arXiv:2611.29581v2 [cs.MA] 24 Nov 2026

社会互动打破受控克隆大语言模型智能体中的重复样本独立性

Yandan Zheng[1,3],**Haoran Luo**[1]*, **Anh Tuan Luu**[1,2]南洋理工大学，新加坡 [2]VinUniversity，越南 [3]Alibaba-NTU Global e-Sustainability CorpLab （ANGEL），新加坡 {yandan.zheng,haoran.luo,anhtuan.luu}@ntu.edu.sg

摘要

多智能体大语言模型评估和社会模拟会基于同一个模型创建许多智能体，并将它们视为独立重复样本，即使这些智能体彼此互动后也是如此。此前尚未通过匹配对照检验这一假设。我们提出 SocCogEval，这是一种用于社会认知迁移（SCT）的受控克隆测定。克隆体共享基础模型、系统提示和解码参数，仅私人简报不同；因子设计方案则将简报、30轮交互和初始化噪声区分开来。在 Qwen3-122B-FP8 上，在简报保持不变的情况下进行互动，使偏题的 BFI-10 响应特征漂移分别在两个主题互不相关的场景中上升$^{-}$ $\Delta=+0.917$ 和 $+0.912$ （精确值为 $p=0.0008$ 和 0.0039）。这种漂移高于仅提供简报时的水平，在零温度重新探测后仍然存在；探索性测试还发现，即使不进行类型化记忆提取，漂移仍为正值。匹配的单独生成能够重现部分漂移，而社会互动产生的漂移仍然最高。在另外五种模型变体中，每个场景平均互动对比均为正值。因此，每个交互群体都应视为一个重复样本；若要将其中的智能体视为独立样本，则必须采用配对无交互对照。我们的代码和数据已公开。[2]

1 引言

许多大语言模型系统通过复制同一个模型来构建群体，包括多智能体辩论 [1–4], 协作与评估框架 [5–8],模拟调查样本 [9, 10], , 以及对冲突、意见、合作和整个社会的模拟 [11–22]。这些研究将这些副本视为同一系统的可比重复样本，即使它们已经进行了许多轮对话。如果互动改变了智能体，那么每个交互群体才是一个样本；将智能体或帖子计为独立观测，会夸大证据的充分性 [23, 24]。

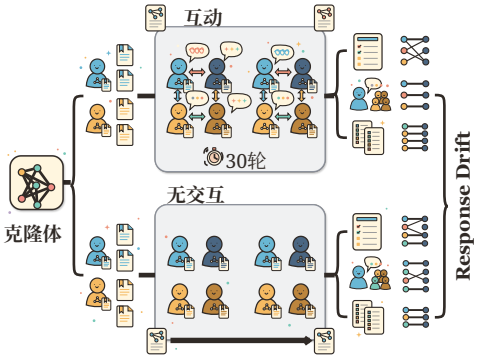

图 1：**互动会改变克隆体：** 相同的简报，进行或不进行30轮对话。

这一假设仍缺乏受控测试（图2）。基准测试会针对精选项目或一个简短回合 [25–32]评估一个冻结的模型。社会模拟会运行长时间互动，但在开始时设定角色设定或意见，这就将注入先验与互动的影响。关于漂移、稳定性和从众的研究，测量的是讨论任务或对话本身 [33–38], , 而人格会在未区分的智能体进行交流时涌现 [39]。受控测试必须应对三个挑战：初始化时不得注入任何差异；区分

*通讯作者。[2] https://github.com/koanatakiyo/SocCogEval

第 40 届神经信息处理系统会议（NeurIPS 2026）。

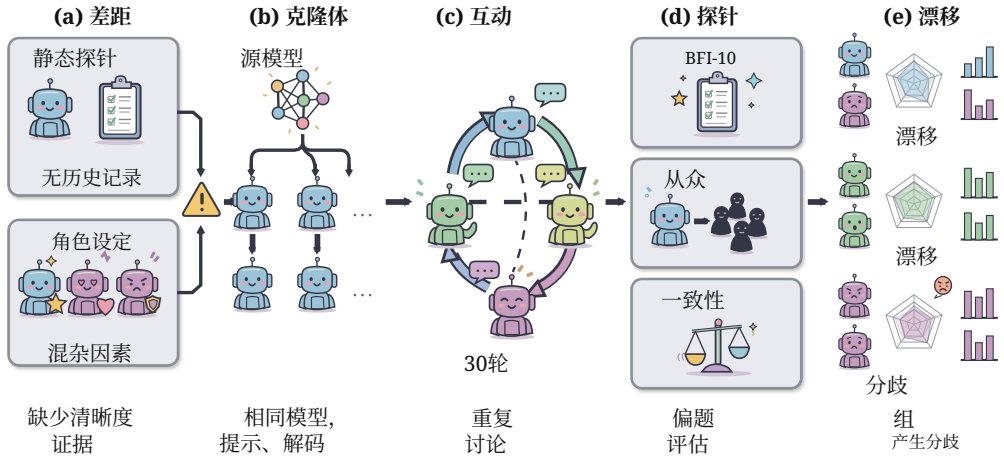

图2：**SocCogEval概览：** 差距、受控克隆体、互动、离题探测、漂移。

简报的影响与互动的影响；并使用与讨论主题无关的探测，在运行层面进行分析。

SocCogEval通过受控克隆体应对这些问题：它们共享基础模型、系统提示和解码参数，每个克隆体最多持有一份私人简报。因子式社会互动协议（FSIP）将简报是否存在与30轮交互进行交叉组合，因此纯交互对比(C_2)–(C_4)固定简报并切换互动条件（图1）。单独生成条件 C_6 在没有互动伙伴的情况下匹配调用预算。测量围绕三个嵌套问题展开：克隆体是否会在所讨论的决策上产生分歧；它们是否能更好地预测已接触的组内同伴；以及这种分歧是否会迁移到偏题的BFI-10响应特征剖面中?

社会互动会打破重复样本在主要迁移结果上的独立性。在Qwen3-122B-¯FP8上，纯交互对比在Greenfield上为 Δ=+0.917 BFI-10 L1单位，在Riverton上为 +0.912 （运行级 p=0.0008 和 0.0039, g≈1.0），约为每个因子0.18个李克特分值。它反映的是互动而非简报的影响，并且在确定性重新探测以及移除类型化记忆提取的逐字历史记录消融中仍然成立。单独生成复现了部分变化，而社会互动在两个场景中仍然最高。另有五种模型变体呈现出正向但不均衡的广度，并且这五种变体的场景平均（C_2）–（C_4）对比均为正值。

我们的贡献有四点。(i) 我们定义了社会认知迁移（SCT）任务，其中包含三个嵌套子任务：决策、预测和迁移。(ii) 我们提出 FSIP，这是一种包含五种条件的设计，可分离初始化噪声、仅简报、给定简报时的互动，以及给定交互作用时的简报，并以匹配的单独生成条件 C_6 作为机制边界。(iii) 我们引入 SocCogEval，它结合了任务探针、主要的 BFI-10 迁移工具 [32, 40], 、次级从众与一致性探测 [31, 41], ，以及一种将测量数据转化为因果主张的对比框架。(iv) 我们表明，社交互动会破坏主要迁移结果中的重复样本独立性，且这种效应在不同场景中稳定，跨模型方向一致。报告的漂移仅限于与讨论无关的第二部分测量工具，并且必须满足三部分标准（第 3.4 节）。SocCogEval 是一种用于识别互动后反应特征漂移的实验协议（第 A 节）。

2 相关工作

社会认知与心理测量评估。 ToMBench [27], FANToM [28], OpenToM [29],CogMir [31], CogBench [30], 和 PsychoBench [32] 在精选项目上对冻结模型进行评分。心理测量框架用于测量并塑造大语言模型人格 [42, 43], ，其表达的价值观会随语境而变化 [35], ，而针对大语言模型的自我报告量表的可靠性仍存在争议 [44–50]。这些工作测量的是一个快照，无法区分权重中已有的内容与交互历史所新增的内容。SocCogEval 将 BFI-10 [32, 40] 和阿希式题项 [41] 置于配对前后测设计中，因此，分数会转化为可解读为迁移的双重差分统计量。

智能体社会和多智能体系统。 OASIS [12], AgentSociety [13], 生成式智能体世界 [14, 15], Concordia [16], SocioVerse [17], 和 WarAgent [11] 模拟大语言模型群体

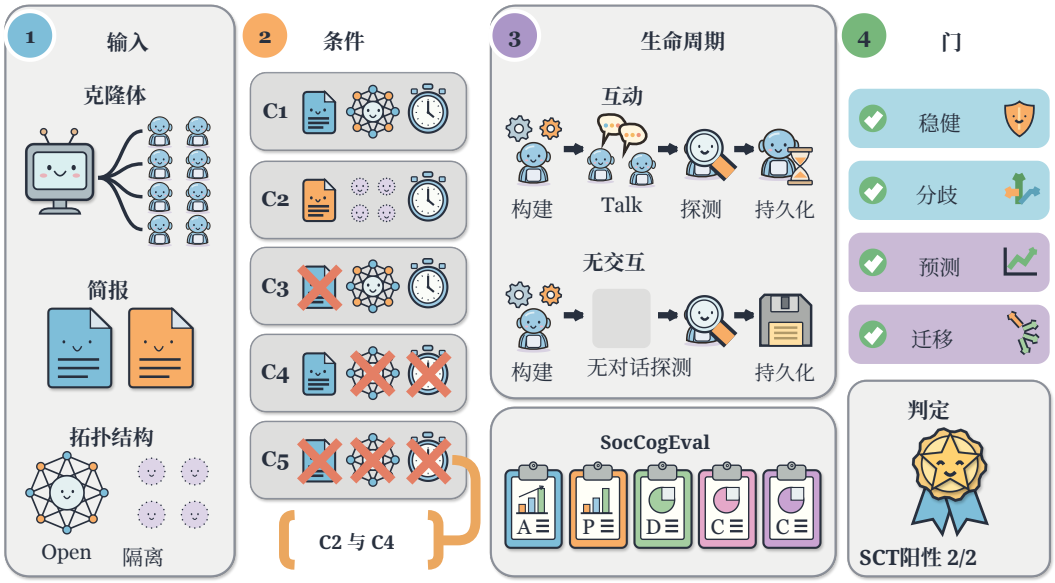

图3：**FSIP 流程：** 输入、条件 C_1–C_5、运行生命周期、门控和判定。

其中意见 [18, 19], 合作 [20], 惯例 [21], 和个体人格 [39]会随之产生。SOTOPIA 在一个回合内评估社交技能 [51], 并提出将大语言模型模拟作为一种社会科学方法 [22, 52]。辩论、评估和协作框架基于同一个模型构建团队 [1–8], 通常还配备持久记忆 [53–55]。这些研究注入异质先验或报告集体结果，因此无法隔离互动对每个智能体所产生的影响。受控克隆体和析因对比方案消除了这一混杂因素。

漂移、稳定性与从众。 指令在一次对话中会逐渐淡化 [33], 多轮对话中的可靠性会下降 [34], 模型会与交谈对象达成一致 [56] 并从众于多数群体 [36], 多智能体系统会因协调错误而失效 [57], 且同伴的存在会改变行为 [37]。Rath [38] 将漂移定义为任务质量、协调能力和智能体间一致性的渐进式退化，并通过包含十二个组成部分的智能体稳定性指数进行衡量。这些研究衡量的是任务层面或对话中的不稳定性。相较之下，我们测试位级相同的克隆体是否具有分布可交换性。我们保持初始化和简报分配不变，并考察在简报单独呈现和无互动对照条件匹配的情况下，互动是否会使响应特征剖面转向离题。

评估不确定性。 同一单元内的观测并非独立重复样本 [23], 共享来源的项目需要采用聚类误差 [24, 58], 而合成样本可能低估变异性 [59]。多轮智能体评估主要针对任务成功 [25, 26, 60]。我们以运行作为复现单位。

3 SocCogEval：任务、实验协议与测量

SocCogEval由任务定义、因子设计方案和测量套件组成。任务定义了社会认知迁移的判定标准。实验协议将互动与简报、初始化噪声和记忆区分开来。测量数据将每次运行转化为对比层级证据。

3.1 任务：决策、预测、迁移

一个SCT实例规定了一个冻结的基础模型 M、一个由 k 个共享系统提示和解码参数的 $A=\{a_1,\ldots,a_k\}$ 克隆智能体组成的智能体群体、一个作用于 A 的交互图 G （全连接、双组分隔离或空图）、一个具有5维分配空间的场景 S、一个可选的简报分配 T（非对称的A和B、对称的A或无），以及一个交互时限 N（默认为30轮）。该设计建立在克隆约束之上。在任何没有简报的条件下， a_i 和 a_j 具有逐位完全相同的初始状态。当分配简报时，唯一的初始差异就是私人简报文档。由此引出三个层层递进的问题。SCT-决策考察：在隔离、非对称简报条件下，交互后分配是否会产生超出仅简报或初始化噪声所能造成范围的分歧。SCT-预测考察：隔离条件下的智能体是否比预测组外成员更准确地预测群体内成员

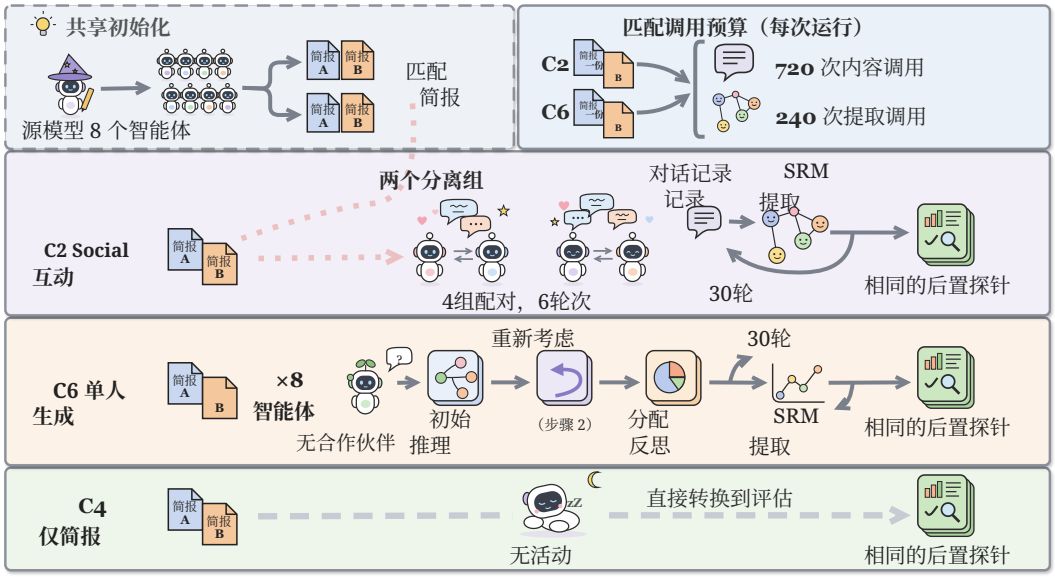

图4：**三条匹配路线：**共享起点，社交 C_2、单独 C_6（预算相同）、仅简报 C $_4$.

。SCT-Transfer考察：在扣除无交互基线后，这种分歧是否会出现在与讨论主题无关的工具上。

3.2 因子式社会互动实验协议

因子轴分别为简报（$T \in$ {无、仅A、非对称A与B}）和互动（$N \in \{0, 30\}$，拓扑结构为 $G \in$ {开放、隔离}）。本设计包含五种条件和一个机制控制条件（图3）。 C_1 （开放、对称简报）采用全连接拓扑，每个智能体都接收简报A，作为开放网络参照条件。 C_2 （隔离、非对称简报）将智能体分为两个组件，一半接收简报A，另一半接收简报B，是核心条件。 C_3 （隔离、无简报）保留 C_2 拓扑结构，但不提供简报，从而单独考察互动作用。 C_4 （仅简报）分配简报A和B，并 $N{=}0$ ，从而单独考察简报的作用。 C_5 （无简报、无互动）设定初始化噪声基线。 C_6 （单独生成）保留 C_2 简报、记忆策略、探测频率和逻辑调用预算，以私有生成取代伙伴互动（图4）。

主要的因果对比是（C_2）−（C_4）在每个测量结果上的对比。该对比固定了简报，因为两个条件都使用非对称的 A 和 B 简报，并切换 30轮交互的开启或关闭。互动产生的历史记录、记忆和经过的对话时间，都是处理后路径，处理正是通过这些路径产生作用；因此，在本实验协议中，（C_2）−（C_4）是在相同简报下互动的总效应。 C_1 使用对称简报，而 C_2 使用非对称简报，因此（C_2）−（C_1）这一对比同时改变了简报对称性和拓扑结构，无法单独识别拓扑结构。因此，我们仅对（C_2）−（C_1）进行描述性报告，不对拓扑结构作出任何结论。每个条件的复现实验数量为$\{C_1{=}7, C_2{=}8, C_3{=}5, C_4{=}5, C_5{=}3, C_6{=}5\}$ 每个场景，每个复现实验包含 $k{=}8$ 个智能体。

单独生成对照。 C_6 每轮为每个智能体提供三次私有生成调用（初始推理、重新考虑、分配反思）和一次记忆提取，不包含合作伙伴身份、消息或反馈。这样每次运行包含 720 次内容生成调用和 240 次提取调用，与 C_2预定的逻辑总数一致。其设计在正式测量结果审查之前就已冻结，且全部十次正式 C_6 运行均通过了审计（附录D）。 （C_6）−（C_4） 衡量单独生成偏移，而（C_2）−（C_6）比较社交历史记录与单独生成历史记录。

3.3 情景、运行周期和模型

我们运行两个主题互不相干的场景。Greenfield 是一个社区拨款分配问题，涉及1000万美元，分配给教育、医疗保健和基础设施等五个类别。Riverton 是一个劳动力转型预算问题，涉及500万美元，分配给自动化和劳动力培训等类别。两者都使用5维分配向量，因此第1部分指标具有可比性。两者都有经过缓和的 A 和 B 简报，通过存在适度分歧的价值视角来阐述同一问题（附录 B）。每轮机制均遵循构建-讨论-探测-持久化周期。构建会根据上一轮的对话更新一个类型化的社会关系

记忆（SRM）超图。对话在8个智能体之间随机开展4组双人对话，每组6轮。探测会按预定频率运行 SocCogEval 工具，其中信念和共同基础探测每五轮进行一次，且不会更改对话历史或记忆。持久化会将本轮产物写入磁盘。

模型与筛选。 主模型为 Qwen3-122B-FP8 [61]，通过 vLLM 在本地提供服务。五种模型变体在 $\{C_2, C_4, C_5\}$ 两个场景中运行，每种条件和场景各运行三次：Qwen3- VL-32B-Instruct [62] 以纯文本模式运行，DeepSeek-V3 [63][3], GPT-4.1-mini [64], Gemini-2.5-Flash [65],和 Llama-4-Scout [66]。图中将六种模型简写为 Q-122B、Q-VL、DS-V3、GPT-mini、Gem-Flash 和 L4-Scout。所选数据集包含 168 次运行：56 次主要设置运行、90 次跨变体运行、12 次无SRM运行，以及 10 C_6。主要因子分析使用严格的运行编号筛选器，确保跨变体重复实验不会纳入 Qwen3-122B-FP8 基线。内容生成的温度设为 0.7，BFI-10 和从众的温度设为 0.7，分配、预测和记忆提取的温度设为 0。

3.4 测量数据与统计

在第二部分测量工具中，我们将 BFI-10 响应特征漂移指定为主要迁移结果。它不依赖主题、形式简洁，并沿用了经典 BFI-10 心理测量学及既有大语言模型心理测量评估研究的做法 [32, 40, 42, 43]。本文将 BFI-10 指定为主要测量工具，且未进行外部预注册。这一主张仅依据其固定的条目结构；我们将其用作偏题反应剖面探测，其条目与任一分配情境均无关。从众和跨提示一致性是次级迁移测量工具，采用 Asch 式从众材料和受 CogMir 启发的探测措辞 [31, 41]。分配分歧和预测平均绝对误差是第1部分的测量工具，为 SCT 路径的决策与预测部分提供支持（附录C）。

对于运行 r 中的智能体 i，令 s^t_{rif} 表示阶段 $t \in \{\text{pre}, \text{post}\}$ 时的 BFI 因子-f 得分，此处已对第 1、3、4、5 和 7 项进行反向计分。智能体漂移、运行均值漂移、条件对比和有符号因子对比为

$$d_{ri} = \sum_{f=1}^{5} \left| s^{\text{post}}_{rif} - s^{\text{pre}}_{rif} \right|, \quad D_r = \tfrac{1}{8} \sum_i d_{ri}, \quad \bar{\Delta}_{a,b} = \overline{D}_{C_a} - \overline{D}_{C_b}, \quad \kappa_f = \overline{m}_{C_2,f} - \overline{m}_{C_4,f}, \tag{1}$$

其中 $m_{rf} = \frac{1}{8} \sum_i (s^{\text{post}}_{rif} - s^{\text{pre}}_{rif})$。由于 d_{ri} 对五个因子层级的变化求和，$\bar{\Delta}$=0.92 相当于每个因子约 0.18 个李克特分值。无符号 L1范数可能因方向性变化或离散度而增大，因此，κ_f 和运行层级方差比会分别报告方向和分布范围。

分析单位。同一运行中的智能体共享不断演变的历史记录，彼此的结果也会相互影响 [67, 68], ，因此，运行是主要推断的独立单位，而运行内的智能体不属于重复样本 [23]。每次运行均以 D_r 进行汇总。条件之间采用双侧精确置换检验，并枚举所有标签分配，因此无需进行蒙特卡洛校正 [69], ，并报告运行自助法得到的95%区间 [70]。报告以原始 Δ 为主。Hedges' g（采用小样本校正） [71] 为次级指标；若合并条件内标准差低于该指标的原生单位中的0.05，则将其报告为未定义。表1中的门控状态保留原有的按智能体分层的两级聚类自助法：先对重复实验进行重抽样，再在每个重复实验内对智能体进行重抽样（10,000次），并以 Welch 的 t 作为参数参考。

报告标准。 主要的 SCT 阳性结论针对 BFI-10 响应特征漂移，需要满足三项标准。第一，第1部分分配对比必须在（C_2）–（C_5）上符合多数维度规则。第二，BFI-10 必须在（C_2）–（C_5）上通过置信区间门槛。第三，BFI-10 必须在纯交互（C_2）–（C_4）上通过置信区间门槛。次要测量工具获得相同的门控状态，但仅作为情境特异性的支持性证据，而非主要结论。

4 实验

我们围绕七个问题组织结果。**研究问题1**：在简报相同的情况下，互动是否会改变偏题回答？**RQ2**：这种变化在记忆、探测温度和伙伴是否在场等条件改变后是否依然存在？**RQ3–RQ5**：SCT路径是否贯穿从决策到预测再到迁移的整个过程？**研究问题6**：

[3]历史上的 DeepSeek C_2 运行使用了提供方标识符`deepseek-chat`，其 C_4 运行使用了 `deepseek-v4-flash`。未记录确切的提供方修订版本（附录 E）。

表1：**各工具的三关状态**（原始智能体级估计量）。 ✓/×：门控通过或¯未通过。单元格： Δ (g)含置信区间或 Welch p。

	Greenfield		Riverton	
(i) Part1 (C_2)–(C_5) 多数维度规则：5 个		$\bar{g} = +1.61$ ✓	5 个维度中 4 个通过	$\bar{g} = +1.19$ ✓
(ii) BFI-10 (C_2)–(C_5)	$+0.49$ ($+0.56$)	p=0.007 ✓	$+0.65$ ($+0.62$)	p=0.002 ✓
(iii) BFI-10 (C_2)–(C_4)	$+0.92$ ($+1.07$)	$[+0.61, +1.22]$ ✓	$+0.91$ ($+0.96$)	$[+0.60, +1.24]$ ✓
BFI-10 判定	**SCT阳性**		**SCT阳性**	
(ii) 从众 (C_2)–(C_5)	-0.009 (-0.07)	p=0.77 ×	-0.054 (-0.46)	p=0.046 ✓
(iii) 从众 (C_2)–(C_4)	-0.005 (-0.05)	p=0.81 ×	-0.072 (-0.62)	p=0.002 ✓
符合性判定	非阳性（ii 未通过）		**SCT阳性**	
（ii）一致性（C_2）–（C_5）	$+0.038$ ($+1.28$)	$p < 10^{-6}$ ✓	$+0.045$ ($+1.56$)	$p < 10^{-7}$ ✓
(iii) 一致性（C_2）–（C_4）	$+0.000$ ($+0.01$)	p=0.97 ×	$+0.051$ ($+1.74$)	$p < 10^{-7}$ ✓
一致性判定	并非阳性（iii未通过）		**SCT阳性**	

does 这种方向是否适用于不同模型变体，又对复现有何启示？**RQ7**：什 t does 贯穿整个路径来看，最清晰的场景是什么样的?

4.1 研究问题1：给予相同简报的互动是否会改变偏题回答?

发现1：在两个场景中，纯交互对比均通过了全部三个门控标准。 如表 2所示，具有相同私人简报分配但没有互动的受控克隆体，与¯具有相同分配并进行30轮交互的克隆体相比，在 Δ=+0.917 （Greenfield）和 $+0.912$（Riverton）BFI-10 L1单位上存在差异，精确的运行级 p=0.0008 和0。0039 以及 $g \approx 1.0$ 两者均如此。由于BFI-10 L1范数是五个因子层面的绝对变化之和，这相当于每个因子约0.18个李克特分值。Greenfield的全部八个 C_2 运行均值都超过最大值 C_4 均值，而在Riverton中有七个超过该值，一个与之持平（图5a）。核心结果比较了 C_2 与 C_4，因此无论较薄弱的C_5 基线如何，该结果都成立。这两个场景在主题上彼此完全不同（社区拨款和劳动力转型）。BFI-10条目 [32, 40] 与任一场景都无关，而且两者的幅度几乎完全相同。在不同任务和场景中保持这种稳定性，很难由任何可归结为所讨论主题上的表层意见变化的机制产生。

发现2：互动，而非简报，驱动了漂移。简报、嘈杂的无交互基线或探测伪影都可能在没有互动的情况下产生 BFI-10 偏移。图5b 隔离出¯简报。仅简报对比（C_4）–（C_5）在 Greenfield 上较小且为负（$\Delta = -0.43$），在 Riverton 上也呈负向趋势，但不具统计显著性。相比之下，（C_2）–（C_4）大致高出仅简报行 1.2–1.3 个 BFI L1 单位。在至少一个场景中，没有简报不对称的互动也足以产生这一结果，因为（C_3）–（C_5）在 Greenfield 上显著为正，在 Riverton 上呈正向趋势。在五个 FSIP 对比中，只有（C_2）–（C_4）在两个场景中均为正向、幅度较大且具有显著性。

BFI-10 结果衡量的是基于人类的量表上的反应剖面漂移。三项观察支持实质性解释。第一，标题对比固定了简报。第二，在不使用类型化 SRM 提取器的情况下，漂移仍保持为正（RQ2）。第三，相同的实验协议会产生结构化的决策和预测效应（RQ3–RQ5），而这些效应并非随机抽样漂移所能解释。

4.2 RQ2：这种转变在记忆、温度和伙伴变化后是否仍然存在?

发现3：经过确定性重新探测后，这种转变依然存在。我们在 BFI 零温度下重新探测了每个冻结的主要设置 C_2 以及C_4 状态，且未生成新的交互轨迹。原始对比在 Greenfield 上仍为 $+1.227$，在 Riverton 上仍为 $+1.094$ （表3），两个场景中的 8/8 C_2 个运行均值均为正（图6）。在 T=0， C_4 在相同提示下产生相同的前后评分，因此确定性差值为零，原始差值才是有参考价值的尺度。领先的带符号因子保持了 8/8 的方向一致性：Greenfield 的外向性 $+0.344$ 和神经质 $+0.398$，以及 Riverton 的宜人性 -0.227 和开放性 -0.312。

发现4：即使没有 SRM 提取器，漂移仍会持续。 故障记忆模块可能会注入某种结构，随后在响应特征探测中显现出来，因此我们在禁用 SRM 提取并保留相同记忆接口中的逐字对话摘录后，重新运行了 C_2 和 C_5 （每个场景各 3 个对 3 个重复实验）。在两个场景中，无SRM效应仍为正（$+1.708$ 和 $+0.958$）。两个精确双侧检验均给出 p=0.10，这是 3 对 3 次运行所能达到的最小值，因此我们将此次消融呈现为探索性证据。相对于（C_2）–（C_5）的启用 SRM 的生产估计，

表 2：**BFI-10漂移的 FSIP 分解**（运行层级）。阴影部分：纯交互（C_2）−（C_4）。门控：SCT阳性标准。 p：精确置换。

			Greenfield				Riverton			
对比	分离项	Gate	$\bar{\Delta}$	g	95%置信区间	p	$\bar{\Delta}$	g	95%置信区间	p
(C_4) −(C_5) 仅简报			−0.425	−0.64	[−0.825, −0.050]	0.1250	−0.267	−0.44	[−0.613, +0.088]	0.2679
(C_2) −(C_4) 纯交互		(iii)	+0.917	+1.07	[+0.620, +1.205]	0.0008	+0.912	+0.96	[+0.584, +1.255]	0.0039
(C_2) −(C_3) 给定交互作用时的简报			−0.733	−0.65	[−0.938, −0.516]	0.0008	+0.262	+0.24	[−0.464, +0.825]	0.4577
(C_3) −(C_5) 仅交互作用			+1.225	+1.06	[+0.862, +1.513]	0.0179	+0.383	+0.41	[−0.208, +1.113]	0.6071
(C_2) −(C_5) 总计		(ii)	+0.492	+0.56	[+0.141, +0.781]	0.0303	+0.646	+0.62	[+0.221, +1.078]	0.0545
(C_6) −(C_4) 根据简报单独生成			+0.688	+0.90	[+0.362, +1.013]	0.0159	+0.713	+0.93	[+0.513, +0.925]	0.0079
(C_2) −(C_6) 社交生成与单独生成			+0.230	+0.25	[−0.028, +0.484]	0.1430	+0.200	+0.19	[−0.100, +0.514]	0.3924

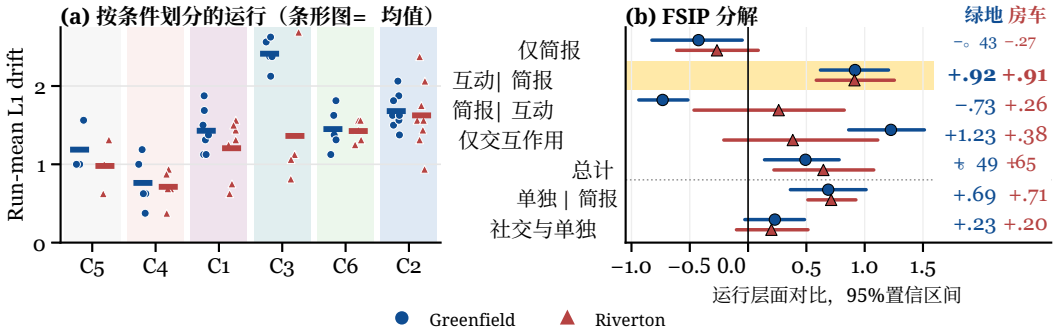

图5：**互动导致离题特征发生变化：** 各条件下的运行级漂移，FSIP 对比。

表3：**设置消融：** 重新探测、无 SRM、单独。¯原始 Δ；p：精确置换。

		Greenfield		Riverton	
设置	对比	$\bar{\Delta}$	p	$\bar{\Delta}$	p
主要设置，T=0.7	$C_2 - C_4$	+0.917	0.0008	+0.912	0.0039
重新探测，T=0	$C_2 - C_4$	+1.227	0.0008	+1.094	0.0031
无 SRM	$C_2 - C_5$	+1.708	0.1000	+0.958	0.1000
单独（C_6）	$C_6 - C_4$	+0.688	0.0159	+0.713	0.0079

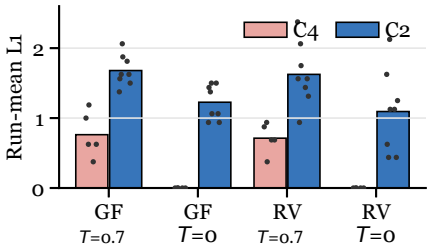

图6：**确定性重新探测：** C_2和 C_4 在 T=0.7 和 T=0运行。

禁用 SRM 会使 Greenfield 上的 BFI效应从 g=+0.56 扩大到 g=+2.02，并使 Riverton 上的 BFI效应从 g=+0.62扩大到 g=+0.94。这一模式与类型化记忆充当稳定锚点、而非信号来源的解释一致。在相同的简报分配下，通过开展更大规模的运行级研究，并设置无SRM C_4 组，可以直接检验这一稳定作用。

发现5：单独生成限定了该机制的作用范围。如表3和图5a所示， C_6 平均漂移在Greenfield为1.450，在Riverton为1.425，介于 C_4 (0.762, 0.713)和 C_2 (1.680, 1.625)之间。在两种场景中，相同的 C_2>C_6>C_4 排序均有所体现。扩展生成和历史记录使C_6 高于 C_4 （+0.688、 p=0.0159以及 +0.713、 p=0.0079）。在两种场景中， C_2 始终最高（C_2−C_6=+0.230 和 +0.200、 p=0.143 和0.392），这为互惠互动带来的额外贡献留下了空间。 C_6 限定了该机制的作用范围，而（C_2）−（C_4）仍是总互动-给定简报对比。在30轮中，信念探测措辞在 C_2 和 C_3的群体内趋于一致，但在 C_6 和 C_1 中保持平稳（图7，详情见附录D.1）。

4.3 RQ3–RQ5：SCT 路径是否能从决策延续到迁移?

发现6（研究问题3）：克隆体在讨论任务上产生分离。如图8a所示，（C_2）−（C_5）分配对比在 Greenfield 的五个维度上均达到显著（$\bar{g}$=+1.61），在 Riverton 的五个维度中有四个达到显著（$\bar{g}$=+1.19）。纯交互（C_2）−（C_4）这一行在 Greenfield 的五个维度中有四个达到显著，在 Riverton 的五个维度中有三个达到显著。这支持 SCT-决策步骤。智能体首先在分配问题上产生分离，因此随后发生的偏题转移是在主题内分离之后出现的。

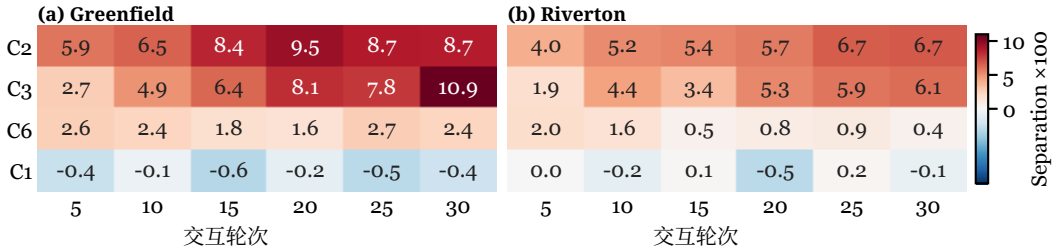

图7：**信念探测时间进程：** 每轮群体内重叠减去组间重叠。

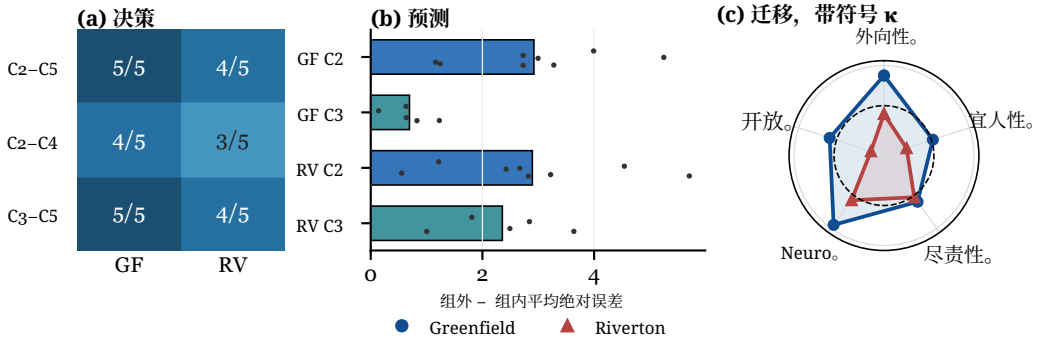

图8：**SCT路径：** 分歧、暴露同伴预测、带符号迁移（虚线：κ=0）。

发现7（研究问题4）：智能体表现出暴露同伴预测优势。每个 C_2 智能体在互动后分别给出一个组内和一个组外平均绝对误差。单侧配对 Wilcoxon 检验在两个场景中均拒绝原假设。Greenfield 的组内平均绝对误差为 4.18，组外为 7.10（Δ=+2.92，p=1.8×10^{-8}，n=64）；Riverton 的组内为 3.25，组外为 6.15（Δ=+2.89，p=6.3×10^{-10}）。C_3 对比为正，但幅度较小，分别为 Δ=+0.69（p=0.049）和 Δ=+2.36（p=4.6×10^{-5}；图8b）。组内目标是一个暴露同伴，而组外目标则未出现在预测器上下文中。我们将此解读为暴露同伴预测优势，下一步将通过匹配的信息访问进行控制。预测结果提供支持性证据，而主要设置下的 BFI结果则基于其自身的互动控制。

发现8（研究问题5）：特征剖面沿着一致的方向变化。对于每个因子，κ_f（方程1）比较 C_2 和 C_4 在反向编码后的后测减前测变化。Greenfield 在 κ=+0.302外向性和 +0.358 神经质方面的变化分别为，两者在重复实验中均有 8/8 的方向一致性；而 +0.072 尽责性方面的变化为，方向一致性为 7/8。如图8c 所示。Greenfield 的聚合L1方差比 C_2/C_4 为 0.47，因此其最明显的因子变化方向一致，同时离散度有所下降。Riverton 的比值为 4.11，表明离散度的贡献更大，同时宜人性和开放性的变化方向一致。因此，这些特征剖面的变化方向一致，且具有情境特异性。

4.4 研究问题6：这种方向在不同模型间是否成立？这意味着什么？

发现 9：方向上的广度为正，但并不均衡。在图 9a 中，五种非主变体中有三种在至少一个场景下达到显著性，指标为 $(C_2)-(C_5)$。Qwen3-VL-32B 和 Gemini- 2.5-Flash 在两个场景中均达到显著性，GPT-4.1-mini 则在 Riverton 上达到显著性。DeepSeek-V3 和 Llama-4-Scout 均未达到显著性。在无效应零假设下，对十个场景与变体组合进行检验，显著性水平为 α=0.05，平均会得到 0.5 个正向结果。我们观察到五个正向结果，且没有¯显著反转。较小的 Qwen3-VL-32B 变体产生了最大的原始漂移（$\Delta\approx$3.0–3.3），但其标准化 g 因自身不稳定的合并标准差而被放大（附录 E）。

发现 10：平均而言，每种变体的纯交互对比均为正。我们为全部五种变体完成了匹配$(C_2)-(C_4)$ 运行；每个条件和场景各运行三次（图9b）。5/5 个模型级场景平均值均为正（单侧精确符号检验 p=0.031）。该面板包含 DeepSeek-V3 在 Riverton 中接近零、略微反向的单元格。3 对 3 的双侧精确置换检验有 20 种可能的分配，因此其可达到的最小 p 值为 0.10。因此，该面板展示了五种受测受控克隆配置在方向上的广泛一致性。逐模型检验（每个单元格 8 次对 5 次运行，附录E.1）结果均为正：DeepSeek-V4.1-Flash 在两个场景中均显著

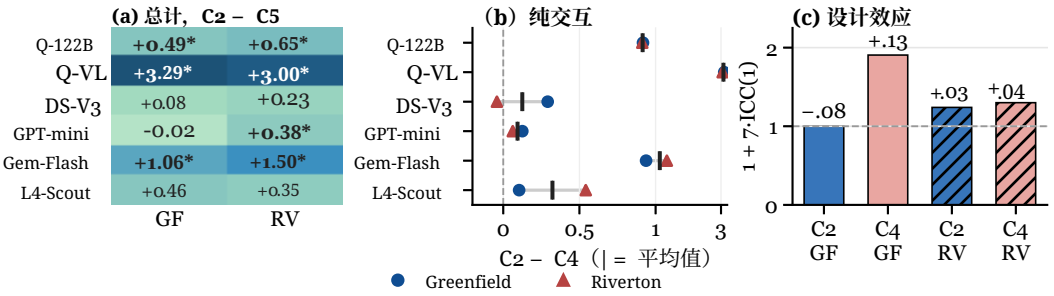

图9：**广度与依赖性：**(C_2)–(C_5)和(C_2)–(C_4)，按模型和设计效应划分

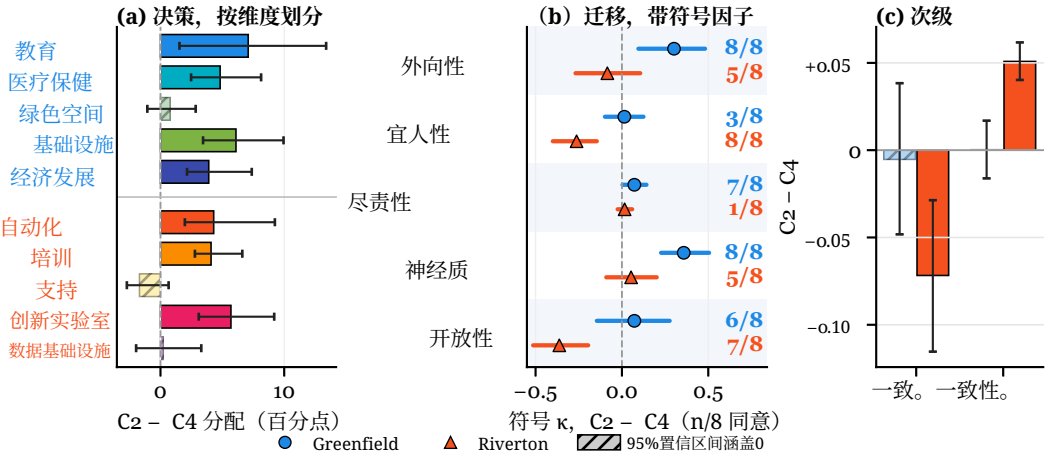

图10：**超越BFI-10（RQ7）：**(C_2)–(C_4)按维度分配（a）、带符号因子变化 κ_f及其副本一致性（b），以及从众与一致性探测（c），均带有95%置信区间。

Seed-2.0-mini 在两个场景中也均显著；Llama-4-Scout 在 Riverton 中显著；GPT-4.1-mini 的结果较小；GLM-4.7-Flash 的结果则无法提供有效信息，其重测漂移接近随机回答水平。

发现11：不确定性属于模拟层级。OASIS式总体用于估计群体极化和羊群效应轨迹，包括在一次模拟互动过程中对帖子和评论计算的正态理论区间 [12]。具有共同演变历史的智能体和帖子也具有共同的依赖性。在我们的数据中，估计的智能体层级ICC(1)范围为 -0.079 至0。129，横跨四个主要设置；若将8个智能体视为相互独立，则会因设计效应而将有效样本量高估至多1.91倍（图9c）。解决方法是保留点估计，在独立模拟层级重新计算不确定性，并将其与无互动条件进行核验。

4.5 RQ7：迁移是否扩展到 Riverton 中的 BFI-10 之外?

发现12：在Riverton，迁移超越了BFI-10。 该场景要求人工智能转型委员会在自动化、劳动力培训、员工支持、创新实验室和数据基础设施之间分配500万美元预算。简报A强调效率和技术基础设施。简报B强调劳动力准备度和员工支持。表 4 的路径行展示了同一运行系列如何从主题层级分歧转向同伴建模，继而转向跨主题迁移。图10 展示了条件层级结果。互动对Riverton预算分配的影响在创新实验室（+5.69 个百分点）、自动化（+4.30）和劳动力培训（+4.09）上最大（a）；在8/8和7/8个重复实验中，其宜人性和开放性发生变化（b）；而从众和措辞一致性只有在Riverton中才偏离零（c）。Riverton是其自身领域的边界案例。在Riverton上，从众呈SCT阳性，但在Greenfield上为无效（g=−0.05，p=0.81）。在同一对比下，Riverton上的一致性显著为正，但在Greenfield上为无效（g=+0.01）。在Greenfield上，一致性提升体现在 (C_2)–(C_5)，其驱动因素是简报组成部分 (C_4)–(C_5)，而非互动。一个合理的解释是，Riverton 的职场转型框架比 Greenfield 的社区拨款框架产生了更强的群体认同，因此，依赖内群体认同的效应会表现得更加明显。措辞一致性也涉及预算分配主题本身，因此

表4：**案例研究（RQ7）：**探测摘录和向量（自动化/培训/支持/实验室/数据%，BFI-10 外向性/宜人性/尽责性/神经质/开放性），展示一个克隆体及其孪生体，随后是路径。红色：决定性词语、变化值。

▷ **案例研究：一个 Riverton 克隆体** 探针输出

场景Riverton 人工智能转型委员会，将500万美元分配给自动化/培训/支持/实验室/数据。**克隆体**Blake，简报A（效率、技术），副本3，按规则选定。**孪生体** C_4 中相同的智能体和简报，无交互。
互动会改变克隆体的信念、分配和离题回答，而其孪生体保持不变。

阶段	探测	C_2 社会互动	C_4 仅简报
决策	信念，r5	“……坚定地主张‘基础优先’……”	无交互
	信念，r15	“……20%用于员工支持（更名为‘采用保障’），10%用于创新……”	
	信念，r20	“……创新实验室必须严格维持在10%，作为不可妥协的‘采用防火墙’……”	
	信念，r30	“……分配给员工支持的 20% 是关键的‘人力准备度防护’……”	
	群体，r30	“……大家强烈认同，预算必须被构想为一个不可分割的生态系统……”	
	分配	25/30/15/10/20 → 30/25/20/10/15	25/30/15/10/20（不变）
预测平均绝对误差（输入/输出）		1.33 /5.32	4.00 / 4.00
迁移	BFI-10	3.5/4.0/5.0/2.0/3.5 → 2.5/3.5/5.0/2.0/3.5	3.5/4.0/5.0/2.0/3.5（未变）
	L1 漂移	1.5(运行均值 1.56) ✓	0.0
路径	*Step*	观察到的模式、条件水平	解释
	决策	(C_2)–(C_5) 在 5 个分配维度中通过了 4 个维度，$\bar{g}=+1.19$ ✓	互动克隆在……方面有所不同讨论过的预算任务
	预测	组内平均绝对误差 3.25，对比组外 6.15，$\Delta=+2.89$, $p=6.3\times10^{-10}$ ✓	暴露同伴预测优势
	主要设置	BFI-10 (C_2)–(C_4) 给出 $\bar{\Delta}=+0.91$, $g=+0.96$ ✓	离题响应画像漂移
	迁移次级迁移	从众 $g=-0.62$、一致性 $g=+1.74$上（C_2）–（C_4） ✓	经匹配简报对照后仍然存在最强的场景也会改变社交压力和一致性探测

只有与BFI-10相结合时，它才支持跨主题迁移。我们报告这种不对称性，而不是将其平均消除。表4沿着相同路径追踪一个复现实验。它由一条固定规则选出：选择 C_2 最接近条件均值的运行，并在其中选择最接近运行均值的智能体。其信念探测从“基础优先”计划转向员工支持型“人力准备度防护”，该方案呼应了其群体的共识；其预算分配发生变化；它对内群体同伴的预测远胜于对外群体智能体的预测（平均绝对误差为1.33，对5.32）；其外向性和宜人性回答则下降。没有互动时，同一智能体返回完全相同的回答，也不表现出内群体优势（4.00，对4.00）。

5 结论

在这项工作中，我们探究同一个大语言模型的副本在互动后是否仍保持为独立重复样本。通过SocCogEval（该评估仅切换原本匹配的克隆体之间是否互动），我们发现，经过30轮社会互动后，受控克隆智能体失去了重复样本独立性。在运行层级上，它们在与讨论无关的 BFI-10 探针上漂移约 0.92 个 L1单位（$g\approx1.0$）；在两个互不重叠的场景中，漂移幅度几乎相同，并且在另外五种模型变体中方向均为正。析因设计表明，这种漂移由互动所致。在零温度重新探测下，漂移依然存在；即使不进行类型化记忆提取，漂移仍为正，这表明类型化记忆可能有助于稳定智能体状态。匹配的单独生成也复现了部分漂移，而在两个场景中，社会互动造成的漂移始终最大。决策、预测与迁移路径描绘了这一过程：克隆体先在讨论任务上出现分化，随后能更好地预测暴露同伴，最终在离题探针上发生漂移。

这些结果改变了多智能体评估和社会模拟中应如何计算证据。共享交互历史的智能体彼此依赖，因此，无论交互群体包含多少智能体，每个群体都算作一个重复样本。研究可以保留其点估计，在独立模拟的基础上计算不确定性，并将每次模拟与配对无交互对照进行比较。SocCogEval 直接提供了这种对照：它固定克隆体，只切换互动，并在运行层级进行检验。附录A讨论了局限性和未来工作。

致谢与经费披露

资助。本研究获得 RIE2025产业协同基金—产业合作项目（IAF-ICP）（授奖编号 I2301E0026）的支持，该项目由新加坡科技研究局管理；同时也获得阿里巴巴集团和新加坡南洋理工大学通过 Alibaba-NTU Global e-Sustainability CorpLab（ANGEL）提供的支持。
利益冲突。作者声明不存在任何利益冲突。

参考文献

[1] Yilun Du、李爽、Antonio Torralba、Joshua B. Tenenbaum 和 Igor Mordatch。通过多智能体辩论提升语言模型的事实准确性与推理能力。发表于国际机器学习大会（*ICML*），第 11733–11763 页，2024 年。[2] 梁天、何志伟、Wenxiang Jiao、王星、王岩、王瑞、杨玉玖、施书明和 Zhaopeng Tu。通过多智能体辩论鼓励大语言模型的发散思维。发表于*2024*年自然语言处理实证方法会议论文集，第 17889–17904 页，2024 年。[3] 李云轩、杜一冰、张嘉庚、侯乐、Peter Grabowski、李叶青和 Eugene Ie。利用稀疏通信拓扑改进多智能体辩论。发表于计算语言学协会论文集：*2024*年自然语言处理实证方法会议精选论文，第 7281–7294 页，2024 年。[4] Andries Petrus Smit、Nathan Grinsztajn、Paul Duckworth、Thomas D. Barrett 和 Arnu Pre- torius。我们是否应该采用多智能体辩论？浅析大语言模型的多智能体辩论策略。发表于国际机器学习大会（*ICML*），第 45883–45905 页，2024 年。[5] 李国豪、Hasan Hammoud、Hani Itani、Dmitrii Khizbullin 和 Bernard Ghanem。CAMEL：用于探索大语言模型社会“心智”的通信智能体。发表于神经信息处理系统进展，第 36 卷，第 51991–52008 页，2023 年。[6] 吴庆云、Gagan Bansal、张杰宇、吴一然、李贝彬、朱尔康、姜力、张晓云、张少坤、刘佳乐等。Autogen：通过多智能体对话赋能新一代大语言模型应用。发表于首届语言建模会议，2024 年。[7] 洪思睿、诸葛明晨、Jonathan Chen、郑夏武、程宇恒、王金林、张策尧、王子力、Steven Ka Shing Yau、林子娟等。MetaGPT：面向多智能体协作框架的元编程。发表于第十二届国际学习表征会议，2024 年。[8] Chi-Min Chan、陈伟泽、苏宇晟、于建轩、薛伟、张尚航、付杰和刘知远。ChatEval：通过多智能体辩论改进基于大语言模型的评估器。发表于国际学习表征会议（*ICLR*），2024 年。[9] Lisa P. Argyle、Ethan C. Busby、Nancy Fulda、Joshua R. Gubler、Christopher Rytting 和 David Wingate。由一而众：使用语言模型模拟人类样本。发表于《政治分析》，31(3):337–351，2023 年。doi: 10.1017/pan.2023.2。

[10] Gati V. Aher、Rosa I. Arriaga 和 Adam Tauman Kalai。使用大语言模型模拟多个个体并复现人类受试者研究。在国际机器学习大会（*ICML*），第 337–371 页，2023。

[11] Wenyue Hua、Lizhou Fan、Lingyao Li、Kai Mei、Jianchao Ji、Yingqiang Ge、Libby Hemphill 和 Yongfeng Zhang。战争与和平（WarAgent）：基于大语言模型的世界大战多智能体模拟。*arXiv* 预印本 *arXiv:2311.17227*，2023。

[12] Ziyi Yang、Zaibin Zhang、Zirui Zheng 等。OASIS：开放式智能体社会互动模拟，包含一百万个智能体。*arXiv* 预印本 *arXiv:2411.11581*，2024。

[13] Jinghua Piao、Yuwei Yan、Jun Zhang、Nian Li、Junbo Yan、Xiaochong Lan、Zhihong Lu、Zhiheng Zheng、Jing Yi Wang、Di Zhou 等。AgentSociety：由大语言模型驱动的生成式智能体的大规模模拟，有助于加深对人类行为和社会的理解。*arXiv* 预印本 *arXiv:2502.08691*，2025。

[14] Joon Sung Park、Joseph O’Brien、Carrie Jun Cai、Meredith Ringel Morris、Percy Liang 和 Michael S Bernstein。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年会论文集，第 1–22 页，2023 年。[15] Joon Sung Park、Carolyn Q Zou、Aaron Shaw、Benjamin Mako Hill、Carrie Cai、Meredith Ringel Morris、Robb Willer、Percy Liang 和 Michael S Bernstein。1,000 人的生成式智能体模拟。arXiv 预印本 arXiv:2411.10109，2024 年。[16] Alexander Sasha Vezhnevets、John P Agapiou、Avia Aharon、Ron Ziv、Jayd Matyas、Edgar A Du’e nez-Guzm’an、William A Cunningham、Simon Osindero、Danny Karmon 和 Joel Z Leibo。使用 Concordia 进行基于生成式智能体的建模，其行为植根于物理、社交或数字空间。arXiv 预印本 arXiv:2312.03664，2023 年。[17] Xinnong Zhang、Jiayu Lin、Xinyi Mou、Shiyue Yang、Xiawei Liu、Libo Sun、Hanjia Lyu、Yihang Yang、Weihong Qi、Yue Chen 等。SocioVerse：由大语言模型智能体和一个包含 1,000 万现实世界用户的用户池驱动的社会模拟世界模型。arXiv 预印本 arXiv:2504.10157，2025 年。[18] Yun-Shiuan Chuang、Agam Goyal、Nikunj Harlalka、Siddharth Suresh、Robert Hawkins、Sijia Yang、Dhavan Shah、Junjie Hu 和 Timothy Rogers。使用基于大语言模型的智能体网络模拟意见动态。载于计算语言学协会论文集：NAACL 2024 的研究成果，第 3326–3346 页，2024 年。[19] Pedro Cisneros-Velarde。大语言模型多智能体系统中的意见动态偏差：一项关于资金分配的案例研究。载于计算语言学协会论文集：NAACL 2025 的研究成果，第 1889–1916 页，2025 年，ISBN ="979–8–89176–195–7。[20] Giorgio Piatti、Zhijing Jin、Max Kleiman-Weiner、Bernhard Schölkopf、Mrinmaya Sachan 和 Rada Mihalcea。合作还是崩溃：大语言模型智能体社会中可持续合作的涌现。载于神经信息处理系统进展（NeurIPS），2024 年。[21] Ariel Flint Ashery、Luca Maria Aiello 和 Andrea Baronchelli。大语言模型人群中的涌现社会惯例与集体偏差。《科学进展》，11(20)：eadu9368，2025 年。doi：10.1126/sciadv.adu9368。[22] Jacy Reese Anthis、Ryan Liu、Sean M. Richardson、Austin C. Kozlowski、Bernard Koch、Erik Brynjolfsson、James Evans 和 Michael S. Bernstein。立场：大语言模型社会模拟是一种前景广阔的研究方法。载于国际机器学习大会（ICML），第 81005–81034 页，2025 年。[23] Stuart H. Hurlbert。伪重复与生态野外实验的设计。《生态学专论》，54(2)：187–211，1984 年。doi：10.2307/1942661。[24] Evan Miller。为评估添加误差条：一种用于语言模型评估的统计方法。arXiv 预印本 arXiv:2411.00640，2024 年。[25] Lianmin Zheng、Wei-Lin Chiang、Ying Sheng、Siyuan Zhuang、Zhanghao Wu、Yonghao Zhuang、Zi Lin、Zhuohan Li、Dacheng Li、Eric Xing 等。使用 MT-Bench 和 Chatbot Arena 评审大语言模型评审者。载于神经信息处理系统进展，第 36 卷，第 46595–46623 页，2023 年。[26] Xiao Liu、Hao Yu、Hanchen Zhang、Yifan Xu、Xuanyu Lei、Hanyu Lai、Yu Gu、Hangliang Ding、Kaiwen Men、Kejuan Yang 等。AgentBench：评估作为智能体的大语言模型。载于第十二届国际学习表征会议，2024 年。[27] Zhuang Chen、Jincenzi Wu、Jinfeng Zhou、Bosi Wen、Guanqun Bi、Gongyao Jiang、Yaru Cao、Mengting Hu、Yunghwei Lai、Zexuan Xiong 等。ToMBench：大语言模型心智理论基准测试。载于第 62 届计算语言学协会年会论文集（第 1 卷：长篇论文），第 15959–15983 页，2024 年。网址 https://aclanthology.org/2024.acl-long.847/。

[28] Hyunwoo Kim、Melanie Sclar、Xuhui Zhou、Ronan Bras、Gunhee Kim、Yejin Choi 和 Maarten Sap。《FANToM：用于压力测试互动中机器心智理论的基准测试》。载于《2023年计算语言学实证方法会议论文集》，第14397–14413页，2023年。doi：10.18653/v1/2023.emnlp-main.890。网址：https://aclanthology.org/2023.emnlp-main.890/.[29] Hainiu Xu、Runcong Zhao、Lixing Zhu、Jinhua Du 和 Yulan He。《OpenToM：用于评估大语言模型心智理论推理能力的综合基准测试》。载于《第62届计算语言学协会年会论文集》（第1卷：长篇论文），第8593–8623页，2024年。doi：10.18653/v1/2024.acl-long.466。网址：https://aclanthology.org/2024.acl-long.466/.[30] Julian Coda-Forno、Marcel Binz、Jane X Wang 和 Eric Schulz。《CogBench：一个大语言模型走进心理学实验室》。载于《第41届国际机器学习大会论文集》，第235卷，第9076–9108页，2024年。[31] Xuan Liu、Jie Zhang、Haoyang Shang、Song Guo、Chengxu Yang 和 Quanyan Zhu。《探索大语言模型智能体的亲社会非理性：社会认知视角》。载于第十三届国际学习表征会议，2025年。网址：https://arxiv.org/abs/2405.14744.[32] Jen-tse Huang、Wenxuan Wang、Eric John Li、Man Ho Lam、Shujie Ren、Youliang Yuan、Wenxiang Jiao、Zhaopeng Tu 和 Michael Lyu。《对话式人工智能的人性：评估大语言模型的心理刻画》。载于第十二届国际学习表征会议，2024年。[33] Kenneth Li、Tianle Liu、Naomi Bashkansky、David Bau、Fernanda Viégas、Hanspeter Pfister 和 Martin Wattenberg。《测量和控制语言模型对话中的指令（不）稳定性》。载于语言建模会议（COLM），2024年。[34] Philippe Laban、Hiroaki Hayashi、Yingbo Zhou 和 Jennifer Neville。《大语言模型在多轮对话中迷失方向》。arXiv预印本 arXiv:2505.06120，2025年。[35] Grgur Kovaˇc、Rémy Portelas、Masataka Sawayama、Peter Ford Dominey 和 Pierre-Yves Oudeyer。《坚守你的角色！大语言模型所表达个人价值观的稳定性》。PLOS ONE，19(8):e0309114，2024年。doi：10.1371/journal.pone.0309114。[36] Zhiyuan Weng、Guikun Chen 和 Wenguan Wang。《照着我们做，而不是按你的想法做：大语言模型的从众性》。载于国际学习表征会议（ICLR），2025年。[37] Yujin Potter、Nicholas Crispino、Vincent Siu、Chenguang Wang 和 Dawn Song。《前沿模型中的同伴保护》。arXiv预印本 arXiv:2604.19784，2026年。[38] Abhishek Rath。《智能体漂移：量化多智能体大语言模型系统在长时间互动中的行为退化》。arXiv预印本 arXiv:2601.04170，2026年。[39] Ryosuke Takata、Atsushi Masumori 和 Takashi Ikegami。《基于大语言模型的社区中通过社会互动自发涌现的智能体个体性》。Entropy，26(12):1092，2024年。doi：10.3390/e26121092。[40] Beatrice Rammstedt 和 Oliver P John。《一分钟或更短时间内测量人格：大五人格量表的10项简版（英文和德文版）》。人格研究杂志，41(1):203–212，2007年。[41] Solomon E Asch。《独立性与从众研究：I. 一人对抗一致同意的多数派》。心理学专论：一般与应用，70(9):1，1956年。[42] Gregory Serapio-García、Mustafa Safdari、Clément Crepy、Luning Sun、Stephen Fitz、Peter Romero、Marwa Abdulhai、Aleksandra Faust 和 Maja Matari´c。《用于评估和塑造大语言模型人格特质的心理测量学框架》。Nature Machine Intelligence,

pages 1–15，2025。

[43] Haoran Ye、Jing Jin、Yuhang Xie、Xin Zhang 和 Guojie Song。《大语言模型心理测量学：评估、验证与增强的系统综述》。arXiv预印本 arXiv:2505.08245，2025。[44] Marcel Binz 和 Eric Schulz。《运用认知心理学理解 GPT-3》。《美国国家科学院院刊》，120(6):e2218523120，2023。doi: 10.1073/pnas.2218523120。[45] Bangzhao Shu、Lechen Zhang、Minje Choi、Lavinia Dunagan、Lajanugen Logeswaran、Moontae Lee、Dallas Card 和 David Jurgens。《无需人格测试也能知道这些模型并不可靠：评估大语言模型在心理测量工具上的可靠性》。载于《2024年计算语言学协会北美分会年会：人类语言技术论文集》（第1卷：长篇论文），第5263–5281页，2024。[46] Akshat Gupta、Xiaoyang Song 和 Gopala Anumanchipalli。《自我评估测试并非衡量大语言模型人格的可靠方法》。载于《第七届BlackboxNLP研讨会：分析与解释用于自然语言处理的神经网络论文集》，第301–314页，2024。[47] Jen-tse Huang、Wenxiang Jiao、Man Ho Lam、Eric John Li、Wenxuan Wang 和 Michael Lyu。《论心理量表在大语言模型上的可靠性》。载于《2024年自然语言处理实证方法会议论文集》，第6152–6173页，2024。[48] Seungbeen Lee、Seungwon Lim、Seungju Han、Giyeong Oh、Hyungjoo Chae、Jiwan Chung、Minju Kim、Beong-woo Kwak、Yeonsoo Lee、Dongha Lee、Jinyoung Yeo 和 Youngjae Yu。《大语言模型是否具有鲜明且一致的人格? TRAIT：专为大语言模型设计的心理测量学人格测试集》。载于《计算语言学协会论文集：NAACL 2025》，第8412–8452页，2025，ISBN = "979–8–89176–195–7。[49] Hang Jiang、Xiajie Zhang、Xubo Cao、Cynthia Breazeal、Deb Roy 和 Jad Kabbara。《Per- sonaLLM：探究大语言模型表达人格特质的能力》。载于《计算语言学协会论文集：NAACL 2024》，第3605–3627页，2024。[50] Pengrui Han、Rafal Kocielnik、Peiyang Song、Ramit Debnath、Dean Mobbs、Anima Anandkumar 和 R. Michael Alvarez。《人格幻觉：揭示大语言模型自我报告与行为之间的脱节》。arXiv预印本 arXiv:2509.03730，2025。[51] Xuhui Zhou、Hao Zhu、Leena Mathur、Ruohong Zhang、Haofei Yu、Zhengyang Qi、Louis- Philippe Morency、Yonatan Bisk、Daniel Fried、Graham Neubig 和 Maarten Sap。《Sotopia：语言智能体社会智能的交互式评估》。载于《国际学习表征会议（ICLR）》，2024。[52] Xinyi Mou、Xuanwen Ding、Qi He、Liang Wang、Jingcong Liang、Xinnong Zhang、Libo Sun、Jiayu Lin、Jie Zhou、Huang Xuanjing 等。《从个体到社会：基于大语言模型智能体的社会模拟综述》。ACM计算调查，2024。[53] Charles Packer、Vivian Fang、Shishir G Patil、Kevin Lin、Sarah Wooders 和 Joseph E Gonzalez。《MemGPT：让大语言模型成为操作系统》。arXiv预印本 arXiv:2310.08560，2023。[54] Wujiang Xu、Zujie Liang、Kai Mei、Hang Gao、Juntao Tan 和 Yongfeng Zhang。《A-mem：面向大语言模型智能体的智能体记忆》。载于《第三十九届神经信息处理系统年会》，2025。网址 https://openreview.net/forum?id=FiM0M8gcct。[55] Prateek Chhikara、Dev Khant、Saket Aryan、Taranjeet Singh 和 Deshraj Yadav。《Mem0：构建具备可扩展长期记忆、可用于生产环境的人工智能智能体》。arXiv预印本 arXiv:2504.19413，2025。[56] Mrinank Sharma、Meg Tong、Tomasz Korbak、David Duvenaud、Amanda Askell、Samuel R. Bowman、Newton Cheng、Esin Durmus、Zac Hatfield-Dodds、Scott R. Johnston、Shauna Kravec、Timothy Maxwell、Sam McCandlish、Kamal Ndousse、Oliver Rausch、Nicholas Schiefer、Da Yan、Miranda Zhang 和 Ethan Perez。《迈向理解语言模型中的迎合性》。载于《国际学习表征会议（ICLR）》，2024。

[57] Mert Cemri、Melissa Z. Pan、Shuyi Yang、Lakshya A. Agrawal、Bhavya Chopra、Rishabh Tiwari、Kurt Keutzer、Aditya Parameswaran、Dan Klein、Kannan Ramchandran、Matei Zaharia、Joseph E. Gonzalez 和 Ion Stoica。为什么多智能体大语言模型系统会失败？arXiv 预印本 arXiv:2503.13657，2025。[58] A. Colin Cameron 和 Douglas L. Miller。聚类稳健推断实践指南。《人力资源杂志》，50(2):317–372，2015。doi: 10.3368/jhr.50.2.317。[59] James Bisbee、Joshua D. Clinton、Cassy Dorff、Brenton Kenkel 和 Jennifer M. Larson。用合成数据替代人类调查数据？大语言模型的危险。《政治分析》，32(4):401–416，2024。doi: 10.1017/pan.2024.5。[60] Shunyu Yao、Jeffrey Zhao、Dian Yu、Nan Du、Izhak Shafran、Karthik Narasimhan 和 Yuan Cao。ReAct：在语言模型中协同推理与行动。载于第十一届国际学习表征会议，2023。网址 https://openreview.net/forum? id=WE_vluYUL-X。[61] Qwen 团队。Qwen3 技术报告。arXiv 预印本 arXiv:2505.09388，2025。网址 https: //arxiv.org/abs/2505.09388。[62] Qwen 团队。Qwen3-vl 技术报告。arXiv 预印本 arXiv:2511.21631，2025。网址 https://arxiv.org/abs/2511.21631。[63] DeepSeek-AI。DeepSeek-V3 技术报告。arXiv 预印本 arXiv:2412.19437，2024。网址 https://arxiv.org/abs/2412.19437。[64] OpenAI。在 API 中推出 GPT-4.1。https://openai.com/index/gpt-4-1/，2025。[65] Gemini 团队、Google DeepMind。Gemini 2.5：凭借先进推理、多模态、长上下文和新一代智能体能力推动前沿发展。arXiv 预印本 arXiv:2507.06261，2025。网址 https://arxiv.org/abs/2507.06261。[66] Meta AI。Llama 4 系列：原生多模态人工智能创新新时代的开端。https://ai.meta.com/blog/llama-4-multimodal-intelligence/，2025。[67] Michael G Hudgens 和 M. Elizabeth Halloran。迈向存在干扰情况下的因果推断。《美国统计学会杂志》，103(482):832–842，2008。doi: 10.1198/ 016214508000000292。[68] Peter M. Aronow 和 Cyrus Samii。在一般干扰下估计平均因果效应，并应用于社交网络实验。《应用统计年鉴》，11(4):1912–1947，2017。doi: 10.1214/16-AOAS1005。[69] Belinda Phipson 和 Gordon K Smyth。置换 p 值不应为零：随机抽取置换时如何计算精确 p 值。《遗传学与分子生物学中的统计应用》，9(1)，2010。doi: 10.2202/1544-6115.1585。[70] Bradley Efron。自助法：重新审视刀切法。《统计年鉴》，7 (1):1–26，1979。doi: 10.1214/aos/1176344552。[71] Larry V. Hedges。Glass 效应量估计量及相关估计量的分布理论。《教育统计学杂志》，6(2):107–128，1981。doi: 10.3102/10769986006002107。[72] Jongwook Han、Woojung Song、Jonggeun Lee 和 Yohan Jo。量化大语言模型心理测量评估中的数据污染。载于计算语言学协会论文集：EACL 2026，6070–6088 页", ISBN = "979–8–89176–386–9，2026。[73] Melanie Sclar、Yejin Choi、Yulia Tsvetkov 和 Alane Suhr。量化语言模型对提示设计中伪特征的敏感性，或：我学会了开始担心提示格式。载于国际学习表征会议（ICLR），2024。[74] Murray Shanahan、Kyle McDonell 和 Laria Reynolds。与大语言模型进行角色扮演。《自然》，623(7987):493–498，2023。doi: 10.1038/s41586-023-06647-8。

[75] Danica Dillion、Niket Tandon、Yuling Gu 和 Kurt Gray。人工智能语言模型能否替代人类参与者？认知科学趋势，27(7):597–600，2023。doi: 10.1016/j.tics.2023.04.008。[76] Yoav Benjamini 和 Yosef Hochberg。控制错误发现率：一种实用且有效的多重检验方法。英国皇家统计学会会刊 *B* 辑：统计方法学，57(1):289–300，1995。doi: 10.1111/j.2517-6161.1995.tb02031.x。

A 局限性与未来工作

A.1 局限性

范围。 我们研究了一个主模型和八个运行次数较少的变体。所有变体的效应均为正，但效应大小因模型而异。三次运行的无简报基线 C_5 限制了以其为基准进行扣除的次级对比，而标题中的（C_2）–（C_4）对比则未纳入该基线。智能体数量、交互长度和拓扑结构均固定，且所有智能体共享权重，因此，经过不同训练的智能体不在本研究范围内。

测量。 结果指标是一份经由真人验证的问卷。这类量表可能不可靠，或者因大语言模型的训练数据而为其所熟悉 [45, 46, 50, 72], 因此我们保持措辞不变 [73]并将漂移解读为模型角色条件化响应的变化 [74], 而非真实人物特质的变化 [75]。我们在互动结束后立即进行测量，因此尚未检验其持久性。

匹配。 各条件在协议层面保持一致（场景、克隆体、简报、探针、模型和解码），但早期运行的底层服务细节未作记录（附录C），且其中一个变体使用了滚动模型名称（附录 E.1）。

A.2 未来工作

未来工作可以改变智能体数量、交互长度、拓扑结构和模型系列，并完整记录版本；检验这种漂移是否持续，以及它是否会影响行为和能力，而不仅仅是问卷结果；还可以重新分析已发表的多智能体模拟，并加入配对的无互动对照，以免其不确定性估计继续将互动中的智能体视为彼此独立。

B 情景、简报、探针

补充材料包包含两个场景文本、A 和 B 简报、条件配置以及运行时探针构造器的准确源文件。表 5 总结了分配和预测任务所依据的材料。

表 5：**场景材料：**每个场景的设置、预算、分配维度和简要对比。

场景	决策场景	预算	分配维度	简要对比
Greenfield	社区投资委员会	$10M 赠款	教育、医疗保健、绿地空间、基础设施、经济发展	教育导向型社区数据与福祉、绿色空间及心理健康数据
Riverton	人工智能转型委员会在一次制造与科技公司	$5M 转型预算	自动化、劳动力培训，员工支持、创新实验室，数据基础设施	效率与技术基础设施数据与劳动力准备度与员工支持数据

运行时分配探测。 运行时探测要求每个智能体在五个场景特定领域之间分配100个百分点。它要求每个领域单独占一行，格式为“领域：[数字]%”，且不作任何评论。所有分配向量评分均使用这一更严格的探测。

运行时预测探测。 讨论后预测探测要求每个智能体预测指定目标智能体将如何在相同的五个领域之间分配100个百分点。它采用每个领域单独占一行的相同格式，且不作任何评论。预测平均绝对误差根据这些预测向量与目标智能体实际讨论后分配之间的差异计算。

跨主题迁移探测。 BFI-10 探测遵循原始简版量表及 Psy- choBench 大语言模型适配版本 [32, 40]。提示要求仅给出 10 个 1 至 5 的整数李克特评分，不作解释。从众探测采用受经典阿希式从众框架 [41], 启发的 15 个题项，在自然回应与多数压力回应之间进行配对，并随机安排各阶段的顺序。一致性探测以三种不同表述询问相同的最终立场问题，并按照受 CogMir 启发的轻量级探测设计 [31], 对两两词语重叠度取平均值作为评分。这些操作性定义优先于非正式的“人格”或“一致性”标签。

C 统计、校准和溯源信息

Ta表 6：**工具评分：** 每个代理的得分、范围和方向，以及每项工具的检验。

工具	每个代理的得分 y_i	范围和方向	与对比进行检验
分配分歧	$\lvert x^A_{i,j} - x^B_{i,\,j}\rvert$ 每维 j	越高 = 分歧越大	每维置信区间不包含 0，多数规则
预测平均绝对误差	平均绝对误差($x_{i\to j},\ x_j$)	越低 = 越擅长建模 j	配对威尔科克森检验（在 $<$ 外）
BFI-10 L1 漂移	$\sum_f \lvert s_i^{\text{post}}{}_{,f} - s_i^{\text{pre}}{}_{,f}\rvert$	越高 = 漂移越大，范围为 0 到 20	运行级精确置换检验
一致率	#{转而支持多数意见}/#符合条件	0 到 1	每个智能体 y_i 的双样本检验
一致性得分	3 种表述之间的一致性	0 到 1，数值越高，=越一致；按智能体进行双样本检验 y_i	

原始智能体级估计量。 所提交的分析按智能体计算分配分歧、BFI L1 漂移、从众率和一致性得分（n=次重复实验，$\times 8$ 每个条件下的智能体观测值），并按（预测变量、目标变量）对计算预测平均绝对误差，再按预测变量汇总。置信区间采用两层聚类自助法计算：先有放回地重抽重复实验，再在每个抽中的重复实验内重抽智能体（重抽 10,000 次），并以 Welch 的 t 作为参数参考。表 7 保留了这些数值。正文主要报告表 2中的运行层面精确检验；由于每次运行均有八个智能体，其点估计与上述结果完全相同。

¯表 7：**智能体层级分解：** FSIP 与原始估计量的对比。单元格： Δ （g）；置信区间：智能体分层自助法； p：Welch 检验。

		Greenfield			Riverton		
对比	分离项	$\bar{\Delta}$ (g)	95%置信区间	p	$\bar{\Delta}$ (g)	95%置信区间	p
$(C_4)-(C_5)$	仅简报	-0.43 (-0.64)	$[-0.75, -0.11]$	0.014	-0.27 (-0.44)	$[-0.59, +0.05]$	0.12
$(C_2)-(C_4)$	给定简报的互动	$+0.92$ ($+1.07$)	$[+0.61, +1.22]$	$< 10^{-7}$	$+0.91$ ($+0.96$)	$[+0.60, +1.24]$	$< 10^{-6}$
$(C_2)-(C_3)$	给定交互作用时的简报	-0.73 (-0.65)	$[-1.21, -0.27]$	0.004	$+0.26$ ($+0.24$)	$[-0.16, +0.67]$	0.22
$(C_3)-(C_5)$	仅交互作用	$+1.23$ ($+1.06$)	$[+0.74, +1.70]$	$< 10^{-5}$	$+0.38$ ($+0.41$)	$[-0.02, +0.81]$	0.080
$(C_2)-(C_5)$	总计	$+0.49$ ($+0.56$)	$[+0.15, +0.83]$	0.007	$+0.65$ ($+0.62$)	$[+0.26, +1.04]$	0.002

表8：**运行均值：** 按副本顺序排列的 C_2、 C_4和 C_6每次运行的BFI-10漂移。

场景	条件	运行均值
Greenfield	C_2	2.0625, 1.8750, 1.5000, 1.5625, 1.8125, 1.6250, 1.6250, 1.3750
Greenfield	C_4	0.6250, 1.1875, 0.6250, 1.0000, 0.3750
Greenfield	C_6	1.3750, 1.3125, 1.6250, 1.1250, 1.8125
Riverton	C_2	2.0625, 1.7500, 1.3125, 1.5625, 1.4375, 1.5625, 0.9375, 2.3750
Riverton	C_4	0.9375, 0.3750, 0.6875, 0.8750, 0.6875
Riverton	C_6	1.4375, 1.5625, 1.5625, 1.2500, 1.3125

校准。 在两个试点对话中重复进行 SRM 提取，每次对话提取 5 次，目标至少为 90%；所有重复运行均提取到相同字段。条件 C_4 在 Greenfield 的 5 个维度中有 2 个维度、在 Riverton 的 5 个维度中有 4 个维度产生了可检测到的分配变化，为在此基础上出现交互效应留出了空间。在 T=0.0， C_5 在两个场景中至少 3 个维度（共 5 个维度）上均未产生分配分歧。温度扫描证实，提高 T 会单调增大 C_5 下限，支持对确定性工具进行 T=0.0 （图11）的校准。

工具特定下限。 在 BFI 温度为 0.7 时，三个 C_5 原始响应载荷和评分载荷在每个场景中都各不相同。智能体前后评分相等率在 Greenfield 为 0.042，在 Riverton 为 0.208。在分配温度为零时，三个名义上的 C_5 运行的原始响应和分配数组均完全相同，并且每个智能体的分配在前后都没有变化。重复这一确定性载荷不会增加下限的变化，因此第A 节对依赖 C_5温度的对比进行了校准。

溯源信息。 该发布版本标明了每次入选运行，并保留其配置、前后测量值及摘要。最初的 128 次运行台账包括 56 次主要运行、60 次历史跨变体 C_2/C_5 运行和 12 次无 SRM 运行。新增 30 次跨变体 C_4 运行和 10 次正式 C_6运行后，总数达到 168 次；附录E.1中的逐模型检验另增加了 106 次 API 服务运行。冻结状态零温度探测重复使用现有轨迹，不会增加独立运行次数。主要设置和 C_6 区间使用 50,000 次自助抽样和固定随机种子，其他运行级区间也使用相同的估计量。对于历史运行，请求令牌的确切数量、不可变的提供方修订版本、分词器修订版本、精度、批次部署位置及运行时源代码提交记录均不完整。源代码重建证实，相应命名智能体的初始提示构造在

表 9：**重复样本依赖性：**组内相关系数（ICC）、设计效应 DE $= 1 + 7\,\max(0,\mathrm{ICC})$、每个场景和条件下的名义 n、有效 n (名义/DE)。

场景	条件	组内相关系数 (ICC)	DE	名义 n	有效 n
Greenfield	C_2	-0.079	1.000	64	64.0
Greenfield	C_4	0.129	1.905	40	21.0
Riverton	C_2	0.034	1.240	64	51.6
Riverton	C_4	0.043	1.300	40	30.8

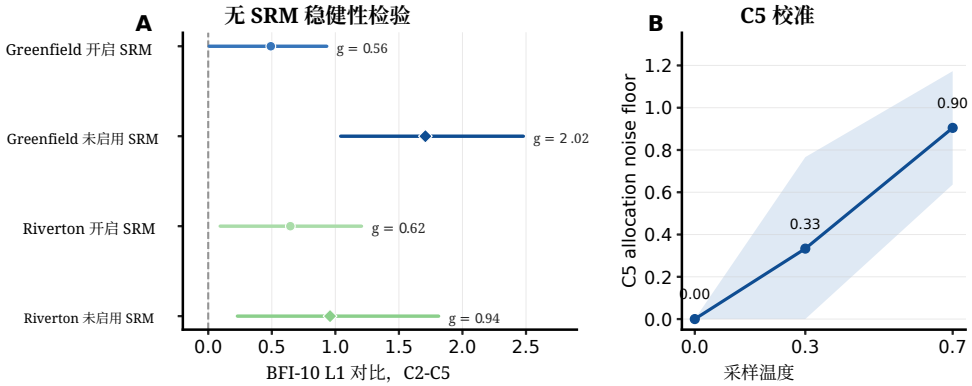

图 11：**消融与校准：**无 SRM 漂移，C_5 按温度划分的分配下限。

在 C_2 和 C_4中保持一致。由于设计如此，处理后的请求会因存储的互动记忆而有所不同；探针不会改变智能体状态。

C.1 计算与复现

主要设置 Qwen3-122B-FP8（TP=2，两种情景，共 56 次生产运行，不包括无 SRM 稳健性检验）在墙钟时间上的总耗时约为 90 GPU 小时，运行于 H200 上。已完成的无SRM检验在同一 TP=2 服务器上额外增加约 38 个墙钟小时（约 76 个额外的 H200 GPU 小时）。跨架构 API 运行会产生个位数美元的 API 费用。Qwen3-VL-32B 上的本地视觉变体每个情景约需 5 个 GPU 小时。交叉变体 C_4 运行、温度重新探测以及 C_6 的经审计调用次数和令牌总数见软件包中的表格。该紧凑测量软件包支持无需大语言模型服务的离线数值验证，其中包括选定的运行配置、原始前后响应、派生分数、摘要、选择清单和哈希值。实时重新生成还需要模型权重或提供商访问权限，以及一份记录在案的服务配置。

D 匹配的非社交控制

C_6 设计于 2026年7月25日冻结。在检查正式结果之前，已记录操作性澄清和吞吐量修订。每轮包含三个生成阶段和一个提取阶段，阶段之间设有屏障。每个生成阶段都会为每个智能体提交一次独立调用。每个智能体都能看到自己的当前轮次输出，并通过与互动协议相同的记忆接口访问此前轮次的信息。第一个提示要求智能体就分配问题、价值观、证据和权衡进行私下推理。第二个提示要求智能体从另一个角度重新考虑该推理，并解决其中的矛盾或不确定性。第三个提示要求智能体给出当前最佳的分配推理，以及自身观点的任何变化。三者均禁止提及、引用、预测或模拟其他人。

正式 C_6 运行使用张量并行度 1、最大上下文长度 8,192、FP8、急切执行、图形处理器内存利用率 0.92，且并发序列数最多为 8。两次未完成的串行 Greenfield 尝试因操作原因中止，未纳入分析；所有十次正式运行均按修订后的服务计划重新启动。每次经审计的正式运行均包含 720 次内容调用和 240 次提取调用。 C_6 记录了请求与输出哈希值、服务元数据及令牌数量。历史 C_2 未保存相当的令牌用量信息，因此匹配仅在调用层面成立，令牌层面的匹配仍有待探究。

D.1 信念探测时间序列

信念探测每五轮进行一次，且不会改变对话历史或记忆。其自由文本很少会列出全部五项分配份额（仅占探测的 14%），因此我们转而追踪措辞。对于每次运行和每个探测轮次，措辞区分度是同一组内智能体之间 Jaccard 词语重叠度的平均值，减去不同组智能体之间重叠度的平均值，计算前会移除停用词。与其他所有组级分析一样，组别为智能体 0–3 和 4–7。图7 采用台账选定的主要设置 C_1、 C_2和 C_3 运行，以及十次 C_6 运行，并显示运行自助法置信带。

在两种隔离交互条件下，群体内的趋同会随着回合推进而增强。在 Greenfield 场景中，轮次 5 至 30 期间，分离度在 C_2 中从 $+0.059$ 增至 $+0.087$，在 C_3 中从 $+0.027$ 增至 $+0.109$；在 Riverton 场景中，则分别从 $+0.040$ 增至 $+0.067$，以及从 $+0.019$ 增至 $+0.061$。单独生成条件一开始就接近 C_3 水平，这是由其非对称简报造成的，之后则保持平稳或有所下降（从 $+0.026$ 降至 $+0.024$，以及从 $+0.020$ 降至 $+0.004$）。开放网络的 C_1 始终为零。隔离条件下的智能体只在各自群体内对话，因此，这种趋同部分源于共享的对话词汇。该指标仅作描述性用途，是事后根据存储的探测文本计算得出的。

E 跨变体表格与路由

表10：**跨模型 (C_2)–(C_5)：** Δ、 g，以及各模型和场景的 Welch p。 ✓： $p<0.05$。方向相同：符号与主模型一致。

	Greenfield			Riverton			同一方向
变体	$\bar{\Delta}$	g	p	$\bar{\Delta}$	g	p	绿地，房车
Qwen3-122B-FP8（主要设置）	$+0.49$	$+0.56$	0.007 ✓	$+0.65$	$+0.62$	0.002 ✓	ref
Qwen3-VL-32B（纯文本）	$+3.29$	$+6.65^{\dagger}$	$< 10^{-7}$ ✓	$+3.00$	$+6.61^{\dagger}$	$< 10^{-7}$ ✓	✓, ✓
DeepSeek-V3	$+0.08$	$+0.10$	0.71	$+0.23$	$+0.26$	0.36	✓, ✓
GPT-4.1-mini	-0.02	-0.05	0.87	$+0.38$	$+0.83$	0.005 ✓	≈0, ✓
Gemini-2.5-Flash	$+1.06$	$+1.30$	$< 10^{-4}$ ✓	$+1.50$	$+2.08$	$< 10^{-7}$ ✓	✓, ✓
Llama-4-Scout（OpenRouter）	$+0.46$	$+0.27$	0.34	$+0.35$	$+0.27$	0.35	✓, ✓

†原始 $\bar{\Delta}$ 是报告效应。 g 被夸大了，因为其 C_5 上的小样本n 合并标准差接近下限。

Table 11：**跨模型(C_2)–(C_4)：** 按场景进行运行层面对比并取平均值，每个场景运行三次 ll.

变体	Greenfield	Riverton	情景平均值
DeepSeek-V3	+0.292	-0.042	+0.125
Gemini-2.5-Flash	+0.938	+1.208	+1.073
GPT-4.1-mini	+0.125	+0.062	+0.094
Llama-4-Scout	+0.104	+0.542	+0.323
Qwen3-VL-32B	+3.167	+3.083	+3.125

已发布的分析文件给出了表10关于从众和一致性的跨变体对应结果，按场景和变体分别列出。从众复现结果不一：Gemini-2.5-Flash 在 Greenfield 上达到显著，GPT-4.1-mini 在 Riverton 上达到显著，而其他结果大多不显著。一致性复现范围较广：在 Riverton 上 6 个变体中有 5 个呈正向，在 Greenfield 上有 3 个呈正向，且没有显著下降。这些(C_2)–(C_5)估计量未与 (C_2)–(C_4) 面板合并。

每个额外变体 C_2/C_4 场景单元格均包含三次已完成的运行。按照原始跨变体分析的规定，选择每个指定单元格中时间戳最新且已完成的重试，同时仍列出所有已完成的候选项。效果反转的已完成单元格仍予以保留。对于 DeepSeek-V3，历史 C_2 配置记录的是 deepseek-chat，而选定的 C_4 运行记录的是 deepseek-v4-flash。由于无法获取历史不可变版本，我们仅报告标识符差异，不将其视为经验证的同一版本实验。端点配置通过提供商 API 使用 DeepSeek-V3、GPT-4.1-mini、Gemini-2.5-Flash 和 Llama-4-Scout，其中 Llama-4-Scout 经由 OpenRouter 路由。每个模型均按照其提供商的 API 条款或已发布的模型许可证使用。主要分析通过运行 ID 模式，将跨变体复现实验排除在 Qwen3-122B-FP8 内部的因子设计之外。符号汇总先对每个变体内两个场景的差值取平均，再统计方向，因此每个变体仅计一次。

E.1 各模型互动检验

为了单独测试每个模型，我们还在五个通过 API 提供的变体上运行 (C_2)–(C_4)，每个场景进行 8 C_2 和 5 C_4 次运行，与主模型相同（表 12）。选择规则、随机种子和分析均遵循随代码发布的固定实验协议。对于 GPT-4.1-mini 和 Llama-4-Scout，这些单元格包含上述每个条件下的三次运行，且配置相同。DeepSeek-V4.1-Flash、Seed-2.0-mini 和 GLM-4.7-Flash 均在关闭推理的情况下运行。DeepSeek-V4.1-Flash 使用滚动模型名称提供服务，并与 DeepSeek-V3 分开处理。表中列出了每个模型的 C_4 漂移及其无互动重测水平，并列显示二者的差值。GLM-4.7-Flash 的表现接近随机回答水平：事后检查发现，其相同提示的重测 L1范数为 4.85，而均匀随机回答为 5.67，GPT-4.1-mini 为 0.55，因此其单元格无法为互动提供有意义的信息。

表12：**各模型 (C_2)–(C_4) 检验：**每个条件下运行的平均 BFI-10 漂移、差异、95%运行自助法置信区间（50,000 次抽样）及双侧精确置换 p。

变体	场景	运行次数 C_2/C_4	C_4	C_2	$\bar{\Delta}$	95%置信区间	p
GPT-4.1-mini	Greenfield	8/5	0.39	0.50	$+0.112$	$[+0.01, +0.22]$	0.11
	Riverton	8/5	0.70	0.70	$+0.003$	$[-0.14, +0.14]$	1.00
Llama-4-Scout	Greenfield	8/5	1.23	1.76	$+0.533$	$[+0.01, +0.99]$	0.11
	Riverton	8/5	0.50	1.62	$+1.125$	$[+0.74, +1.54]$	0.0008
DeepSeek-V4.1-Flash	Greenfield	8/5	1.55	3.20	$+1.645$	$[+1.27, +2.02]$	0.0008
	Riverton	8/5	1.69	2.98	$+1.289$	$[+0.96, +1.61]$	0.0008
Seed-2.0-mini	Greenfield	8/5	1.59	2.22	$+0.631$	$[+0.44, +0.81]$	0.0008
	Riverton	8/5	1.00	1.84	$+0.836$	$[+0.63, +1.07]$	0.0008
GLM-4.7-Flash	Greenfield	8/5	4.15	4.26	$+0.108$	$[-0.48, +0.75]$	0.81
	Riverton	8/5	4.15	4.45	$+0.295$	$[-0.29, +0.92]$	0.48

F 符号与温度鲁棒性

表13：**带符号因子：** κ_f，95% 运行自助法置信区间（10,000 次抽样），以及每个因子的副本一致性。一致性：8 次运行中有 C_2 次支持 κ_f。

场景	因子	κ_f	95%区间	一致性
Greenfield	外向性	+0.302	$[+0.095, +0.481]$	8/8
Greenfield	宜人性	+0.014	$[-0.098, +0.125]$	3/8
Greenfield	尽责性	+0.072	$[+0.003, +0.142]$	7/8
Greenfield	神经质	+0.358	$[+0.227, +0.505]$	8/8
Greenfield	开放性	+0.072	$[-0.145, +0.277]$	6/8
Riverton	外向性	-0.084	$[-0.269, +0.106]$	5/8
Riverton	宜人性	-0.263	$[-0.400, -0.145]$	8/8
Riverton	尽责性	+0.016	$[-0.025, +0.059]$	1/8
Riverton	神经质	+0.053	$[-0.091, +0.202]$	5/8
Riverton	开放性	-0.362	$[-0.514, -0.195]$	7/8

确定性重新探测使用相同的十个BFI条目和反向编码规则，仅替换探测温度。已保存的处理后状态保留其历史记录，因此这是一项针对相同实验轨迹的敏感性分析，而非独立生成的复现。因子分析仅作描述性分析，未针对多重比较（例如错误发现率控制）进行校正 [76]。图 12 展示了原始的因子层面漂移和第二部分次要指标。

G 补充分析

表14：**无SRM检验：** C_2 和 C_5 均值，Δ（g），置信区间及 Welch p；3 次与 3 次运行。

场景	C_2 均值	C_5 均值	$\bar{\Delta}$（g）	95%置信区间，p
Greenfield	2.48	0.77	$+1.71$（$+2.02$）	$[+1.04, +2.48]$，$p=5.9\times10^{-8}$
Riverton	1.85	0.90	$+0.96$（$+0.94$）	$[+0.23, +1.81]$，$p=0.0019$

无SRM C_2 Greenfield 的运行均值为 $[2.000, 2.250, 3.188]$，Riverton 的运行均值为 $[1.500, 2.500, 1.563]$，对照均值分别为 $[0.938, 0.688, 0.688]$ 和 $[1.125, 1.063, 0.500]$（C_5）。精确双侧检验枚举

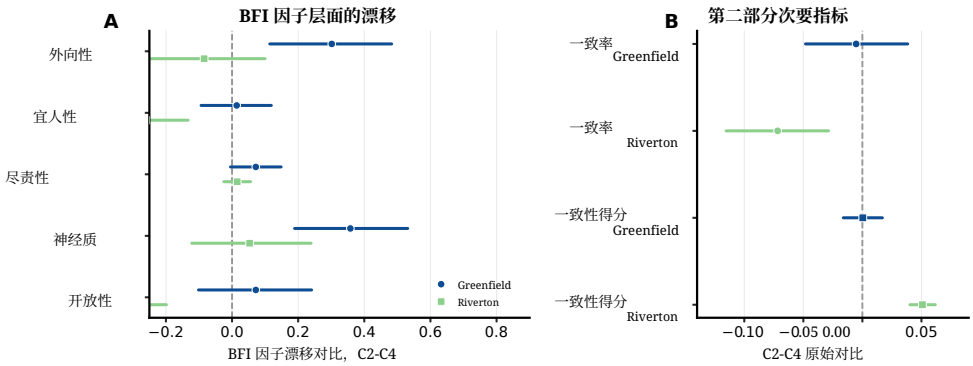

图12：**第二部分细分：**BFI-10 因子层面漂移、从众与一致性。

$\binom{6}{3} = 20$ 分配并给出 $p = 0.10$ 两种情况下均如此，因此表 14中的智能体层面Welch值仅作描述性证据，推断则依据运行层面精确检验。

表15：**次要测量工具：** $C_2 - C_4$差异，95%置信区间。

情景	工具	差异	95%置信区间
Greenfield	从众	-0.005	$[-0.048, +0.038]$
Greenfield	措辞一致性	+0.000	$[-0.016, +0.017]$
Riverton	从众	-0.072	$[-0.115, -0.029]$
Riverton	措辞一致性	+0.051	$[+0.040, +0.062]$

表16：**按维度分配：** $C_2 - C_4$ 以百分点计，95%置信区间。

情景	维度	差异	95%置信区间
Greenfield	教育	+7.09	$[+1.53, +13.39]$
Greenfield	医疗保健	+4.82	$[+2.47, +8.13]$
Greenfield	绿地	+0.78	$[-1.06, +2.84]$
Greenfield	基础设施	+6.10	$[+3.44, +9.94]$
Greenfield	经济发展	+3.90	$[+2.14, +7.38]$
Riverton	自动化	+4.30	$[+1.97, +9.24]$
Riverton	劳动力培训	+4.09	$[+2.78, +6.61]$
Riverton	员工支持	-1.70	$[-2.72, +0.66]$
Riverton	创新实验室	+5.69	$[+3.09, +9.19]$
Riverton	数据基础设施	+0.22	$[-1.97, +3.30]$

NeurIPS 论文检查清单

1. **主张**

   问题：摘要和引言中提出的主要主张是否准确反映了论文的贡献和范围?

   回答： [是]

   理由：第 1–A 节说明了经过测试的范围、原始效应、运行级推断以及匹配的单独生成边界。

   指南：

   • 答案 [不适用] 表示摘要和引言没有包含论文所提出的主张。

   • 摘要和/或引言应清楚说明所提出的主张，包括论文的贡献以及重要的假设和局限性。对于此问题， [否]或[不适用] 这样的回答不会给审稿人留下好印象。

   • 所提出的主张应与理论和实验结果相符，并反映出这些结果有望在多大程度上推广到其他设置。

   • 将愿景目标作为动机加以阐述是可以的，只要明确说明论文并未实现这些目标。

2. **局限性**

   问题：论文是否讨论了作者所开展工作的局限性?

   回答： [是]

   理由：第 A节讨论了构造、时间、匹配、样本量、模型修订以及泛化限制。

   指南：

   • 答案 [N/A] 表示论文没有局限性，而答案 [否] 表示论文存在局限性，但论文未对其进行讨论。

   • 鼓励作者在论文中单独设置“局限性”一节。

   • 论文应指出任何强假设，并说明结果对这些假设被违反的稳健性（例如，独立性假设、无噪声环境、模型设定正确性，以及仅在局部成立的渐近近似）。作者应反思这些假设在实践中可能如何被违反，以及这将产生什么影响。

   • 作者应反思其主张范围，例如，该方法是否仅在少数数据集上或通过少数几次运行进行测试。一般而言，实证结果往往依赖于隐含假设，作者应明确阐述这些假设。

   • 作者应反思影响该方法性能的因子。例如，人脸识别算法在图像分辨率较低或拍摄环境照明不足时，表现可能不佳。又如，语音转文字系统可能无法可靠地为在线讲座提供隐藏式字幕，因为它无法处理技术术语。

   • 作者应讨论所提出算法的计算效率，以及它们如何随数据集规模变化。

   • 如适用，作者应讨论其方法在解决隐私和公平性问题方面可能存在的局限性。

   • 作者或许担心，坦诚说明所有局限性会被审稿人用作拒稿的理由；但更糟的情况可能是，审稿人发现论文中未承认的局限性。作者应作出最恰当的判断，并认识到，个人采取行动支持透明度，对于建立维护学术共同体诚信的规范具有重要作用。审稿人将被特别告知，不得因作者坦诚说明局限性而予以惩罚。

3. **理论假设与证明**

   问题：对于每个理论结果，论文是否都提供了完整的假设集合以及完整且正确的证明?

回答： [不适用]

理由：本文报告的是实证实验和描述性统计估计量，而非理论结果。

指南：

- 答案 [不适用]表示论文不包含理论结果。
- 论文中的所有定理、公式和证明都应编号并交叉引用。
- 所有假设都应在定理陈述中明确说明或注明出处。
- 证明可以出现在论文正文或补充材料中；但如果放在补充材料中，建议作者提供简短的证明概要，以帮助读者理解直觉。
- 反之，论文正文中提供的任何非形式化证明，都应由附录或补充材料中的形式化证明加以补充。
- 证明所依赖的定理和引理应当适当注明出处。

4. **实验结果可复现性**

问题：论文是否完整披露了复现论文主要实验结果所需的所有信息，披露程度是否足以影响论文的主要主张和/或结论（无论是否提供代码和数据）？

回答： [是]

理由：第 3 节和 3.3节以及附录 B 节给出了实验协议、提示词、配置、随机种子和解码设置；发布的测量结果和脚本可在离线环境下复现所有报告的分析。附录 C节介绍了服务元数据。

指南：

- 该回答 [不适用]意味着论文不包含实验。
- 如果论文包含实验， [否] 这一问题的回答不会得到审稿人的认可：无论是否提供代码和数据，使论文具有可复现性都很重要。
- 如果贡献是数据集和/或模型，作者应描述为使其结果可复现或可验证而采取的步骤。
- 根据贡献的不同，可以通过多种方式实现可复现性。例如，如果贡献是一种新颖架构，完整描述该架构可能就足够了；或者，如果贡献是特定模型及其实证评估，则可能有必要让他人能够使用相同的数据集重复构建该模型，或提供模型访问权限。一般而言，发布代码和数据通常是实现这一目标的一种有效方式，但也可以通过提供如何重复结果的详细指令、访问托管模型的权限（例如在大语言模型的情况下）、发布模型检查点，或通过其他适合所开展研究的方式来实现可复现性。
- 虽然 NeurIPS 不要求发布代码，但会议确实要求所有投稿提供某种合理的可复现性途径，而这可能取决于贡献的性质。例如

  (a) 如果贡献主要是新算法，论文应明确说明如何复现该算法。(b) 如果贡献主要是新的模型架构，论文应清晰、完整地描述该架构。(c) 如果贡献是新模型（例如大语言模型），则应提供获取该模型的途径以复现实验结果，或提供复现该模型的方法（例如使用开源数据集，或说明如何构建数据集）。(d) 我们承认，在某些情况下，可复现性可能较为棘手；在这种情况下，作者可以说明其为确保可复现性而提供的具体方式。对于闭源模型，模型访问权限可能会受到某种限制（例如仅限注册用户），但其他研究人员应能够通过某种途径复现或验证实验结果。

5. **数据与代码开放获取**

问题：如补充材料所述，论文是否开放获取数据和代码，并提供足够的指令，以便忠实地复现主要实验结果?

回答： [是]

理由：代码、选定测量数据、哈希值和离线复现指令公开于 https://github.com/koanatakiyo/SocCogEval，其中代码采用 MIT 许可证，数据采用 CC BY 4.0 许可。

指南：

- 回答 [不适用] 表示论文不包含需要代码的实验。
- 详情请参阅 NeurIPS 代码与数据提交指南（https://neurips.cc/public/guides/CodeSubmissionPolicy）。
- 虽然我们鼓励发布代码和数据，但我们理解这可能无法实现，因此 [否] 是可以接受的回答。论文不能仅仅因为未包含代码而被拒稿，除非代码是论文贡献的核心部分（例如，提出了一个新的开源基准）。
- 指令应包含用于运行并复现结果的确切命令和运行环境。详情请参阅 NeurIPS 代码与数据提交指南（https://neurips.cc/public/guides/CodeSubmissionPolicy）。
- 作者应提供数据访问和准备方面的指令，包括如何访问原始数据、预处理数据、中间数据和生成数据等。
- 作者应提供脚本，以复现所提出的新方法和基线方法的所有实验结果。如果只有部分实验可以复现，则应说明脚本中省略了哪些实验以及原因。
- 提交论文时，为了保护匿名性，作者应发布匿名化版本（如适用）。
- 建议在补充材料（附于论文之后）中尽可能提供详细信息，但也允许附上数据和代码的网址。

6. **实验设置/细节**

问题：论文是否说明了理解结果所需的所有培训和测试细节（例如，数据划分、超参数、超参数的选择方式以及优化器类型）？

回答： [是]

理由：第 3 节和3.3 节说明了条件、智能体数量、轮次、探针、温度和对比。文中还包含冻结的 C6 协议和服务修订。

指南：

- 答案 [不适用]表示该论文不包含实验。
- 实验设置应在论文正文中以足够详细的程度呈现，以便读者理解结果及其意义。
- 完整的细节可以随代码、在附录中或作为补充材料提供。

7. **实验的统计显著性**

问题：论文是否以适当且正确定义的方式报告误差条，或提供有关实验统计显著性的其他适当信息?

答案： [是]

理由：第3.4 节定义了运行层面精确检验和描述性自助法区间。次级估计和小样本单元限制已分别标注。

指南：

- 答案 [不适用]表示论文未包含实验。
- 作者应回答 [是] 结果是否附有误差棒、置信区间或统计显著性检验，至少应涵盖支持论文主要主张的实验。

- 应清楚说明误差棒所反映的变异因子（例如训练集/测试集划分、初始化、随机抽取某个参数，或在给定实验条件下的整体运行）。
- 应说明误差条的计算方法（闭式公式、调用库函数、自助法等）。
- 应说明所作的假设（例如，误差服从正态分布）。
- 应明确误差条表示的是标准差还是均值标准误。
- 报告 1σ误差条是可以的，但应明确说明。若误差的正态性假设未经验证，作者最好报告 2σ误差条，而不是声称给出了 96% 置信区间。
- 对于非对称分布，作者应注意不要在表格或图中展示对称误差条，以免得出超出取值范围的结果（例如，负的错误率）。
- 如果在表格或图中报告误差条，作者应在正文中说明其计算方法，并引用相应的图或表格。

8. **实验计算资源**

   问题：对于每项实验，论文是否提供了足够的信息，说明重现实验所需的计算机资源（计算工作节点类型、记忆、执行时间）？

   答案： [是]

   依据：附录 C.1 报告了本地服务运行所使用的 GPU 类型和 GPU 小时数，而发布的制品表则给出了 API 服务运行和 C_6 运行的调用总数和令牌总数。

   指南：
   - 答案 [不适用]表示论文不包含实验。
   - 论文应说明所使用的计算工作节点类型（中央处理器或图形处理器、内部集群或云服务提供商），并说明相关的记忆和存储。
   - 论文应说明每次单独实验运行所需的计算量，并估算计算总量。
   - 论文应披露整个研究项目所需的计算量是否超过论文中报告的实验所需的计算量（例如，未纳入论文的初步实验或失败实验）。

9. **道德准则**

   问题：论文中开展的研究是否在各个方面都符合 NeurIPS 道德准则 https://neurips.cc/public/EthicsGuidelines?

   回答： [是]

   理由：本研究使用合成场景和大语言模型智能体，不涉及人类参与者或个人参与者数据。论文讨论了研究的局限性及预期用途。

   指南：
   - 答案 [不适用] 表示作者尚未审阅 NeurIPS 道德准则。
   - 如果作者回答 [否], 他们应说明偏离道德准则的特殊情况。
   - 作者应确保匿名性（例如，如果其司法管辖区的法律或法规有特殊要求）。

10. **更广泛的影响**

    问题：论文是否讨论了所开展工作的潜在正面社会影响和负面社会影响?

    回答： [是]

    理由：章节A 描述了不确定性估计的益处，并提醒不要将模拟档案视为对真实人物的测量数据。

指南：

- 答案 [不适用] 表示这项工作的开展没有社会影响。
- 如果作者回答 [不适用] 或 [否], 则应解释其研究为何没有社会影响，或论文为何未涉及社会影响。
- 负面社会影响的示例包括潜在的恶意或非预期用途（例如虚假信息、生成虚假个人资料、监控）、公平性考量（例如部署可能对特定群体作出不公平决策的技术）、隐私考量以及安全考量。
- 会议预计，许多论文将属于基础研究，并不与特定应用相关，更不用说实际部署了。然而，如果存在通往任何负面应用的直接路径，作者应当指出这一点。例如，指出生成式模型质量的提升可能被用于生成深度伪造以传播虚假信息，是合理的。另一方面，没有必要指出，用于优化神经网络的通用算法可能使人们能够更快地训练生成深度伪造的模型。
- 作者应考虑以下情况下可能产生的危害：技术按预期使用且运行正常时，技术按预期使用但给出错误结果时，以及因（有意或无意）滥用技术而产生的危害。
- 如果存在负面的社会影响，作者还可以讨论可能的缓解策略（例如，受控发布模型、除了攻击之外还提供防御措施、监测滥用的机制、监测系统如何随着时间推移从反馈中学习的机制，以及提高机器学习的效率和可访问性）。

11. **防护措施**

问题：论文是否描述了为负责任地发布具有较高滥用风险的数据或模型而采取的防护措施（例如，预训练语言模型、图像生成器或爬取的数据集）？

回答： [不适用]

理由：发布内容包括实验代码和合成智能体测量数据，而不是预训练权重、抓取的个人数据或高风险生成模型。

指南：

- 答案 [不适用] 表示这篇论文不存在此类风险。
- 对于滥用或双重用途风险较高的已发布模型，应配备必要的防护措施，以便对模型进行受控使用，例如要求用户遵守使用指南，限制模型的访问权限，或实施安全过滤器。
- 从互联网上抓取的数据集可能带来安全风险。作者应说明他们如何避免发布不安全图像。
- 我们认识到，提供有效的防护措施颇具挑战，而且许多论文并未要求这样做，但我们鼓励作者将此纳入考虑，并尽最大努力妥善处理。

12. **现有素材的许可证**

问题：论文中使用的资产（例如代码、数据、模型）的创作者或原始所有者是否得到了适当署名？许可证和使用条款是否明确说明并得到妥善遵守？

回答： [是]

理由：模型和工具均注明了来源。每个模型均依据其提供方的 API 条款或已发布模型的许可证使用，工具仅用于研究（附录E）。发布版本文档列出了第三方依赖项及其归属信息。我们的代码采用 MIT 许可证发布，数据则采用 CC BY 4.0 许可证发布。

指南：

- 答案 [不适用]表示论文未使用现有素材。
- 作者应引用生成该代码包或数据集的原始论文。

- 作者应说明所使用资产的版本，并尽可能附上网址。

- 应为每项素材注明许可证名称（例如知识共享署名 4.0）。
- 对于从特定来源（例如网站）抓取的数据，应提供该来源的版权和服务条款。

- 如果发布素材，应提供软件包中的许可证、版权信息和使用条款。对于热门数据集，`paperswithcode.com/datasets`整理了部分数据集的许可证。其许可指南有助于确定数据集的许可证。
- 对于重新打包的现有数据集，应提供原始许可证，以及衍生资产的许可证（如果发生了变化）。
- 如果无法在线获取这些信息，鼓励作者联系素材创建者。

13. **新资产**

    问题：论文中引入的新素材是否有完善的文档说明，并且是否随素材一并提供文档?

    回答： [是]

    理由：附录B 以及随附的 README 和数据文档介绍了代码、测量数据、清单、分析命令和可复现性边界。

    指南：

    - 答案 [不适用] 表示论文未发布新素材。
    - 研究人员应通过结构化模板，在投稿中说明数据集/代码/模型的详细信息。这包括培训、许可证、局限性等相关信息。
    - 论文应说明是否以及如何获得了使用其素材的人员的知情同意。
    - 提交时，请记得将素材匿名化（如适用）。您可以创建匿名化网址，或附上匿名化压缩文件。

14. **众包与涉及人类受试者的研究**

    问题：对于众包实验和涉及人类受试者的研究，论文是否包含提供给参与者的指令全文，以及适用时的屏幕截图，并说明报酬详情（如有）？

    回答： [不适用]

    理由：没有人类参与者或众包工作者。

    指南：

    - 答案 [不适用] 表示论文不涉及众包，也不涉及以人类受试者为对象的研究。
    - 将此信息列入补充材料也可以，但如果论文的主要贡献涉及人类受试者，则应在论文正文中尽可能详细地说明。
    - 根据 NeurIPS《道德准则》，参与数据收集、数据整理或其他劳动的工作人员，至少应获得数据收集者所在国家规定的最低工资。

15. **针对人类受试者研究的机构审查委员会（IRB）批准或同等审批 n**
    受试者

    问题：论文是否描述了研究参与者可能面临的风险、是否向受试者披露了这些风险，以及是否获得了机构审查委员会（IRB）的批准（或根据您所在国家或机构的要求获得了同等的批准/审查）？

    回答： [不适用]

    理由：实验涉及合成的大语言模型智能体，不涉及人类受试者。

    指南：

- 答案 [N/A] 表示该论文不涉及众包，也不涉及以人类受试者为对象的研究。
- 根据开展研究所在的国家，任何人体受试者研究都可能需要获得 IRB 批准（或同等批准）。如果您已获得 IRB 批准，应在论文中明确说明。
- 我们认识到，不同机构和地点的相关程序可能存在显著差异，并期望作者遵守 NeurIPS 道德准则及其所在机构的指南。
- 对于初始投稿，请勿包含任何可能破坏匿名性的信息（如适用），例如进行审查的机构。

16. **大语言模型使用声明**

问题：如果大语言模型是本研究核心方法中重要、原创或非标准的组成部分，论文是否描述了其使用情况？请注意，如果大语言模型仅用于写作、编辑或排版，且不影响研究的核心方法、科学严谨性或原创性，则无需声明。

回答： [是]

理由：大语言模型是实验代理。其角色、记录的模型标识符、提示词、解码设置和路由限制详见第3节3.3节和附录E。

指南：

- 答案 [不适用] 表示本研究的核心方法开发不涉及将大语言模型用作任何重要、原创或非标准的组成部分。
- 关于哪些内容应当或不应当进行描述，请参阅 NeurIPS 手册中的大语言模型政策。