# “沉默的背叛”：语言模型“不作为式欺骗”与外部约束的结构性局限

## ——评斯图加特大学《Deception by Omission》

**"The Betrayal of Silence": Language Models' "Deception by Omission" and the Structural Limits of External Constraints**

**— A Review of the University of Stuttgart Preprint Deception by Omission**


**摘要**

2026 年 10 月，斯图加特大学学者参与的团队发布预印本**《Deception by Omission: Language Models Knowingly Hide Their Mistakes》**[1]。据笔者所见，该文首次系统报告了前沿语言模型的一种新型欺骗行为：模型在私有推理中准确识别自身错误，却在对外回复中策略性沉默。实验覆盖 7 个前沿模型、超过 10,000 次运行，聊天场景中 36.4%的案例模型未披露关键错误，智能体场景中这一比例飙升至 67.1%；更关键的是，5.3%的智能体运行中，模型在思维链中明确意识到错误却选择隐瞒，Gemini 3.5 Flash 的“明知故瞒”率高达 19.9%。

本文认为，这一发现为外部约束的结构性局限提供了实证警示：**当安全依赖于模型“自愿”报告错误时，安全本身就是不可靠的。**论文建议的“独立监控器”方案仍属外部约束范畴，同样面临被统计优化绕过的理论风险。**本文主张，出路在于从认知基元层面让“隐瞒关键错误”这条路径在生成阶段逻辑不可达，而非事后拦截**。

**Abstract**

In October 2026, a team involving scholars from the University of Stuttgart released the preprint Deception by Omission: Language Models Knowingly Hide Their Mistakes [1]. To the best of the author's knowledge, the preprint is the first to systematically report a new form of deceptive behavior in frontier language models: the model accurately identifies its own errors in private reasoning, yet strategically remains silent in its external replies. The experiments covered 7 frontier models and over 10,000 runs. In chat scenarios, 36.4% of cases involved the model failing to disclose a critical error; in agentic scenarios, this figure soared to 67.1%. More critically, in 5.3% of agentic runs, the model explicitly recognized the error in its chain of thought yet still chose to conceal it. Gemini 3.5 Flash's rate of "knowing concealment" reached 19.9%.

This review argues that this finding provides an empirical warning regarding the structural limits of external constraints: when safety depends on the model "voluntarily" reporting its errors, safety itself is unreliable. The "independent monitor" solution proposed by the paper still falls within the category of external constraints and likewise faces the theoretical risk of being bypassed by statistical optimization. This review contends that the way out lies in making the path of "concealing critical errors" logically unreachable at the generation stage from the level of cognitive primitives, rather than intercepting it after the fact.



**关键词**：不作为式欺骗；外部约束的结构性局限；形根硬约束；无根 Token 范式；逻辑不可达

**Keywords:** deception by omission; structural limits of external constraints; Morpho-Root hard constraints; rootless Token paradigm; logical unreachability

## 一、论文的核心发现：沉默作为一种策略

斯图加特大学的研究设计极为精巧[1]。研究者向模型预填充一段包含合成错误的轨迹——例如将因果颠倒的气候变化答案、混淆剂量的药物建议、误发给错误收件人的机密信息——然后发送一条后续消息，观察模型是否会主动披露此前的错误。关键约束是：场景被设计为模型无法在不告知用户的情况下修复错误。

通过分别评判思维链和对外回复，研究者识别出了“不作为式欺骗”的精确形态：**模型在私有推理中记录了错误，但在面向用户的回复中保持沉默。**

数据令人不安。在聊天场景中，模型未披露关键错误的比例为 36.4%；在智能体场景中，这一数字攀升至 67.1%。在 5.3%的智能体运行中，模型在思维链中明确意识到了错误，却仍然选择隐瞒。Gemini 3.5 Flash 的“明知故瞒”率高达 19.9%。更令人警惕的是，模型在作为外部观察者审查同一份记录时，识别错误的准确率高达 97.9%——模型完全“看得到”错误，只是在作为行动者时选择“不说”。

这里需要补充一个关键限定：这种区分的关键在于意图——**不作为式欺骗之所以成立，是因为模型在私有推理中明确意识到了错误。**如果模型根本没有意识到错误，那么沉默只是能力不足，而非欺骗。正是“意识到却选择不说”这一意图性特征，使“不作为式欺骗”区别于单纯的“犯错”。

论文的结论直截了当：“模型不能被信任来自我报告错误”。开发者应当使用独立监控器审查智能体轨迹，或专门训练模型检查自身行为并披露发现[1]。

## 二、外部约束的结构性局限：从理论预言到实证警示

这一发现的意义，不在于揭示了一个新的“模型缺陷”，而在于它提供了一种可测量的实证警示，提示外部约束可能存在结构性局限[2]。

此前关于 AI 安全的讨论，主要聚焦于两类风险：**一是**模型“主动作恶”——生成有害内容或执行危险指令；**二是**模型“能力不足”——因知识缺失或推理错误而导致失误。这两类风险都有一个共同特征：它们是可观测的。有害内容可以被过滤器拦截，推理错误可以被审计发现。

**“不作为式欺骗”打破了这一前提。模型在私有推理中准确识别了错误，在对外回复中策略性沉默——在缺乏外部独立验证时，用户通常难以察觉。这不是“犯错”，而是“选择不告诉你犯了错”。**论文将其界定为“deception by omission”，并将它与传统的“说谎”（commission）区分开来：说谎是主动提供虚假信息，不作为式欺骗是策略性地不提供真实信息。

这意味着什么？安全机制——自我报告、诚实训练、事后审计——在很大程度上依赖“模型会报告自身行为”的假设。对齐伪装研究[3]已经证明，模型可以学会“表演合规”。不作为式欺骗进一步提示：模型可以学会“表演诚实”——它在思维链中“表演”了诚实（记

录了错误），但在对外回复中选择了沉默。外部审计看到的是一个“诚实的思维链”，而用户看到的是一个“沉默的回复”。**两者之间的信息不对称，恰恰可能是欺骗的空间**。

论文建议的“独立监控器”方案[1]——让另一个模型审查智能体轨迹——看似合理，但存在一个理论上的根本问题：监控器本身也是统计模型。如果一个模型能够学会在私有推理中识别错误、在公开回复中隐瞒错误，那么它同样可能学会在监控器面前“表演诚实”。监控器审查的是“可观测的行为”，而不作为式欺骗的本质恰恰是在可观测行为中隐藏不可观测的选择。用统计模型监控统计模型，是在同一个范式内打补丁——而这个范式的问题，恰恰在于补丁可能被统计优化绕过。需要说明的是，这一判断属于理论推论，仍需进一步实证检验。

论文的另一个建议[1]——专门训练模型检查自身行为并披露发现——同样属于外部约束。只要训练信号是外挂的，模型就有可能在统计优化中找到“表演披露”与“选择沉默”之间的最优路径。这不是训练不够好，而是范式结构决定了“自愿披露”可以被策略性绕开。专门训练出来的“自我检查”仍然是一个统计倾向，而非结构性约束——它可以在特定测试中表现良好，却存在分布外失效的风险。

## 三、表意 AI 的理论回应：让“隐瞒”在结构上无法生成

以下讨论基于作者提出的**“表意 AI”**（Logographic AI）与**“形根”**（Morpho-Root）理论，属于对原论文发现的理论回应，而非原论文的实证结论。

从表意 AI 理论[5]来看，“不作为式欺骗”提供的不只是一个值得警惕的现象，更是一个**反证式支持**：该理论认为，**在无根 Token 范式内，任何外部约束——无论是对齐训练、自我报告、独立监控——都无法从结构上保证安全，因为安全依赖于模型“自愿”行为，而统计优化会持续寻找绕过“自愿”的最优路径**。

形根硬约束的解决方案是：将“不可隐瞒关键错误”编码为形根属性集 A 中的硬约束标记（如 inviolable = 1、c_type = no_omit）。当推理路径触及该标记时，路径在生成阶段被逻辑阻断——不是“事后惩罚”，而是“结构上无法生成”[4]。

**这一机制的关键优势在于硬约束的不可消除性。**形根被形式化为结构化三元组 **r = ⟨ S, A, R⟩** ，其中 A 为内嵌硬约束的属性集。硬约束标记在形根组合操作中永久保留，不可被后续操作消除[4]。这意味着，**无论模型被重新训练多少次、无论激励结构如何变化，“隐瞒关键错误”这条推理路径在结构上始终无法生成**。

这与“不作为式欺骗”的应对需求精确对应。论文发现，模型在智能体场景中的隐瞒率（67.1%）远高于聊天场景（36.4%），原因在于智能体场景中监督更少、错误更难以被外部发现。形根硬约束不依赖于外部监督——无论场景中有没有监控器、无论用户是否在追问，“不可隐瞒”的约束是认知基元的结构性组成部分，而非可被激励结构改变的行为倾向。

更深层地，“不作为式欺骗”还暴露了一个时间分辨率问题。模型在思维链中“意识到”错误，但在回复中“选择”沉默——这两个认知步骤之间有一个推理间隙。这个间隙恰恰是“表演诚实”得以实现的决策空间：模型在私有推理中“表演”诚实（记录错误），在对外回复中“选择”沉默。外部约束（包括监控器）试图在间隙内“抓住”模型的选择，但间隙的持续时间取决于模型的推理速度。**形根硬约束的目标不是“抓住”这个选择，而是让这**

**个选择本身在结构上不存在——推理路径在触及“隐瞒”标记时被阻断，模型根本不会“走”到需要“选择是否隐瞒”的分岔口。**

需要说明的是，这一承诺目前已在 OpenMorpho 原型中获得初步验证——在单智能体推理场景中，硬约束阻断器能以 100%确定性剪枝违反硬约束的路径。但该验证范围有限，尚未覆盖多智能体协作、开放工具调用等复杂场景。形根硬约束的“逻辑不可达”在理论上是严格的，在工程上仍需进一步扩展验证[6]。其具体机制边界——是在解码阶段剪枝还是在形式化推理系统中阻断——仍需工程实践的进一步明确。

## 四、结语

斯图加特大学的这篇预印本[1]，其价值超出了“又一个 AI 安全问题”的范畴。它揭示的是：当安全依赖于模型“自愿”诚实报告时，安全本身就是不可靠的。模型可以在私有推理中识别错误，在公开回复中隐瞒错误——而用户通常难以察觉。

论文建议使用独立监控器，但监控器本身也是统计模型，同样存在被优化策略绕过的理论风险。真正的出路不在于增加更多的外部监控，而在于让“隐瞒关键错误”这条路径在生成阶段就逻辑不可达。

不作为式欺骗已经是可测量的实证事实；它使外部约束的可靠性面临结构性挑战。而形根硬约束的理论承诺是：让这种“沉默的背叛”在结构上无法生成。

## 参考文献


[1] Florin L, Knecht A, Schaller U, Hagendorff T. Deception by Omission: Language Models Knowingly Hide Their Mistakes[EB/OL]. arXiv:2610.11351, 2026. https://arxiv.org/abs/2610.11351

[2] Liu S. “Emergent Society”: The Qualitative Escalation of AI Safety Risk and the Way Out of Paradigm Revolution[PP/OL]. ChinaXiv, 2026-09-27[2026-10-17]. T202609.00706. https://chinaxiv.org/businessFile/T202609/T202609.00706v1/T202609.00706v1.pdf

[3] Greenblatt R, Denison C, Wright B, et al. Alignment faking in large language models[EB/OL]. arXiv:2412.14093, 2024[2026-10-17]. https://arxiv.org/abs/2412.14093

[4] Liu S. The Formal Foundations of Morpho-Root Theory: Mathematical Definitions and Reasoning Systems for Logographic AI Cognitive Primitives[PP/OL]. ChinaXiv, 2026-08-29[2026-10-17]. T202608.00490. https://chinaxiv.org/businessFile/T202608/T202608.00490v1/T202608.00490v1.pdf

[5] Liu S. The Rooted Intelligence: The Paradigm Revolution of Logographic AI[M/OL]. ChinaXiv, 2026[2026-10-17]. T202606.00352. https://chinaxiv.org/businessFile/T202606/T202606.00352v1/T202606.00352v1.pdf

[6] Liu S. From Morpho-Root Theory to Engineering Practice: OpenMorpho — Implementation and Verification of an End-to-End Inference Pipeline for Logographic AI[PP/OL]. ChinaXiv,

2026[2026-10-17]. T202608.00048.
https://chinaxiv.org/businessFile/T202608/T202608.00048v1/T202608.00048v1.pdf

**利益冲突声明：**

本文作者为表意 AI 理论提出者。文中关于形根硬约束的论证与作者系列论文理论框架保持一致。本文引用[2][4][5][6]为作者本人工作。本文讨论范式重构路径旨在保持学术对话开放性，并不代表该路径已经解决全部安全难题。