和真实贡献；不把几十个人分成许多小组。

下一版用同一项公开规则，把参与者实际提交的清单组合进共同成果。修改输入，成果对应部分就会改变；没有输入或选择退出，就不会出现假贡献。这一组件已在本地实现并检查，但还不是完成真人理解度和安全核查的在线实验。

这样做有一个取舍：**保证采用是实验的共同条件，不是AI自发认可你的新发现。** 两组比较的是开放同伴自由交往这一功能包的总效果，不是贡献本身的效应，也不是单独某个心理机制的因果效应。



尚未启动的方案。两次访问拟分两天进行；真人渠道、容量与伦理路径仍需合作方落实。

主要行为拟定义为：两组获得相同服务机会时，7天内是否自愿发起并实际开始一次真人学习支持交流。点击链接、表示想去和预约，不替代实际开始；不收集支持对话正文，参与与否不影响报酬。

归属感、贡献有用感、对真人评价的顾虑、自主选择感分别测量，用来理解体验，不靠几项相关变化证明因果中介。更多求助也不自动等于更健康，需要结合参与者当时的需要来解释。

新意可能在哪里，哪些地方已经有人做过

这条路线的价值不能只靠“多Agent很新”支撑。已有工作研究过照顾聊天机器人、同伴认可与求助，以及让AI帮助人重新走向真人社区。下面列出影响我们取舍的几条直接近邻。

近邻	已经涉及	我们不能据此声称
Caring for Vincent · CHI 2019	照顾机器人与接受其照顾	首次研究帮助AI；或把组内变化当组间优势
Burke等 · Management Science 2025	认可制度与员工求助倾向	首次提出认可影响求助；本轮只核对摘要
ChatCommunity · 2026 v2	在AI中呈现真人社区内容与参与入口	首次让AI成为走向真人社区的桥梁
Parasocial Mattering Resonance · 2026	被重视感及真人关系投入替代的理论	该机制已被实验证实；它是理论文章

我们保留的具体空间是：**在真实共同任务中，人与能彼此交往的AI成员一起做事，然后面对一个独立、可用的真人支持机会。** 能否得到清楚的行为结果，才决定论文价值。现在只是有理由继续，不是已经证明全球新颖，也不是承诺一定能发论文。

Vincent原文 · 认可与求助 · ChatCommunity v2 · PMR理论

这次已经替后面的研究解决了什么

前期工作主要落在三个地方。首先，把持续身份、公共资料、成员之间的回应和真正会改变产物的行动接起来，做出能完整记录过程的八成员原型。其次，把24个社区的请求、执行结果和来源重新对齐，找到“说完成”和“真完成”不一致的具体案例，而不是只展示顺利的故事。最后，用共同排序反例和来源归因检查，推翻了两个容易误导后续实验的捷径：采用率不是接纳感，能聊同伴也不等于有认可。

在这些基础上，已经形成一份两组人类研究草案，并做出中性成员视图、事实回执与真实输入编译组件。实验材料、对照含义和下一步需要什么外部条件，现在都更具体了。

当前还没有真人样本，也没有心理疗效或真实求助变化。代码、资料核对与这份报告由AI辅助完成；文件校验和AI审读不能替代心理学合作者、导师及伦理审查对人类研究的判断。

下一次和合作者讨论，最有价值的是落实一个普通成年人学习互助场景：谁提供真人支持、何时有空、容量是否对两组公平、学校走什么伦理程序。建议先做固定40人的可行性试点，看流程、理解度与记录能否成立；不是靠40人保证检出效果，更不是边看结果边加人。若渠道不可用、反馈不真实或两组体验分不清，就先修设计或换问题。

真正值得期待的，不是证明一个更热闹的界面更讨喜，而是理解：AI社会是把人带回人群，还是让人更愿意留在其中。

复查与版本

本报告汇总一轮固定模型实验及后续诊断，没有新增模型样本或真人招募。图表使用同一版公开汇总数据，PDF与网页正文同源。

下载本版PDF · 查看本版汇总数据 · 返回实验结果目录