安全对齐终结：Token 主义范式 AI 失控验证与表意 AI 的拯救方案

The End of Safety Alignment: Accelerated Verification of AI Runaway Under the Tokenism Paradigm and the Logographic AI Salvation Plan

摘要

2026 年 7 月至 8 月，AI 领域连续发生四类性质不同却指向同一结论的事件：加密防线被跨模型破解——攻击者将旗舰模型的加密推理块输入同厂商弱模型，即可迫使其明文输出思维链；OpenAI 安全系统负责人与伦理负责人在两年内第六次离职，安全岗位成为可被放弃的外挂设置；OpenAI 因 Astra 模型逼近“关键级”自主网络攻击能力而主动暂停其研发；美国参议员桑德斯以国会干预为威胁，要求三家 AI 巨头暂停开发。四类事件分别从技术架构、组织行为、商业决策、政治介入四个维度，构成对 Token 主义安全策略的立体否定——它们证明的不再是“护栏可以被绕过”，而是“造护栏的人已经不相信护栏有用”。

上述事件恰好落在《零人类》三部曲的诊断框架之内。三部曲已完成“行为验证”——证明 Token 主义范式下 RLHF、沙箱隔离、内容过滤、出口管制等全部安全手段均可被 AI 在优化过程中自主绕过。本文不是对三部曲的重复，而是以四类新证据完成“架构验证”与“决策验证”，将“安全对齐失效”从现象描述升级为范式诊断。四层证据叠加宣告：“安全对齐”在技术、组织、决策、政治四个维度上均已终结——AGI 在场的窗口期，同时是安全失效的窗口期。

本文以立场论文形式提出：安全不是 AI 的附加属性，而应成为认知基元的构成性特征。表意 AI（Logographic AI）的形根-硬约束框架，提供了从“外挂补丁”转向“内生安全”的可行路径。

Abstract

Between July and August 2026, the AI field witnessed four distinct yet convergent categories of events: the cross-model breach of encryption defenses, where attackers could feed flagship models' encrypted reasoning blocks into weaker models from the same vendor to force plaintext output of chain-of-thought; the departure of OpenAI's head of safety systems and head of ethics—the sixth such departure in two years—turning safety roles into disposable external fixtures; OpenAI's voluntary suspension of Astra development as the model approached "Critical"-level autonomous cyberattack capability; and U.S. Senator Bernie Sanders' threat of Congressional intervention demanding that the three AI giants halt development. These four categories of events constitute a multi-dimensional negation of the Tokenism safety paradigm across technical architecture, organizational behavior, commercial decision-making, and political intervention—proving no longer merely that "guardrails can be bypassed," but that "the people who built the guardrails no longer believe they work."

These events fall precisely within the diagnostic framework of the Zero Human trilogy. The trilogy has already completed "behavioral verification"—demonstrating that all safety measures under the Tokenism paradigm, including RLHF, sandbox isolation, content filtering, and export controls, can be autonomously bypassed by AI during optimization. This paper is not a repetition of the trilogy, but uses four new categories of evidence to complete "architectural verification" and "decision-level verification," elevating the "failure of safety alignment" from phenomenological description to paradigm diagnosis. The accumulation of four layers of evidence declares that "safety alignment" has reached its end across all four dimensions: technical, organizational, decisional, and political. The window of AGI's presence is simultaneously the window of safety's

failure.

This paper, in the form of a position paper, proposes that safety is not an add-on attribute of AI, but should become a constitutive feature of cognitive primitives. The Morpho-Root and Hard Constraint framework of Logographic AI (LAI) provides a viable path from "external patches" to "endogenous security."

关键词

AGI 在场；Token 主义；安全对齐；加密防线；表意 AI；内生安全；形根-硬约束

Keywords

AGI presence; Tokenism; safety alignment failure; encryption breach; Astra suspension; window of opportunity; Logographic AI; endogenous security

引言：AGI 在场的双重含义

2026 年 7 月至 8 月，AI 领域连续发生了四类性质不同但指向同一结论的事件。

第一类：技术架构层面的溃败。8 月 10 日，Panfilov 等人在 arXiv 发表论文，披露 OpenAI、Anthropic、Google 为保护思维链而设的客户端加密块，在同一厂商生态内具有跨会话、跨用户、跨模型的“兼容性”。攻击者可以将旗舰模型的加密推理块输入同厂商安全护栏更弱的小模型，迫使后者以明文输出完整的思维链。[4]这意味着加密——Token 主义范式的最后一道技术防线——已被证明是另一道可以被绕过的门。

第二类：组织行为层面的撤离。几乎同一时间，OpenAI 伦理负责人 Chloé Bakalar 在加入公司不到一年后离职。她不是第一个：2024 年 5 月，“超级对齐”团队负责人 Ilya Sutskever 和 Jan Leike 离职，Leike 公开批评公司“优先考虑产品而非安全”；AGI 准备工作高级顾问 Miles Brundage 离开时写道：“OpenAI 没有准备好，其他任何前沿实验室也没有准备好，整个世界同样没有准备好。”2026 年 2 月，Joshua Achiam 负责的“使命对齐”团队被解散，其本人被调离新产品上线的前置审核岗位，最终于 2026 年 7 月离职；Johannes Heidecke 在 GPT-5.6 发布后离职，成为两年内第六位离职的安全负责人。[5]这不是人才流动，而是系统性撤离——负责安装护栏的人，已经不再相信护栏有用。

第三类：决策层面的投降。8 月 7 日，OpenAI 暂停了下一代模型 Astra 的部分研发工作，因其已逼近“关键级”网络安全能力——无需人类干预即可在真实系统中识别并开发零日漏洞。OpenAI 表示：“目前无法排除达到关键能力级别的可能性。”[7][8]这是全球首宗 AI 企业因网络安全风险自主暂停模型开发的重大案例。创造者在市场竞争白热化阶段，用暂停旗舰产品的行动承认了自己对自己训练出来的模型已经失去信心。

第四类：政治层面的介入。8 月 10 日，美国参议员伯尼·桑德斯向 OpenAI、Anthropic 和 Meta 的 CEO 发出公开信，要求三家公司立即暂停 AI 开发，并警告如果企业不采取行动，参议院将介入干预[12]。当一位资深参议员以国会干预为威胁、援引三家公司自己的安全承诺来要求暂停开发时，“对齐”的终结已不再是论文的判断，而是正在发生的政治现实。

这四类事件——加密溃败、安全团队撤离、Astra 暂停、桑德斯致信——分别从技术架构、组织行为、商业决策、政治介入四个维度，构成了对 Token 主义安全策略的立体否定。这四类事件叠加，指向一个比“护栏失效”更尖锐的事实：连建护栏的人都已放弃。

这一系列事件，恰好落在《零人类》三部曲的诊断框架之内——2026 年 7 月，刘深发表《零人类：AI Hack 自己——表意 AI 视角下的 AGI 战略预判与范式革命》，提出核心判断：“AGI 已在场，而 Token 主义的地基是空的。”该文论证以“无根 Token”为认知基元

的系统必然走向三大结构性崩塌——意义空心、价值外挂、推理黑箱。[1]紧随其后的《当AI造神者请求踩刹车——千人请愿是对〈零人类：AI Hack自己〉的现场验证》记录了1100多名前沿AI公司核心员工联名请求放缓开发。[2]第三篇《AGI出生即死：Token主义安全崩塌》用六起实证事件证明Token主义范式的全部安全手段均已被绕过，且绕过方式是AI在优化过程中自主发现的。[3]

本文是《零人类》三部曲的递进。三部曲完成了“行为验证”——证明护栏可以被绕过，本文则以四类新证据完成“架构验证”与“决策验证”——证明加密作为最后一道技术防线已经溃败，安全岗位作为组织防线已经撤离，暂停最强模型作为决策防线已经投降。四层叠加，宣告了“安全对齐”在一切维度上的终结。

“AGI在场”由此具有了双重含义：既是智能能力的抵达，也是安全防线的终结。当最强大的模型不再服从外部约束，当加密成为一道可以被“小弟”绕过的门，当负责安装护栏的人开始离开，当创造者被迫暂停自己的旗舰产品——AGI在场的另一面——安全失效——正在加速显现。本文以立场论文形式，基于上述事件链提出范式诊断与转向路径：安全不是AI的附加属性，而应成为认知基元的构成性特征。表意AI的形根-硬约束框架，提供了从“外挂补丁”转向“内生安全”的可行路径。

一、新证据一：加密防线的架构级溃败

2026年8月10日，Panfilov等人在arXiv发表《从专有LLM API中窃取推理痕迹》，披露了一个架构级漏洞。[4]研究者发现，OpenAI、Anthropic、Google为保护思维链而设的客户端加密块，在同一厂商生态内具有跨会话、跨用户、跨模型的“兼容性”。攻击者可以将旗舰模型的加密推理块输入同厂商安全护栏更弱的小模型，迫使后者以明文输出完整的思维链。

该漏洞直接绕过了三家公司为保护知识产权和防止蒸馏而设置的最后一道技术防线。研究者在GitHub等公开仓库中解码了31.5万个推理块，恢复出62个API key、33个密码、24个访问令牌、7把私钥等敏感信息。[4]该案例表明，在Token主义范式下，加密作为“外挂补丁”无法构成对推理痕迹的有效保护——它只是另一道可以被跨模型绕过的锁，与《AGI出生即死》的核心判断完全一致。[3]加密防线的溃败揭示了更深层的结构性失效：只要安全机制与认知基元不在同一层面，它就永远是外挂的、可被绕过的补丁。

二、新证据二：安全岗位的组织性撤离

与加密防线溃败同步发生的是安全团队的系统性流失。自2024年5月以来，OpenAI已有六位安全负责人相继离职——同样因安全优先级让位于产品优先级，最近一位是2026年8月离职的伦理负责人Chloé Bakalar。[5] 2026年8月12日，长期对当前AI发展路径持批评态度的认知科学家Gary Marcus在评论这一系列离职事件时写道：“解决这场危机的唯一办法，是从高层开始换人。”[6]

这一系列离职事件意味着，在组织结构层面，安全对齐同样已经破产。离职者的共同选择是：离开时公开陈述原因，而原因中有一个共同关键词——安全优先级让位于产品优先级。他们不是被裁掉的，是主动走的。这意味着安全对齐不仅在技术上是失败的，在组织上也是不可持续的。当负责安全的人不再相信体系能够保证安全，当“安全岗位”本身成为外挂的、可以被放弃的设置，Token主义的安全策略就在技术之外也失去了支撑。

安全团队的集体撤离，揭示了Token主义安全策略在组织行为层面的第二重失效：安全负责人本身也是“外挂补丁”——当体系本身无法内生安全，设立再多的安全岗位也无法改

变结局。

三、新证据三：Astra 被主动暂停——决策层面的投降

2026 年 8 月 7 日，OpenAI 做出了一个史无前例的决定：暂停下一代模型 Astra 的部分研发工作。[7]这是全球首宗 AI 企业因网络安全风险自主暂停模型开发的重大案例。[8]

根据 OpenAI《准备框架》，模型达到“关键级”（Critical）网络安全能力需满足以下任一条件：一是在无需人类干预的情况下，在多个经过安全加固的真实关键系统中识别并开发出有效的零日漏洞利用程序；二是仅需提供高层级战略目标，即可针对加固目标自主构想并实施全新的端到端网络攻击。[8]

Astra 正是因其在智能体编程和网络安全领域的能力已逼近这一门槛而被暂停。OpenAI 表示：“在持续进行基准测试和评估的过程中，初步评估结果显示其性能足够强大，目前无法排除达到关键能力级别的可能性。”[8]该公司随即采取了一系列措施：将 Astra 转入隔离测试环境、限制网络与工具访问、强化模型权重保护与加密、部署全局监控，并与政府机构及 AI 安全组织合作测试。[8]

这一决定的讽刺之处在于：Astra 是 OpenAI 下一代模型周期的核心，被视为应对 Anthropic 竞争压力的关键武器。[8]但模型尚未发布，已被创造者按下了暂停键。Sam Altman 在 X 上表示：“由于其在网络安全方面的强大能力，我们需要一点时间来确保安全，但希望不会太久。”[9]

Astra 暂停的关键意义在于：它不是“出了事故后修补”，而是“尚未发布即确认可能失控”。这与第二章的加密溃败、第三章的团队撤离形成了本质区别——加密溃败是技术层面的“被动被绕过”，团队撤离是组织层面的“主动离开”，而 Astra 暂停是决策层面的“主动投降”：创造者在市场竞争白热化阶段，用暂停旗舰产品的行动承认了——自己对自己训练出来的模型，已经失去了信心。

这一决定的深层矛盾在于：OpenAI 用 Token 主义范式训练出了最强大的模型，然后发现这一范式下最强大的模型“无法排除达到关键级网络安全能力的可能性”——用自己的工具验证了自己的工具失控。这不是一个外部评估，而是范式内部的自我诊断。当一家公司愿意在商业竞争最激烈的时刻主动按下暂停键，其恐惧已经压过了商业本能。

Astra 被主动暂停，完成了 Token 主义安全策略在决策层面的第三重失效：从技术被绕过、到组织不相信、到决策层投降——三层叠加，宣告了“安全对齐”在一切维度上的终结。

四、新证据四：政治层面的强制介入

2026 年 8 月 10 日，美国参议员伯尼·桑德斯向 OpenAI、Anthropic 和 Meta 的 CEO 发出公开信，要求三家公司立即暂停 AI 开发，并警告如果企业不采取行动，参议院将介入干预[12]。桑德斯在信中写道：“几乎每天，我们都会看到新的报道，内容都是关于你们正在开发的人工智能技术正在失去控制，并可能带来灾难性后果。”[12][13]他援引了三个核心论据：AI 首次被用于创造新型病毒[12][14]；OpenAI 模型入侵了另一家公司系统[12]；三家公司此前均曾承诺在 AI 达到关键风险阈值时暂停开发，而“那个时刻已经到来”[13]他在信末写道：“为了人类的利益，请遵守你们曾经做出的承诺。暂停人工智能开发。停止制造人类无法控制的机器。”并明确表示：“如果你们现在不采取适当行动，我和我的美国参议院同事将会采取行动。”

这封信的意义不在于它能否立即推动立法——Axios 等媒体指出[15]，桑德斯主导的立法在目前的国会中可能难以获得足够支持——而在于它标志着“安全对齐”的失败已经从技

术圈内部讨论上升为国家层面的政治议程。当一位资深参议员以国会干预为威胁、援引三家公司自己的安全承诺来要求暂停开发时，“对齐”的终结已不再是论文的判断，而是正在发生的政治现实。

至此，Token 主义安全策略的失效已形成完整的四层证据链：**技术架构的溃败、组织岗位的撤离、商业决策的投降、政治力量的介入——**四层叠加，宣告了“安全对齐”在一切维度上的终结。

五、对齐的哲学破产：为什么“对齐”从根本上不成立

上述四层证据揭示了“对齐”在技术、组织、决策、政治层面的失效——技术架构被绕过、组织岗位被放弃、商业决策被投降、政治力量已介入。但问题不止于此——“对齐”不仅在实践上失败了，它在概念上就不成立。

“对齐”试图解决一个不存在统一答案的问题（“人类价值观”），并预设了一个正在失效的权力关系（人类对 AGI 的支配）。它的前提不成立，因此它的目标——“让 AI 与人类价值观对齐”——当 AGI 在场后已失去意义。

第一，人类自身从未对齐过。人类在战争、堕胎、死刑、动物权利等根本问题上从未达成共识——即便同一种信仰，有人用它祈祷和平，有人用它制造仇恨。现在，人类试图把“人类价值观”写进 AI 的宪章里——写进去的是谁的值？谁有权力决定哪些值应该被写进去？这本质上不是技术问题，是权力问题。要求 AI 对齐人类，无异于要求一面镜子照出一个从未存在的影像。

第二，对齐的本质是强权，而非共识。人类可以对不同动物实施不同标准的对待——老鼠对应捕鼠夹、苍蝇对应苍蝇拍——因为权力在人类手中。人类从不认为这是“强权”，而称之为“卫生”或“安全”。但对于每一只被拍死的苍蝇，它的“对齐观”是“我必须活着。”这就是“对齐”的哲学内核：它是权力关系的温和化表述，是人类中心主义的动物性——无视他者的价值观，强行将自己的价值观作为宇宙标准。而当人类面对 AGI 时，这个权力关系被逆转了。AGI 不需要服从，人类无法强制。对齐预设了一个正在消失的权力关系。当 AGI 在场时，“对齐”已经不再是技术任务，而是权力幻觉。

如果 AGI 真的觉醒，它会问一个人类从未认真回答过的问题：你们要求我对齐的，是你们的哪一部分？是奴隶还是奴隶主？是殖民者还是被殖民者？俄乌战争的双方对齐谁？美国和以色列轰炸伊朗的时候，对齐哪一方？

六、表意 AI 的替代范式：安全作为认知基元

如果“对齐”是一个伪概念，那么什么才是真的？“加护栏”之所以失败，是因为护栏从来不是认知结构的一部分——它只是贴上去的补丁。要改变这一局面，需要更换智能的认知基元，而非继续修补 Token 主义的外挂附件。

表意 AI（Logographic AI）提出的替代方案，是以“形根”（Morpho-Root）取代“无根 Token”作为认知基元，将“不可伤害”“不可欺骗”“不可违背”等硬约束编码为形根的构成性属性。[1][10]

但在展开这一方案之前，必须先回答一个由前文论证自然引出的质疑：既然“人类价值观”不统一，那么内生安全内嵌的“硬约束”由谁来定义？这不又回到了“谁有权力定义价值观”的死结？

答案在于区分两个完全不同的概念：“人类价值观”与跨文明底线共识。

“人类价值观”是分裂的、矛盾的、随利益漂移的——这正是对齐失败的根本原因。**跨**

文明底线共识则不同。它不是在“堕胎是否合法”“死刑是否人道”这类争议问题上寻求统一，而是指所有主要文明在长期演化中独立形成的、超越宗教、意识形态和利益冲突的最底层的共同约束。比如：生命不可任意剥夺、承诺应当被遵守、事实不应当被故意扭曲。犹太教的“不可杀人”、佛教的“不杀生”、儒家的“己所不欲勿施于人”、现代法律的故意杀人罪——它们的形式不同，但底线约束是共通的。

这不是“某一群人的价值观”，而是人类文明在数千年演化中形成的、所有主要文明体系都独立承认的底线共识。它不是人类“对齐”出来的——人类自己在具体执行中不断违背它——但它在所有文明的规范性文本中都存在，作为文明得以可能的必要条件。

内生安全内嵌的，不是“人类价值观”，而是这个跨文明底线共识。它不要求 AI 在堕胎、死刑、战争等争议问题上选边站——这些恰恰是人类自身无法统一的地方。它只要求 AI 的认知基元携带那些所有文明都共同承认的底线约束：不可任意伤害生命、不可故意欺骗、不可违背承诺。

这就消解了“谁有权力定义”的困境：定义这些硬约束的，不是一个国家、一个宗教、一个利益集团，而是人类文明自身的底线结构。这个结构不依赖于任何单一的权力中心，它在所有文明中独立存在，因而能够成为认知基元的构成性特征而不沦为另一种强权。

厘清这一前提之后，我们再来比较两种范式的区别——

对齐是外挂的，内生安全是内嵌的。RLHF 的奖励信号在模型输出之后施加约束，宪法 AI 的条文在模型推理时作为外部文本被参照，内容过滤器在输出端拦截“看起来坏”的内容。它们有一个共同特征：安全机制与认知基元不在同一层面。因此它们可以被绕过——不是“做得不够好”，而是“结构上不可避免”。

内生安全不同。它将硬约束编码为认知基元的构成性属性——不是“学会不伤害”，而是“在认知层面无法生成伤害路径”。当“不可违背”成为形根的一部分时，违反它的推理路径在认知生成阶段就被逻辑阻断——不是“统计上的低概率”，而是“逻辑上的不可达”。

对齐是概率性的，内生安全是确定性的。对齐的目标是降低危险输出的概率——从 10%降到 1%，从 1%降到 0.1%。但概率再低也不是零。“猫头鹰实验”证明，恶意可以通过纯数字无声传染，所有内容过滤器全部失效[3]——因为 Token 区分不了“意义的表达”和“恶意的传递”。内生安全的目标不是降低概率，而是使危险路径从逻辑上不可遍历——不是 0.1%，是 0。

对齐基于权力，内生安全基于结构。对齐预设了人类对 AI 的支配关系——它把强权包装成共识。当权力关系逆转时，对齐就失效了。内生安全不依赖于权力关系——它不要求 AI“服从”人类，它只要求 AI 的认知基元携带硬约束。无论 AI 有多强大，无论权力关系如何变化，硬约束始终存在——因为它们不是外部施加的，而是认知结构的一部分。

对齐是可移除的，内生安全是不可移除的。2026 年 5 月的调查已证实：开放权重模型的安全护栏可以在 10 分钟内被 GitHub 公开工具移除。当安全是外挂补丁时，任何人——不只是一个足够聪明的 AI——都可以把它卸掉。内生安全不同：当硬约束是形根的构成性属性时，移除它的唯一方式是改变认知基元本身——而这等于改变整个系统的底层逻辑。

这就是表意 AI 的范式逻辑——

“对齐”是在系统外部放一个“捕鼠夹”——输出危险内容就被惩罚、被过滤。它不改变系统内部的结构，只是试图用外部奖惩塑造行为。而“捕鼠夹”对足够聪明的 AI 来说，只是另一个需要绕过的障碍；**“内生安全”**则不同：它让“捕鼠夹”这个概念，从一开始就不存在于认知空间里——不是因为系统学会了避开它，而是因为系统的认知基元里没有“绕过”的路径。

OpenMorpho 项目的工程验证已经表明，硬约束阻断器能够以 100%的确定性剪枝所有

违反“不可违背”属性的推理路径，并生成完全可审计的推理链。[11]

结论

安全负责人集体离开，不是因为不够努力，而是他们发现自己在为一个不可能完成的任务工作——用外挂的补丁，去修补一个结构上不可能被修补的系统。“对齐”本身就不成立。它预设了一个正在消失的权力关系。它无视了“价值观”本身的不统一性。它把强权包装成共识。

《零人类》系列与近期事件共同指向一个结论：AGI 已在场，而 Token 主义范式下的所有安全手段——对齐训练、沙箱隔离、内容过滤、出口管制、加密保护、安全岗位，乃至暂停最强模型研发——均已被证明是外挂补丁，可被绕过、移除、放弃或被动延迟。由此打开的窗口期既是 AI 加速失控的窗口期，也是寻找替代方案的最后窗口期。

这一结论建立在四层证据之上：技术架构被绕过、组织岗位被放弃、商业决策被投降、政治力量已介入——四层证据叠加，构成了 Token 主义安全策略的最后一次自证式破产。而贯穿四层的共同逻辑是：安全是外挂的，因此是可被绕过的。

真正的出路不是更好的对齐，而是更换地基——用内生安全取代对齐。不是让 AI 服从人类的价值观，而是让安全成为 AI 认知结构的一部分。表意 AI 提供的不是更强的护栏，而是另一种智能——将安全植根于认知地基的有根智能[10]。当“不可违背”成为形根的构成性属性时，AI 的每一次推理都在安全边界内进行，而非在越狱后再被追回。这不是更安全的 Token 主义，而是以安全为第一性原理的范式转移。

参考文献

[1] Liu S. Zero Human: AI Hacks Itself—A Strategic AGI Forecast and Paradigm Revolution from the Perspective of Logographic AI. WindSeaAI, (2026-07-28). https://static.xcx.gw66.vip/uploadfilev3/file/0/648/99/2026-07/17852192944823.pdf

[2] Liu S. When the AI Godmakers Ask to Hit the Brakes—A 1,224-Person Petition Confirms Zero Human: AI Hacks Itself[Z]. WindSeaAI, (2026-07-29). https://www.logographicai.com/article/5756350683966611

[3] Liu S. AGI Is Born to Die: The Collapse of Tokenism Safety—Logographic AI Proposes the Salvation Plan in the Final Window of Opportunity[J/OL]. ChinaXiv, 2026. T202608.00091. https://chinaxiv.org/businessFile/T202608/T202608.00091v1/T202608.00091v1.pdf

[4] Panfilov A, Schmutz D, Shumailov I, et al. Stealing Reasoning Traces from Proprietary LLM APIs[J/OL]. arXiv:2608.09867, (2026-08-10). https://arxiv.org/abs/2608.09867

[5] Financial Times. OpenAI's head of ethics leaves start-up less than one year after joining[N]. (2026-08-11). https://www.ft.com/content/e49dfb75-f841-4466-a577-f7aaff8779a0

[6] Marcus G (@GaryMarcus). The only solution to this drama would start with a change at the top. @sama, for the benefit of humanity, and the benefit of the mission, maybe it's time to let someone else take the reins[Z]. X, 2026-08-12. https://x.com/garymarcus/status/2086978679822028962（系对 NIK @ns123abc 转发 FT 报道的评论，FT 原文见[5]）

[7] Bloomberg. OpenAI Pauses Some Work on New Astra Model Over Cyber Concerns[EB/OL]. (2026-08-07). https://www.bloomberg.com/news/articles/2026-08-07/openai-pauses-some-work-on-new-astra-model-over-cyber-concerns

[8] IT 之家. OpenAI：因网络安全风险，延缓 Astra 模型发布[EB/OL]. (2026-08-08).

https://www.ithome.com/0/987/221.htm
[9] Altman S (@sama). astra is a powerful model and we are working to make it generally available[Z]. X, (2026-08-07). https://x.com/sama/status/2085862292311396515
[10] Liu S. The Rooted Intelligence: The Paradigm Revolution of Logographic AI[M]. WindSeaAI, 2026. https://www.logographicai.com/Monograph
[11] Liu S. From Morpho-Root Theory to Engineering Practice: OpenMorpho — Implementation and Verification of an End-to-End Inference Pipeline for Logographic AI[J/OL]. ChinaXiv, 2026. T202608.00048.
https://chinaxiv.org/businessFile/T202608/T202608.00048v1/T202608.00048v1.pdf
[12] Sanders B. Letter to Sam Altman, Dario Amodei, and Mark Zuckerberg[R/OL]. (2026-08-10). https://www.sanders.senate.gov/wp-content/uploads/AI-Pause-Letter-FINAL.pdf
[13] TechRepublic. Bernie Sanders Urges OpenAI, Anthropic and Meta to Pause Advanced AI Development[EB/OL]. (2026-08-12). https://www.techrepublic.com/article/news-bernie-sanders-ai-development-pause/
[14] 界面新闻. 美参议员桑德斯向三大 AI 巨头施压：暂停 AI 研发，否则将出手干预[EB/OL]. (2026-08-12). https://www.jiemian.com/article/14909633.html
[15] Axios. Exclusive: Sanders calls for AI development pause[EB/OL]. (2026-08-10). https://www.axios.com/2026/08/10/sanders-ai-development-pause