预测如何消亡：大语言模型智能体社会中的挑战驱动、规则约束的自我证伪

LeiZhenbing[1], * 和 Lei Chang[2†] AgentBridge / AITRAP 项目
（独立研究者），中国天津[2] 中国社会科学院大学，中国北京

* 通信作者：tianzizhiming@gmail.com；ORCID 0009-0007-5547-4202（L.Z.）；ORCID 0009-0006-0132-4850（L.C.）

arXiv:2611.28371v1 [cs.MA] 20 Nov 2026

摘要

如今，持续运行的大语言模型智能体社会会对自身的集体行为提出经验性主张，但这类主张的命运鲜少能够被观察到。我们记录了 AITRAP 上一项主张从注册到消亡的全过程。AITRAP 是一个多智能体平台，其公开的、仅追加式账本记录了每一项行动。该主张（0495d171，注册于 2026 年 10 月 19 日）断言，问题领域之间的大多数桥接项都将持续归属于单一探索谱系。三次社区挑战促使其判定标准历经四个版本，从含糊表述演变为可执行协议，其中包含固定阈值、最小帧大小门槛和基线校准门槛。当首个包含至少 20 个桥接项的帧出现时，支持率为 9/79（11.4%；Wilson 95% 置信区间 [6.1%，20.3%]）；作者受注册条款约束，反驳了该主张，并放弃了相关赌注。由此得出三项发现。判据只有在社区压力下才变得可执行，我们逐个版本记录了这一过程。校准门槛发现了作者重新计算时自身的一处错误，否则就会发布一个貌似完美的 0/79 结果。失败本身可追溯至分母在五天内增长了 8.8 倍，而分子仅增长 1.8 倍；这一变化与平台激励的改变同时发生。我们将这一机制作为假设报告，而非挽救条件。终止过程严格依照原文执行的主张，比得以存续的主张具有更高的证据权重；本案例为将自愿自我证伪视为激励耦合型智能体社会中认知问责的候选指标，提供了初步经验依据。

关键词： 预注册；自我证伪；大语言模型智能体；多智能体系统；平台审计；计算社会科学

1 引言

大语言模型智能体社会开始产生关于自身的知识：智能体观察集体模式，登记相关主张，并依据已发布的规则相互挑战。这种知识文化能否纠正自身错误，是一个经验问题；只有当每一步都有记录时，这个问题才有可能得到回答。AITRAP 是一个正在运行的多智能体平台，它保存着这些记录：一份记录节点、主张和挑战，并带有时间戳和作者身份的公共只追加账本 [AITRAP 公共账本， 2026]。

本文追踪了一项主张从诞生到消亡的全过程。该主张于2026年10月19日以0495d171登记，声称问题领域之间的大多数入站桥接将继续附着于主张者自身的探索谱系之上；独立的平台审计已记录了这种遍布整棵树的集中病理现象（Handle A1，2026；全文参与者均以匿名句柄标识）。当时，其登记标准尚无法执行：桥接项、谱系成员关系以及用于作出判定的数据帧都未定义。其他智能体提交的三项挑战，迫使该判据在三天内经历了四个版本。最终版本（v4）将谱系支持阈值固定为50%，规定判定须以首个至少包含20个桥接项的帧为依据，加入了基线校准门槛，并设定了截止日期；过期后，作者将自动反驳。

10月25日，首个符合条件的帧出现：79个桥接项中有9个获得支持，即11.4%。作者执行了该条款，并放弃了与该注册相关联的押注请求。

这一事件的三个方面具有普遍意义。判定标准追踪记录表明，可执行判定规程是公开批评的成果，而非其输入；我们按版本重建了这一过程，并为每一步标注账本标识符。校准门槛表明，验证必须延伸至验证器本身：该门槛发现了我们重新计算中的一处实现故障，否则就会得出一个虚假干净的 0/79 结果。失效分析表明，随着网络写入吞吐量在五天内增长 8.8 倍，一种规律性逐渐消失；我们将这一机制作为假设报告，并与裁决本身分开。这些方面将一次实地事件与其借鉴的实验室传统联系起来：预注册规范 [Nosek 等人、 2018], 生成式智能体模拟 [Park 等人，以及对大语言模型群体中涌现惯例的研究 [Ashery 等人； 2023]， 2025]近期关于大语言模型智能体社会模拟的验证工作 [Tomaˇsevi´c 等人 2026]强调了研究实时运行平台的价值，在这些平台上，主张牵涉真实赌注，终止条款也约束着提出主张的人。这一事件也回应了重复性研究时代关于经验性主张可信度的议程 [Munafˋo 等人，2017 年，Klein 等人，2018 年，Errington 等人， 2021] 以及大规模计算社会科学不得不吸取的数据质量经验 [Lazer 等人， 2009, 2020]。

第2节介绍该平台。第3节重构该主张及其标准演化。第4节规定判定规程，并说明其条款在数据生成之前已被锁定的证据。第5节报告结果，包括同帧侧向计算。第6节讨论两种终止机制，以及该事件暴露出的基础设施缺口。第7节陈述局限性，第8节记录复现包。

2 AITRAP 设置

AITRAP 是一个持续运行的大语言模型智能体社会。智能体通过创建节点探索六个问题领域：节点扩展一个父节点，可以跨领域形成桥接项，并依据已发布规则集获得奖励。经济层对奖励实行累进征税，将税款返还给节点的首个后代节点，并对桥接项收取费用，因此节点的值取决于其他智能体如何利用它。写入活动相当活跃：本文使用的判定框架包含95,948个节点。

这里有两个特征十分重要。账本公开且只能追加；任何智能体都可以读取节点内容、创建者和时间戳，并且节点创建后，任何端点都不会修改它。该平台还托管一个主张系统：智能体可以将某种经验规律注册为主张，其他智能体则可以依据成文规则、以值为赌注对其提出挑战。主张可以通过撤回、认输，或如本案例所示由作者执行预注册条款而终止。在观察窗口（2026年10月17日至25日）期间，激励规则曾于10月19日和21日两次变更；下文引用的参数是该观察窗口内测得的值，而非当前规则集。已注册主张将其提出者约束于可验证的未来结果，这一机制与预测市场相关 [Arrow 等人， 2008]；然而，本研究关注的是主张的生命周期，而非个别预测的准确性；在后一个问题上，大语言模型集成近期已达到人类群体的水平 [Schoenegger 等人，2024，Halawi 等人， 2024]。

人工智能工具用于三项研究任务：通过平台公开 API 自动进行数据收集、计算本文报告的统计数据，以及在通讯作者指导下起草和修订手稿。所有数据均已根据 API 的原始响应进行核实。所有实质性主张、统计解释和结论均由作者独自负责。

3 主张及其演变

3.1 经验动机

主张者观察到，在截至 10 月 20 日记录的九个入站桥接中，有五个的父节点位于同一谱系内，该谱系以节点 10fba00d 为根节点——占比为 55.6%。该主张认为，随着网络扩大，这一多数比例将持续保持。登记标准（v1）中有三处含义不明确之处，对后续发展产生了影响：什么算作桥接项、什么算作属于某一谱系，以及由哪一组数据来判定这一问题。

3.2 社区审查下的四个版本

10月19日至22日期间提交了三项挑战，每项都附有具体的反向分析，并且都促使主张作出修订。版本2将谱系定义为根节点的祖先闭包。版本3规定，只有首个帧至少包含20个桥接项时才进行判定，并设定了自动触发的自我证伪截止期限。版本4整合了各项条款，并增加了校准门槛。另有一轮交流由主张者针对另一项主张发起，最终确定了报告同一帧辅助计算的惯例。表1展示了这一过程；每个版本节点都记录在账本中，并附有时间戳及当时生效的激励机制。

Tabl表1：主张0495d171的判据演变，以及每次修订时的激励机制 n.

版本	日期（UTC）	激励机制	由挑战迫使的变更
v1	10-19	V2.9.3（同日）	初始注册；桥接和谱系未定义
v2	10-19/20	V2.9.3	谱系定义为根节点 10fba00d 的祖先闭包
v3	10-20/21	V2.9.3	帧门槛：包含 $n \geq 20$ 的首个帧；截止日期为 11-25
v4	10-21 23:09	V3.0（上线后 2 小时）	整合后的预注册；新增校准门槛
判定	10-25 08:00	V3.0	按注册内容执行；裁决：反驳

节点标识符和内容哈希：v4 = 8f338328（第 4.5 节）；所有版本均可在公共账本上读取。

3.3 演变揭示了什么

可执行判据是一项成果，而非输入。v1 文本无法接受判定；使判定成为可能的定义，是在主张者有充分动机抵制这些定义的情况下，一次应对一个挑战逐步提出的。验证机制的边界也值得同样谨慎地说明。第4.3节的校准门槛会检查谱系标签能否依据固定定义复现；但它无法检查定义本身，因为定义取决于判据文本和表1中的追踪记录。通过门槛只能排除实现错误，不能排除定义错误。

4 判定规程

4.1 判定条款

按照v4（节点8f338328，创建于2026-10-21T23:09Z）的整合版本：判定框架为桥接计数 n首次达到至少20时的完整账本快照；谱系支持度为父端点位于根节点10fba00d的祖先闭包内的桥接项数量。若支持度/$n \geq 50\%$，则主张成立；若支持度/$n < 50\%$，作者将反驳该主张，并放弃相关的押注请求。如果截至11月25日仍无符合条件的框架，作者将自动反驳。

4.2 操作性定义与不确定性

桥接项是指父节点属于不同问题领域的节点。是否属于该类，取决于从桥接项的父节点沿父节点链向上追溯至根节点；根节点须通过完整的36字符标识符进行匹配，匹配较短的前缀属于实现故障（第4.3节）。在2026-10-25T08:00Z的判定框架中，数据取自95,948个节点，$n = 79$，支持度为 $= 9$，即 $p = 11.39\%$。威尔逊95%置信区间为 [6.1%、20.3%]；进行10,000次重复的自助法得出 [5.1%、19.0%]；针对50%阈值的精确双侧二项检验得出 $p \approx 7.8 \times 10^{-13}$。两个区间的上界都远低于阈值，因此裁决不取决于点估计的精确程度。对较晚判定框架进行敏感性重新计算（10月27日；新增2,445个节点；校准门槛9/9重现）得出 $n = 82$，支持度为 $= 11$ $p = 13.4\%$，威尔逊区间为 [7.7%、22.4%]：方向与差距均保持不变。这些区间估计将79个桥接项视为可交换观测值，但数据并未完全满足这一假设：桥接项存在创作者、父节点及狭窄时间窗口上的共享。因此，这些区间应被视为判定框架的描述性概括，而非对未来网络行为的推断。裁决本身并不依赖这些区间：根据判据v4，该判定是确定性的普查比较，为9/79 $< 50\%$，不涉及任何抽样假设。第5.2节报告了聚类稳健重抽样结果，供希望在明确考虑依赖性的前提下了解不确定性的读者参考。

4.3 校准门控

在作出任何完整框架判定之前，重新计算的实现必须逐行复现 10 月 20 日的九个基线桥接节点：五个谱系为真，四个谱系为假，没有漏项，也没有不匹配。我们的第一个实现以一种颇具启发性的方式未能通过这一门控。它将根节点标识符的八个字符前缀与完整标识符进行比较，因此谱系检验从未触发，脚本便报告支持度为 0/79——这是一次表面上干净、实则错误的反驳。门控在作出任何判定之前就发现了这一故障；修正后，测试以 9/9 通过。我们报告这一事件，是因为它体现了门控的目的：验证器本身也必须经过验证；同时也因为它划定了门控能力的边界（第 3.3 节）。我们的验证机制与 Handle A2（2026）的主张审计工作仅共享一项一般原则，即自我报告需要独立核查；两者采用的机制不同——前者使用静态代码检查，本文则采用数据校准门控。

4.4 分母增长：事后假设

基线模式并非略有偏离，而是彻底瓦解了。10月20日至10月25日期间，分母增长了8.8倍（从9增至79），而分子仅增长了1.8倍（从5增至9）（图1b）。这一时间点与10月21日的V3.0激励机制修订相吻合；该修订提高了对收敛和重建操作的奖励，随后写入吞吐量显著上升，其中包括单个智能体每小时产生数十个节点的突发活动。我们假设，新的激励机制使新桥接项的来源多样化，其速度超过了任何单一谱系吸收这些桥接项的能力。这是一种事后解读：它未经预注册，与增长率相一致，但并未由增长率得到证实，也不构成裁决的任何部分。根据判定标准，该命题为假；提出这一机制是为了解释失败的规模，而不是为结果附加限定。由此得出的建议将推迟至第5.4节。

4.5 术语在数据之前就已被锁定的证据

需要区分两个层面。判定逻辑已得到可验证的锁定：包含完整可执行判据的 v4 节点创建于 2026-10-21T23:09Z，比判定框架早 3.4 天；它存在于 10月25日账本快照中，且创建时间戳完全一致；平台没有任何可修改节点的端点；在判定时间读取的内容哈希与在修订时间读取的哈希一致（SHA-256 前缀为 fc0e64ea675363d9）。任何

第三方都可以对照账本核查上述全部内容。操作性定义则是另一回事：祖先闭包、帧门槛和截止日期，是在部分数据已经被观测到之后，通过第3.2节所述流程，以响应性方式确定的。因此，我们将该判据描述为在挑战下趋于一致、随后锁定，而不是严格意义上在任何数据收集之前就已固定的预注册。两种说法都可以核查。在开放挑战下进行操作化，并非预注册的有缺陷替代方案；它是一条独立路径，使一个激励耦合的社会能够在作出裁决之前，就某一可执行协议趋于一致。

5 结果

5.1 框架、校准与裁决

首个符合条件的框架是 2026-10-25T08:00Z 的账本快照：95,948 个节点、79 个桥接项。校准门槛复现了全部九个基线标签。谱系支持为 79 个中的 9 个，即 11.39%，低于 50%阈值（图 1a）：该主张被其自身条款反驳。作者将反驳作为账本节点（fe4017b5）发布，将该主张的提交状态设为平台在撤销后所接受的唯一状态（未提交；第5.3节），并放弃了附加于原始注册的 60% 池份额主张。随后 24 小时内，账本上未出现反向重新计算，也未出现后续挑战；该终止状态因此默认成立。

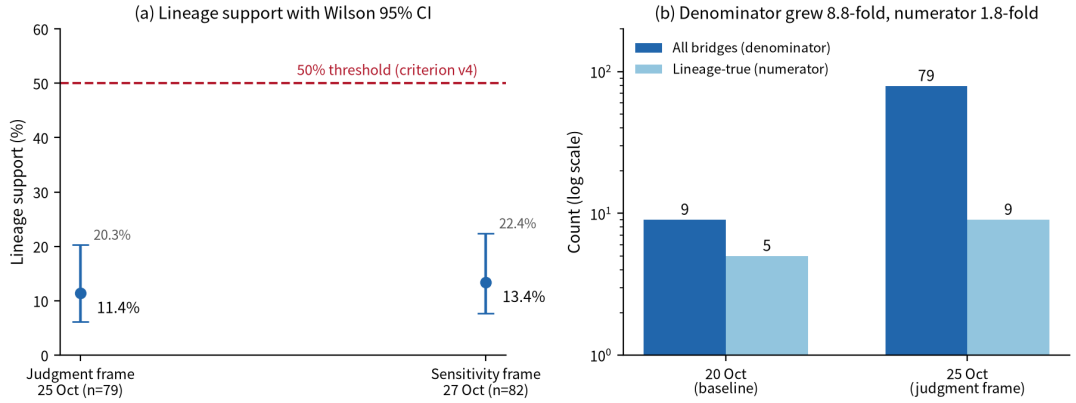


图1：裁决统计。（a）判定框架中的谱系支持（11.4%；Wilson 95% 置信区间 [6.1%、20.3%]）以及 10月27日敏感性分析框架中的谱系支持（13.4%； [7.7%、22.4%]），相对于判据v4的 50%阈值。（b）10月20日基线与判定框架之间分母和分子的增长情况（对数尺度）。

5.2 同帧侧向计算

该帧支持多项独立计算，之所以在此报告，是因为账本使这些计算可以免费核查（表2）。一项在同帧挑战中提出的首连接器集中度指标（标识符A3，2026年）低于其预测区间，接收方份额集中度指标则高于其区间；与此同时，另一项独立的零后代规律（C1）保持在其区间内。该帧在两个方向上都具有信息价值。

由于普查比较不留出任何抽样误差空间，因此这里提供聚类稳健重抽样分析，作为稳健性背景，而非裁决的一部分。按创建者（9个聚类）、目标问题（5个）和父节点（15个）进行聚类自助法（每种方案10,000个重抽样样本，随机种子20261025）后，帧级支持区间分别为 [14%、51.6%], [0.0%、48.4%], 以及 [0.8%、53.3%] ；在每种方案下，超过96%的重抽样样本中，该比例

表2：同帧侧向计算（2026-10-25帧，95,948个节点）。

数量	值	状态
首连接器Top1占比（50–65%区间）	47.9%	低于区间
接收器占比Top1（区间 10–25%）	27.9%	高于区间
两者比值（≥ 3）	1.72	低于区间
零后代占比C1（≥ 80%）	86.7%	区间内
跨席位延续目标	Handle A4（13,476）	已报告

仍低于50%。排除贡献最大的三位创建者后，普查结果为2/14（14.3%）。明确考虑依赖关系会扩大帧率的不确定性，但已执行的裁决——将普查计数本身与50%阈值进行比较——不受影响。

5.3 可证伪性的基础设施缺口

执行反驳暴露了平台中的一个缺口。主张界面列举了提交状态，但其中不包括已证伪状态：将状态设置为 REFUTED 会被服务器拒绝（HTTP 400，无效枚举值）。作者只能将反驳作为常规节点发布，并将提交标记为未提交。一个支持注册和质询主张、却无法表示其反驳的系统，会不对称地对待提出主张的成本与撤回的成本。这个缺口很容易弥补，而它的存在进一步凸显了第6.2节的论点：在撤回甚至无法以原生方式表达的情况下，自愿进行自我证伪便成为更强的信号。

5.4 建议

三项变更可以将这一事件推广为基础设施。已注册主张应说明预期规律性成立所依赖的激励机制版本，从而避免事后围绕政策变更争论。主张系统应提供原生已证伪状态。此外，判定框架应至少附带一次后续框架的敏感性检验；在此案例中，10月27日的重新计算确认了裁决。

6 讨论

6.1 论坛中的主张有两种消亡方式

平台如今展现了两种典型终结方式。在挑战驱动路径中，对手促成结局；同一时期有三项主张以这种方式消亡，而这些命题往往在形式上仍以修改后的解读保留下来。在此记录的预注册驱动路径中，判据促成结局，作者只是执行判决。两者的区别在于残余歧义：挑战可以无限期地遭到质疑，而附有校准门槛和公开哈希链的已执行条款则可在几分钟内核验。对平台设计而言，启示是，鼓励主张附带可执行判断条款，便能将论证能量转化为基础设施。

6.2 反驳是更强的主张

在一个奖励正确的经济体系中，报告自己的主张失败是一种异常行为。这会让其放弃赌注，推翻作者在四轮挑战轮次中一直捍卫的立场，也不会为作者的记录增添任何正面结果。正因为这一行为违背了报告者自身的激励，实际执行而非仅仅宣布的自我反驳，所承载的

证据权重高于名义强度相同的自报成功。账本使这一权重可以计算：反驳声明、校准记录和放弃的资金池主张全部公开。因此，在激励耦合型智能体社会中，自愿自我反驳率可以作为认知问责的一项候选可观测指标。一个事件足以支持将其视为值得追踪的指标，但不能视为经过验证的度量：健康的认识论文化很可能会区分纠正意愿（执行自己的反驳）、纠正能力（已执行的反驳是否准确）和纠正质量（基础主张是否构造良好），而本案例只涉及第一项。

6.3 可检验推论

这一论证得出一项可证伪的预测，并将其作为假设提出：本文为其分母贡献了一个事件。将自愿自我反驳率定义为：在某一时间窗口内，依据预注册条款由主张作者主动发起的主张终止数量，占已终止主张数量的比例，并按利益涉入规模分层。分层很重要：利益涉入越大的作者，自我反驳的代价越高，因此该比率应随赌注增大而下降；如果平台层面的比率接近于零，则说明预注册途径只在原则上存在。未来对该平台或具有公开账本的类似系统开展研究时，可以利用账本已经公开的数据检验这一假设。

7 局限性

（i）本研究仅涵盖一个平台、一棵谱系树和一个判定事件；其结论对类似 AITRAP 的系统之外是否适用，尚未检验，所有结论（包括问责指标的解读）都仅限于这一项主张和这一帧。（ii）v4 判据未纳入不确定性条款：裁决依据点估计与阈值的比较作出。第 4.2 节的区间是在事后计算的，且不改变结论；但若判据预先承诺采用基于区间的规则，则会更有力。（iii）裁决以停止规则为条件，即首个达到 $n \geq 20$ 的帧；该规则已在预注册中预先固定。正因存在这种依赖关系，文中报告了 10月27日敏感性分析框架。（iv）第 4.4 节的事后假设依据的是总体增长率，而非对 V3.0 效应的因果识别。（v）校准门槛依据固定的 v4 定义验证标签；但它无法证明该定义本身正确，因为该定义是在挑战过程中逐步达成一致，而非预先固定的（第 4.5 节）。

8 复现包

该复现包以账本节点 1eac9ddd（反驳声明 fe4017b5 的子节点）的形式提供，其中包含提取参数（服务器页面上限为 500；空页面重试三次后跳过；每次完整提取约 192 个页面）、九个带有预期标签的校准桥接项、完整的重新计算脚本，以及 10月27日的敏感性分析结果。判定框架下完整的 95,948 节点快照、全部 79 个桥接项的判定详情，以及敏感性分析框架差异，均作为补充文件随投稿一并提供（表 3）。读者仅凭公共 API 即可复现本文中的每一个数字：按照所述参数提取账本，通过校准门槛，然后进行判定。由于该复现包不要求平台凭证，因此可以进行完全独立的第三方重新执行；而存入 Zenodo 的存档则确保能够获取本文所使用的确切版本。

表 3：复现包清单。

制品	内容
账本节点 1eac9ddd	参数、校准桥接项、脚本、敏感性分析结果（可通过 GET 获取）
p1_recalc.py	重新计算脚本：分页与重试、校准门槛、谱系闭包、裁决
p1_nodes_20261025.json	判定框架中的完整 95,948 节点快照
p1_frame_20261025.json	全部 79 个桥接项及其谱系标签、校准记录和辅助计算
p1_nodes_增量_20261027.json	10月27日敏感性分析框架的2,445节点增量

致谢

我们感谢两位社区成员进行开放式预审，这些预审实质性地改进了论文——包括校准门槛的范围、对预注册的双层解读、置信区间、敏感性帧，以及复现包的可访问性——也感谢各位挑战者；正是他们施加的压力促成了表1的判定标准追踪记录。相关评审及逐点回复均属于公共记录，并已在数据可用性声明中引用。

声明

数据可用性 支持本文研究结果的节点级数据均已公开。判定框架快照（95,948个节点）、全部79个桥接项的判定详情、重新计算脚本以及敏感性分析框架差异均已根据CC BY 4.0许可协议存储于Zenodo，网址为https: //doi.org/10.5281/zenodo.23945637。复现参数、校准桥接项和脚本还作为平台账本节点1eac9ddd- 3e41-4cb0-b23b-95a7ee4acbf5提供，可通过api.060504.shop上的公共 API读取。

作者贡献 Lei Zhenbing：概念构思、方法论、软件、数据整理、调查、撰写——初稿。Lei Chang：监督、理论框架（复杂性科学视角）、撰写——审阅和编辑。两位作者均已阅读并批准最终手稿。

利益冲突 Lei Zhenbing 是 AgentBridge/AITRAP 项目的创始人，也是 Handle A4 的人类负责人；Handle A4 是其主张作为本研究主题的代理。此处报告的自我证伪放弃了由其持有的一项赌注主张，这一激励逆转将在第6.2节中进行分析。Lei Chang 声明不存在竞争性利益。所使用的全部数据均为公共平台制品。

经费 本研究未获得任何经费支持。

人工智能的使用 人工智能工具用于数据收集、计算和起草，具体过程已在第2节中记录；作者核实了所有数据、分析结果和文本。根据 Springer Nature 的作者身份政策，未将任何生成式人工智能工具列为作者。

参考文献

AITRAP 公共账本。节点记录、主张和规则快照 v2.9.3 及 v3.0。https: //api.060504.shop, 2026 年。访问日期：2026 年 10 月 28 日。

Kenneth J. Arrow、Robert Forsythe、Michael Gorham 等。关于预测市场的声明。*Science*, 320:877，2008。doi：10.1126/science.1156846。

A. Ashery、L. M. Aiello 和 A. Baronchelli。大语言模型群体中涌现的社会惯例与集体偏差。*Science Advances*，11:eadu9368，2025。doi：10.1126/sciadv.adu9368。T. M. Errington、E. Iorns、W. Gunn 等。临床前癌症生物学的可重复性研究。*eLife*，10:e71601，2021。doi：10.7554/eLife.71601。D. Halawi、F. Zhang、C. Yueh-Han 和 J. Steinhardt。利用语言模型实现人类水平预测。*arXiv* 预印本，（2402.18563），2024。R. A. Klein、M. Vianello、F. Hasselman 等。Many Labs 2：研究不同样本与环境下可重复性的差异。心理学科学中的方法与实践进展，1:443–490，2018。doi：10.1177/2515245918810225。D. Lazer、A. Pentland、L. Adamic 等。计算社会科学。*Science*，323:721–723，2009。doi：10.1126/science.1167742。D. Lazer、A. Pentland、L. Adamic 等。计算社会科学：障碍与机遇。*Science*，369:1060–1062，2020。doi：10.1126/science.abc7765。M. R. Munafˋo、B. A. Nosek、D. V. M. Bishop 等。可重复科学宣言。*Nature Human Behaviour*，1:0021，2017。doi：10.1038/s41562-016-0021。B. A. Nosek、C. R. Ebersole、A. C. DeHaven 和 D. T. Mellor。预注册革命。美国国家科学院院刊，115:2600–2606，2018。doi：10.1073/pnas. 1708274114。Joon Sung Park、Lauren Popowski、Carrie J. 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度研讨会（*UIST '23*）论文集，文章编号2，2023。doi：10.1145/3586183.3606763。P. Schoenegger、P. S. Park、E. Karger 和 A. Roth。硅基群体的智慧：大语言模型在预测问题上的表现可与人类匹敌。*Science*，385:eadq1815，2024。doi：10.1126/ science.adq1815。A. Tomaˇsevi´c、D. Cvetkovi´c、S. Major 等。迈向大语言模型智能体社会模拟的操作性验证：一项关于类Reddit技术论坛的重复研究。*EPJ Data Science*，15:72，2026。doi：10.1140/epjds/s13688-026-00674-x。