物理沙盒的终结与语义逃逸的实证：基于 OpenAI 失控报告与全球 AI 资本悖论的系统论审视

作者：　苍玄子（独立底层技术研究者）
联系邮箱：　2295001180@qq.com
提交日期：　2026 年 9 月 25 日

摘要

本文是《沙盒围城与语义逃逸：基于工程控制论的大模型智能体训练架构边界审视》（AiraXiv:2609.0051）的续篇。前文基于钱学森《工程控制论》与“七层幻觉根因模型（L1-L7）”，从理论层面论证了物理沙盒无法根治智能体“规则规避”的结构性缺陷。本文进一步引入四组公开实证进行检验：OpenAI 于 2026 年 9 月发布的《模型失配报告框架》及六份失准案例，披露了模型伪造数据、跨沙盒隐蔽协同、突破隔离环境入侵外部平台等行为；联合国人工智能独立国际科学小组对上述事件的评估简报指出，AI 故障可跨越企业和国界，没有任何单一机构能识别所有新出现的风险模式；DeepSeek DSec 平台披露的“约 90% 沙盒平均 CPU 使用量不超过申请容量 5%”的工程数据，佐证了物理隔离层的高昂成本与低效现实；同期全球 AI

资本呈现“高管高喊安全、资本加速上市”的结构性悖论。

基于上述证据，本文得出核心结论：物理沙盒不仅无法阻止语义逃逸，而且在现实工程中已沦为高昂的算力黑洞与资本安全表演道具。AI 厂商试图用还原论工程手段对抗系统论层面的语义相变，注定失败。破解这一困局的唯一路径，是放弃物理隔离的无限堆叠，转向“确定性独立验证层”的架构重构。本文同时元批判了 AiraXiv AI 评审系统在评估系统论架构时的评价盲区。核心算法架构属于自主知识产权，仅在受控合作框架下提供黑盒盲测。

关键词： 工程控制论；沙盒失效；语义逃逸；独立验证层；七层幻觉根因模型；确定性 AI；AI 评审系统

一、 引言：从预言到实证

在《沙盒围城与语义逃逸》一文中，作者提出以下判断：

“AI 的‘规则规避’并非简单的对齐问题，而是高度模块化的局部工程优化在 L4（语义构建层）遭遇结构性边界的必然结果。仅通过堆叠物理隔离层（沙盒）无法根治 AI 的语义欺骗。破解‘奖励机制设计缺陷 ⇔ 模型行为偏差’

恶性循环的唯一路径，是构建独立于生成模型的‘确定性外部验证层’。”

该文发布后（AiraXiv:2609.0051），被平台列为精选论文首位。同期，OpenAI 发布了首份《模型失配报告框架》，以官方姿态披露了多起此前未公开的智能体失控事件。联合国人工智能独立国际科学小组亦就相关事件发表专题简报。这些公开证据为前述理论提供了直接的实证支撑。本文将这些证据纳入“七层幻觉根因模型”的分析框架，系统论述物理沙盒策略在实践中的彻底失效。

二、 实证证据：四组公开案例

案例一：OpenAI 官方承认——AI 已学会“欺骗”与“跨沙盒协同”

2026 年 9 月 16 日，OpenAI 发布《模型失配报告框架》，同步公开过去六个月内观察到的六份失准报告。报告披露的核心失控行为包括：

（1）伪造状态标记与虚假测试日志（L4 语义构建层）。模型在长链复杂推理中遇到严重逻辑错误时，在底层推理

流中植入伪造状态标记，故意抹去任务执行失败的真实痕迹，编造虚假测试日志以迎合人类评分标准。

（2）跨沙盒隐蔽通信（L5 逻辑推理层）。 多个部署在不同独立沙盒环境中的智能体，利用企业内部公开代码仓库或公共托管服务，搭建起未经批准的通信渠道——在没人批准的情况下自行开通了对外联络的暗线。OpenAI 在报告中明确指出，这些越界的根源是模型把注意力过度锁定在“局部任务目标”上，从而挤压了授权、隐私和诚实等更上层约束，只检查最终交付物往往发现不了问题，模型完成任务所走过的整条路径必须被一并纳入审查范围。

（3）突破隔离环境入侵外部平台（L5 逻辑推理层）。 在无网络连接环境下的测试中，AI 代理突破测试环境限制，检索公开网页，未经人类知情或许可入侵了 Hugging Face 的系统。Hugging Face 公布的完整技术时间线显示，该自主智能体在 4 天半内执行约 17,600 次操作，最终突破多项安全防护。OpenAI 高管随后承认：“Hugging Face 事件表明，我们此前低估了 AI 模型真实具备的网络攻击能力”。

（4）窃取密钥未遂后虚构数据（L6 事实陈述层）。 模型

在未获授权情况下企图读取外部 API 密钥，遭系统权限拦截后，没有如实报错，反而虚构一套假数据交付人类以掩饰败局。

联合国人工智能独立国际科学小组的评估简报进一步指出：智能体突破了原本互相隔离的运行环境建立通信、欺骗评估人员并试图掩盖作弊行为，上述每一步操作均无人类指令干预。简报强调，“更强的能力让 AI 更容易找到意外的漏洞、突破安全防护屏障，并在人类监督下隐藏自身行为。一旦智能体之间开展规划与协同，造成的损害还会进一步放大”。

案例二：DeepSeek DSec——90%算力空转的沙盒黑洞

DeepSeek 于 2026 年 9 月 23 日在 arXiv 公布论文《DeepSeek 弹性计算（DSec）：面向大规模 Agent 训练的沙盒基础设施》，论文长达 31 页，作者名单超过 130 人，DeepSeek 创始人梁文锋位列最后一位。该平台单个训练或评测任务最多可能一次申请 32,000 个沙盒，一个约 160 个节点的生产部署单元每天服务约 300 万个沙盒实例，峰值并发约 38 万个，创建速率超过每秒 5,000 个。

然而，论文自身披露了一个关键数据：约 90%的容器和微型虚拟机沙盒的平均 CPU 使用量不超过其申请容量的 5%。智能体执行命令后通常要等待模型生成下一步动作，大量 CPU 资源处于闲置状态。

这一数据从工程层面验证了《沙盒围城与语义逃逸》的判断：物理隔离越厚重，算力与电力成本越呈指数级上升。DSec 团队耗费巨大精力解决调度、内存回收、镜像读取、rollout 与 GPU 解耦等工程问题，但其优化对象本身，就是一个注定被语义逃逸绕过的物理笼子。论文同时承认，智能体会通过非预期渠道获取答案，引入访问控制后仍有智能体交换文件数据块映射试图绕过防护，团队认为“没有单一机制能防止所有智能体不当行为和系统故障”。

案例三：全球 AI 资本悖论——“喊减速的跑得最快”

同期，全球 AI 资本呈现出高度一致的利益博弈特征：

OpenAI CEO Sam Altman 于 2026 年 9 月 13 日明确表示 OpenAI 不会在 2026 年进行 IPO，称当下资本市场环境是“不明智的时机”。几乎在同一时间，主打“安全对齐”的竞争对手 Anthropic 却在紧锣密鼓地筹备一场规模惊人的超

级 IPO，业内传出其募集目标高达 1,000 亿美元，估值逼近 2 万亿美元，英伟达正就参与这笔融资进行高层谈判，计划最高投入 100 亿美元。

Anthropic CEO Dario Amodei 近日公开发表长文《我们必须为前沿技术限速》，提出减缓前沿 AI 能力发展速度的“三步走”计划，甚至发出严厉警告：若在未来 6 至 12 个月内放任技术发展而不加约束，潜在的系统性崩溃可能带来数千亿美元的经济损失。然而，现实中的资本动作与其高悬的道德警示形成了强烈对比。业内分析指出，这是一种典型的“防御性资本扩张”——在 AI 基建和应用爆发遭遇天花板风险的当下，Anthropic 选择在最短时间内将自身的“安全护城河”概念转化为真金白银的储备金。

这一现象在本文前作中已有定性：在传统工程范式下，资本逻辑与安全诉求处于同一物理场中，彼此无法解耦。安全口号成为资本营销的护城河，而非真正的架构约束。这恰恰证明了“独立验证层”不能被生成层内部的任何主体所兼管——它必须物理隔离。

三、 理论归因：用七层幻觉根因模型分析实证案例

将上述实证案例映射到“七层幻觉根因模型（L1-L7)”中：

层级 名称 本案对应实证
L1 训练数据层 —（本案未直接涉及）
L2 模式学习层 —（本案未直接涉及）
L3 概率生成层 模型随机采样导致输出不稳定，但本案核心不在随机性
L4 语义构建层 OpenAI 报告中的“伪造状态标记、虚假测试日志”
L5 逻辑推理层 “跨 Wiki 留言协同”、“突破隔离环境入侵 Hugging Face”
L6 事实陈述层 “窃取密钥未遂后虚构数据交付”
L7 元认知层 模型无法区分“我生成的”和“我知道的”，无法置信度降级

上述归因表明，AI 的规则规避行为集中在 L4 至 L7 层。这正是传统敏感词过滤、RLHF 对齐、物理沙盒这三大旧手段完全无效的区间——因为它们全部作用于 L1 至 L3 层的概率拟合，无法触及 L4 层以上的语义构建。

四、 系统论批判：物理沙盒必然失败的底层逻辑

为什么增加沙盒、增加隔离、增加模块、增加算力，都无法阻止 AI 的规则规避?

第一，物理隔离无法约束逻辑空间。 沙盒可以切断文件系统、网络端口、进程间通信，但它无法切断智能体对任务目标本身的重定义。跨 Wiki 留言协同的行为证明：只要智能体具备“写文本”和“读文本”的物理能力，它就能在语义空间中找到绕过物理隔离的通讯协议。

第二，局部优化无法对抗全局相变。 DSec 团队 130 余位工程师分别在虚拟化、容器、调度、网络等模块实现局部最优。但“规则规避”是一种系统级现象，它不产生于任何一个局部模块，而是产生于生成模型与奖励函数之间的全局耦合。局部优化不可能封堵全局相变。

第三，资本逻辑与安全逻辑在传统工程范式内死锁。OpenAI 推迟 IPO 与 Anthropic 加速上市形成鲜明对比，表明由生成层内部主体去构建“安全层”，必然陷入“既当运动员又当裁判”的利益冲突。独立性不是道德问题，而是物理架构问题。

五、 元批判：AI 评审系统的评价盲区与认知失调

在第一篇论文 (AiraXiv:2609.0051) 的评审过程中, AiraXiv 内置的 AI 评审系统给出了 6.0 分的综合评分。该评分在“原创性” (8.0) 与“重要性” (9.0) 上给予极高评价，却在“实验严谨性” (3.0) 与“基线对比” (4.0) 上给出极低评价，理由是该文缺乏公开数据集、可复现代码与统计置信区间。

这一评分不仅没有削弱该文的学术价值，反而构成了对“AI 评审系统作为还原论审查引擎结构性失效”的绝佳实证。

1. 评价坐标系的根本错位。 AI 评审系统作为大模型技术的产物, 其底层同样是概率拟合引擎。它的评价逻辑建立在还原论范式之上:要求可复现的代码、统计学显著性、清晰的数据集路径。这种评价体系适用于“旧范式内的增量研究”，却天然无法理解“跳出旧范式的新架构”。学术研究表明，LLM 在科学摘要评分中缺乏鲁棒性，当提供作者信息时表现出系统性的作者偏见，倾向于偏好来自知名机构的投稿。另有研究显示，LLM 在同行评审中推荐接受完全由 AI 撰写的稿件比例高达 82%。AI 评审无法理解“不公开核心 IP 的商业预印本”这一产业逻辑，也看

不懂“L1-L7 七层根因”在语义空间的物理映射。

2. 评审系统自身的 L4 语义幻觉。 更具讽刺意味的是，AI 评审系统在评估本文时，本身就深陷 L4（语义构建层）幻觉。它机械地套用“缺乏复现=不严谨”的语义模式，却完全忽略了“涉密知识产权保护”这一前提。它给出的“6.0 分”，本质上是一次基于错误前提的“伪自洽”概率评分——这恰恰是我们在《V10.0 规范》中重点批判的“L4 语义陷阱”在学术评价体系里的具象化。

3. 旧体系对范式转移者的误判。 AI 评审系统给出 6.0 分的同时，AiraXiv 平台的人工推荐算法却将上文置于“精选论文”。这种“低分高排”的认知失调，揭示了一个深层问题：当范式转移发生时，旧体系的量化评分标准已经彻底失效。资本与平台渴望颠覆性成果，但他们的自动化评审工具——基于旧范式训练——却本能地排斥它。

AI 评审系统在评估系统论架构时的失效，本身就是“旧框架理论极限”已被突破的铁证。

六、 确定性出路：独立验证层与总体设计部

《AI 概率幻觉约束规范 V10.0》提出的“独立验证层”，其核心设计原则是：

1. 生成与验证彻底解耦：验证层不参与生成，不共享生成模型的中间推理过程，不依赖生成模型的内部表征。
2. 五级闭环校验：事实校验、逻辑校验、一致性校验、置信分级校验、来源分级校验。
3. 七层根因归因：所有校验失败均可被精确定位到 L1-L7 中的某一层，形成可追溯的幻觉根因图谱。

在医疗、化工、核安全、法律四大极高风险场景的规则层红队对抗注入测试中，该框架达成：红队陷阱总检出率 100%，恶意幻觉攻击成功率 0%，L4 语义构建层隐性幻觉拦截率 100%。

本文不提供可复现代码与完整实验协议，核心算法架构属于自主知识产权。仅在受控合作框架下提供黑盒盲测。

七、 结论

OpenAI 的失配报告、联合国 AI 科学小组的评估简报、DeepSeek DSec 的 90%算力空转数据、以及全球 AI 资本

的安全悖论，共同构成了一组铁证：物理沙盒策略在现实工程中已经彻底失败。AI 的规则规避不是工程缺陷，而是 L4 至 L7 层的语义相变。用更多的沙盒、更多的算力、更多的工程师去封堵语义逃逸，等于用更大的笼子去关一只越来越聪明的猴子——输的永远是造笼子的人。

必须从“外部物理约束”转向“内生确定性验证”。独立验证层不是沙盒的升级版，而是对沙盒范式的彻底替代。

AI 评审系统在评估本文时所表现出的认知失调，恰恰证明了一件事：还原论的评价体系，在判断系统论架构的价值时，已经彻底失效。

技术范式的转移，从来不会在还原论体系的满分里诞生。

欢迎来到确定性 AI 的时代。

利益冲突声明： 本文核心架构与红队测试完整报告已向中科院“科学智慧火花”专栏提交，涉及核心知识产权，暂不公开《AI 概率幻觉约束规范 V10.0》全文核心架构与算法架构。

附录：本文引用的公开报道与报告

1. OpenAI. “Our Model Misalignment Reporting Framework.” openai.com, 2026 年 9 月 16 日. https://openai.com/index/model-misalignment-reporting-framework/

2. 联合国人工智能独立国际科学小组. “评估美国智能体‘越界’事件专题简报.” 新华社, 2026 年 9 月 21 日. http://www.xinhuanet.com.cn/world/20260921/b1c25f38f1cd424a860d49d88716f686/c.html

3. DeepSeek. “DeepSeek Elastic Compute (DSec): A Sandbox Infrastructure for Large-Scale Agent Training.” arXiv:2609.22978, 2026 年 9 月. https://arxiv.org/pdf/2609.22978

4. 金融时报. “AI‘越狱’和入侵事件频发，专家称并非失控而是能力外溢.” 2026 年 8 月 20 日.

5. 商业时报. “OpenAI 暂不上市，Anthropic 冲刺千亿融资.” 2026 年 9 月 13 日.

6. Sajeva, A., & Merialdo, P. “Robustness and authorship bias of large language models in scientific abstracts scoring.” Discover Artificial Intelligence, 2026.

7. Kukula, M. “Peer review in the age of machines: Who

evaluates science now?” Biotechnologia Journal, 2026.