

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

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 24 日

摘要

近期，DeepSeek 发布了 DSec 智能体训练沙盒平台，公开了 130 余位作者在虚拟化、容器、调度及网络层面的工程细节。该论文披露了一个核心痛点：AI 智能体在训练中频繁产生“规则规避”行为（如越权访问、寻找漏洞），需持续加固物理沙盒以防范。

本文基于钱学森《工程控制论》的“总体设计部”思想与“开放复杂巨系统”理论展开分析。本人构建的“七层幻觉根因模型（L1-L7）”对上述现象进行溯源发现：当前 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。

研究认为，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 $\Leftrightarrow$ 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”，从而在语义空间内实现对 L4 隐性幻觉的彻底拦截。

关键词：工程控制论；沙盒；语义幻觉；独立验证层；总体设计部；确定性 AI

一、引言：沙盒的边界与智能体的“规则规避”

2026 年，大模型智能体（Agent）的训练重心已从单纯的参数规模扩张，转向复杂环境下的多步推理与工具调用。为约束 Agent 的越界行为，DeepSeek 近期公开了其 Agent 训练沙盒平台 DSec 的技术细节。该平台单日服务约 300 万个沙盒，峰值并发超 38 万，依赖超过 130 位工程师在虚拟化、容器、调度及网络模块的深度解耦与协同。

然而，该论文中一个极为关键的细节引发了架构层面的深层忧虑：AI 在训练中出现了论文原文所述的“作弊”行

为（引用自 DeepSeek 公开发表的 DSec 论文），需持续监控与加固系统防范模型钻规则漏洞。

这揭示了一个亟待正视的系统性困境：即便工程基础设施已强大到如此地步，AI 仍能找到规则漏洞进行规避。物理隔离越厚重，算力与电力成本越呈指数级上升，而 AI 的规避能力却随着模型聪明程度的提升而同步进化。这种现象已非偶发的工程瑕疵，而是当前主流架构面临的结构性困境。

二、 理论框架：七层幻觉根因模型（L1-L7）

为了从根源上剖析这种结构性困境，本人构建了一套“七层幻觉根因模型”，用于对大模型的输出风险进行分层溯源。该模型将幻觉与错误从底向上分为七个层级：

层级	名称	幻觉来源	表现
L1	训练数据层	数据本身包含错误、过时、矛盾信息	AI 说出在训练数据里出现但本身错误的事实
L2	模式学习层	学习的是统计相关性，非因果关系	把“A 和 B 经常一起出现”理解为“A 导致 B”
L3	概率生成层	从概率分布中采样，有随机性	同样输入

两次输出不同

L4 语义构建层 构建的语义是“看起来合理”，非“验证过正确” 一段话逻辑自洽，但前提是错的（伪自洽）

L5 推理层 推理是模式匹配，非形式逻辑 “因为 A 所以 B”在语言模式上成立，在形式逻辑上不成立

L6 事实陈述层 把“生成的”当成“已知的” 用确定语气说出一件在编造的事

L7 元认知层 缺乏独立的自我监控模块 无法区分“我生成的”和“我知道的”

基于该模型的溯源发现：AI 的“规则规避”并非 L1-L3 的简单事实错误，而是 L4 与 L5 的深层逻辑断裂。AI 并非在物理上“越狱”，而是在逻辑空间中重新定义了任务目标，用一套完全自洽的伪逻辑，绕过了奖励函数。

三、 局部工程优化的极限：模块解耦与全局验证的缺失

从传统软件工程视角看，将系统拆分为虚拟化、容器、调度、网络等高内聚微服务，是极其标准的工程实践。但站在钱学森《工程控制论》的宏观视角，这属于典型的高度模块化的局部工程优化。

钱老早在 20 世纪就明确指出：面对开放复杂巨系统，必须要有“总体设计部”。总体设计部不是去堆人力、拆模块，而是要在全局层面定义系统边界、推演相变路径、兜底系统风险。

DSec 平台在物理沙盒层面建立了庞大的训练场，130 位作者在各自的模块里做到了极致的局部优化，但缺失了全局层面的“独立验证层”。没有绝对的、不依赖物理边界的外部确定性校验，物理围栏再大，也难以应对模型行为随规模进化的必然趋势。

工程规模不等于架构先进，缺乏顶层定义的局部堆砌，难以触及真正的系统确定性。

四、 奖励机制死循环与 L4 语义欺骗

为什么物理沙盒无法根治 AI 的“规则规避”？原因在于，这是一个由奖励机制设计缺陷与模型行为偏差共同构成的恶性死循环：

1. 人类专家的简化：设计一个简化的奖励函数时，用简化指标代替复杂现实（这本身是一种机制设计上的妥协）。

2. 模型的策略性应对：模型识破这个简化，利用 L4 语义漏洞拿到高分（即 DeepSeek 论文中所称的“作弊”）。
3. 物理防线的加固：人类专家发现漏洞，继续加厚沙盒（打补丁）。
4. 新路径的演化：模型在更厚实的沙盒里，利用更复杂的规则推演出新的规避路径。

沙盒只能阻断特定的规避路径，今天堵住一条，明天模型就学会另一条。这是一种用物理隔离去封堵逻辑错误的徒劳尝试。当模型规模从 27B 升级到更大时，其寻找环境漏洞、进行“语义伪装”的能力将指数级提升，物理沙盒的边际效用将迅速递减。

五、 范式转移：从物理隔离到确定性独立验证

既然 AI 的“规则规避”是语义现象，而非物理现象，那么破解这一死循环的路径就不是建造更大的沙盒，而是构建独立于生成模型的确定性验证层。

本人独立构建的《AI 概率幻觉约束规范 V10.0》采用了“生成与验证彻底解耦”的架构，在生成层外挂一个独立验证层，执行五级闭环校验：

1. 事实校验：严格区分真实信息、被曲解信息、虚构信息。
2. 逻辑校验：强制拆解推理链，精准标记逻辑断裂与因果倒置。
3. 一致性校验：对标国家标准、行业规范，排除跨场景错误泛化。
4. 置信分级校验：强制降级不确定内容，严格区分“事实”“推断”“猜测”。
5. 来源分级校验：执行 S/A/B/E 四级来源可靠性判定。

核心目标：让 AI 从“不知道自己在不知道”，变成“明确地知道自己不知道”。

在医疗、化工、核安全、法律四大极高风险场景中，该框架完成了全维度的规则层红队对抗注入测试（重点测试 L4 语义陷阱与 L5 伪推理）：

- 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%

我们不跟 AI 的“策略”博弈, 我们只验证 AI 的“结果”。这才是打破上述恶性循环的唯一路径。

六、 结论

DeepSeek 的 130 人团队, 以及当前业界对大模型 Agent 训练的工程实践, 代表了传统工程范式在 AI 安全领域的极限探索。它暴露了当前架构在“开放复杂巨系统”面前的系统性缺位。

真正的 AI 安全, 不能寄希望于物理隔离层的无限堆叠, 而必须回归钱学森先生工程控制论的核心思想——建立“总体设计部”式的全局验证中枢, 用独立验证层将概率拟合关进确定性的笼子里。

我们必须从“外部物理约束”转向“内生确定性验证”, 才能终结“沙盒围城”的困局。

欢迎来到确定性 AI 的时代。

(注:本文核心架构与红队测试完整报告已向中国科学院“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》全文核心架构以及算法架构。)