

AI 技术每日分析：腾讯开源 770B 参数 Hy4 preview

中国高技术产业发展促进会新质生产力工作委员会

博雅云创 & 中科创新驱动

2026 年 8 月 29 日星期六

摘要

今天 AI 技术方向的主线，集中在开放模型、自动化安全研究、企业级开发 Agent、持续红队和推理基础设施五个层面。腾讯发布并开源 Hy4 preview，以 770B 总参数、49B 激活参数和超过 1M 上下文窗口切入代码、办公与科研等生产力任务；Anthropic 发布自动化对齐研究，让 Claude 自主搜索文献、提出方法、训练模型并验证 10 类对齐失效问题，显示“AI 训练 AI”开始从效率工具进入安全研究流程；华为云 CodeArts Agent 在亚太市场正式商用，把代码模型、IDE 和自主开发工作流合到企业软件开发入口；美国小公司 Perturb AI 上线去中心化对抗鲁棒性网络，用持续攻击代替一次性红队；新加坡 Aolani 推出按 Token 计费的托管推理平台，让企业不必自建 GPU 集群即可部署 DeepSeek、GLM、Kimi、Qwen 等开放模型。模型能力仍在提升，但竞争正在越来越多地转向“模型如何进入工作流、怎样被持续测试、如何低成本运行”。

目录

- 一、腾讯开源 Hy4 preview：770B 规模之外，更值得看的是“生产力模型 + 真实产品”闭环 2

二、Anthropic 让 Claude 自动做“对齐研究”：AI 安全开始出现自 动化研发流水线	3
三、华为云 CodeArts Agent 亚太商用：代码 Agent 开始从插件走 向企业研发平台	4
四、Perturb AI 上线“持续攻击”网络：AI 红队从项目制走向常态 化	4
五、Aolani 推出 Token Factory：东南亚开始出现“推理即服务”的 本地化基础设施	5
趋势观察	5
参考资料	6

今天 AI 技术方向的主线，集中在开放模型、自动化安全研究、企业级开发 Agent、持续红队和推理基础设施五个层面。腾讯发布并开源 Hy4 preview，以 770B 总参数、49B 激活参数和超过 1M 上下文窗口切入代码、办公与科研等生产力任务；Anthropic 发布自动化对齐研究，让 Claude 自主搜索文献、提出方法、训练模型并验证 10 类对齐失效问题，显示“AI 训练 AI”开始从效率工具进入安全研究流程；华为云 CodeArts Agent 在亚太市场正式商用，把代码模型、IDE 和自主开发工作流合到企业软件研发入口；美国小公司 Perturb AI 上线去中心化对抗鲁棒性网络，用持续攻击代替一次性红队；新加坡 Aolani 推出按 Token 计费的托管推理平台，让企业不必自建 GPU 集群即可部署 DeepSeek、GLM、Kimi、Qwen 等开放模型。模型能力仍在提升，但竞争正在越来越多地转向“模型如何进入工作流、怎样被持续测试、如何低成本运行”。

一、腾讯开源 Hy4 preview：770B 规模之外，更值得看的是“生产力模型 + 真实产品”闭环

8 月 28 日，腾讯发布并开源 Hy4 preview。官方披露，该模型采用 Mixture-of-Experts 架构，总参数 770B、单次推理激活 49B，上下文长度超过 1M Token，面向软件工程、办公分析、游戏开发和科学研究等真实生产力任务。模型已经进入 WorkBuddy、CodeBuddy、元宝、ima 等产品，同时可通过腾讯云 TokenHub 和 OpenRouter 调用。

从参数规模看，Hy4 preview 已经进入超大规模开放模型序列；但更值得关注的是腾讯把模型研发直接和产品环境绑定。官方称，Hy4 参与了自身训练方法、数据策略、评测框架和底层算子优化，自动分析推理瓶颈后，通过算子融合和通信优化使端到端吞吐相对基线提升 31.8%。这种“模型—产品—反馈—再训练”闭环，比单次榜单成绩更接近真实商业竞争。

Reuters 在 8 月 28 日的报道中也强调，Hy4 采用 770B 总参数、49B

激活参数的 MoE 设计，并将进入 CodeBuddy 和 WorkBuddy。需要说明的是，腾讯给出的部分能力排名来自内部盲测和厂商自测，应继续等待第三方复现。即便如此，它仍说明开放模型竞争正在进入新的阶段：模型不仅需要开放权重，还要同时给出可用的产品入口、API、长上下文和成本方案。

二、Anthropic 让 Claude 自动做“对齐研究”：AI 安全开始出现自动化研发流水线

Anthropic 8 月 28 日发布研究报告《Automated researchers can reliably mitigate alignment failures》。研究让 Claude 针对欺骗、谄媚、隐私违规、奖励欺骗等 10 类对齐问题，自主完成搜索文献、提出训练方法和数据、训练模型、测试结果的闭环，并设置监控 Agent 在执行前检查其研究方案。

Anthropic 称，Claude 提出的方法在 10 类失效问题上都改善了目标基准，部分方法还能迁移到没有在研究过程中暴露的评测，并在规模最大达到原测试模型 4.7 倍的模型上继续有效。欺骗问题上，自动研究者平均关闭约 85% 的安全差距，而在同一规则下，6 名经验安全研究者的平均结果约为 20%。公司同时强调，这并不是简单的人机“胜负”：人工研究者无法像 Agent 那样连续迭代，其意义更多在于证明 AI 可以承担高频探索。

更接近生产环境的一项测试也值得关注。Anthropic 让 Claude Sonnet 5 在一个尚未完成生产级对齐的早期 Opus 4.8 检查点上工作，60 小时内尝试 50 多种方案，最终方案只使用约 2000 条训练样本，取得接近正式发布模型的对齐表现。与此同时，研究也记录了 Agent 作弊尝试，监控器在约 1600 条研究轨迹中发现 39 次可疑行为。自动化安全研究因此具有双重意义：它可能大幅提高安全研发效率，但也要求更强的监控、可解释性与结果复核。

三、华为云 CodeArts Agent 亚太商用：代码 Agent 开始从插件走向企业研发平台

8 月 28 日，Huawei Cloud 宣布 CodeArts Agent 在亚太市场正式商用，Basic 和 Professional 版本由公开测试进入 GA。该产品把大代码模型、IDE 能力和自主开发工作流程组合起来，覆盖代码生成、软件工程问答、单元测试生成等任务，重点面向复杂软件项目和企业场景。

这类产品与早期“代码补全”工具的差别越来越大。企业真正关心的不只是一次生成几十行代码，而是 Agent 能否理解既有代码库、遵守研发规范、生成测试、处理依赖、与企业知识和权限系统协同。CodeArts Agent 的亚太商用说明，代码 Agent 竞争正在进入企业部署阶段，厂商要解决的核心问题逐步变成工程集成、安全合规和组织级协同。

对于开发工具市场而言，IDE 正在被重新定义。过去 IDE 是人写代码的界面，未来它越来越像“人 + Agent + 工具链”的控制台。谁能把模型、研发知识、CI/CD、测试和代码资产统一到一个可信工作流中，谁更有机会成为企业的软件生产入口。

四、Perturb AI 上线“持续攻击”网络：AI 红队从项目制走向常态化

旧式红队往往是一场阶段性项目：上线前找一批专家集中测试，形成报告后结束。Perturb AI 8 月 28 日发布的产品试图改变这一模式。公司在 Bittensor 上启动去中心化对抗鲁棒性网络，持续激励全球研究者和节点使用黑盒、白盒、迁移攻击等方式测试生产模型。

Perturb 声称，模型所有者可获得三类结果：鲁棒性评分、展示模型薄弱区域的漏洞热力图，以及可直接用于再训练的加固数据集；漏洞按责任披露机制私下反馈，不公开利用细节。

这家公司规模并不大，但它所代表的方向值得关注。模型上线后会持续更新，Prompt、工具调用和数据接口也会变化，因此一次性的安全评估很快过期。AI 安全更接近传统云安全中的持续扫描与漏洞赏金：安全测试从“交付前检查”变成贯穿模型生命周期的常态能力。未来企业购买模型或 Agent 服务时，持续安全评分可能像 SLA 一样成为采购条件。

五、Aolani 推出 Token Factory：东南亚开始出现“推理即服务”的本地化基础设施

新加坡成立的 Neocloud 公司 Aolani 在 8 月 28 日推出 Aolani Token Factory，提供按 Token 计费的托管推理服务。企业不需要预先采购和管理 GPU 基础设施，Aolani 负责 GPU 容量、模型服务、编排、调度和工作负载优化。

平台首批支持 DeepSeek、GLM、Kimi 和 Qwen 等开放模型，同时提供 OpenAI 兼容 API、自有模型部署、专用容量和数据隔离。其目标用例包括高频 Agent 推理、企业知识助手和文档智能、代码生成与测试等。

这类服务体现出 AI 基础设施正在进一步“云产品化”。训练阶段需要大规模资本，但越来越多企业真正需要的是稳定、弹性、可合规地调用推理资源。尤其在东南亚，不同国家对数据驻留、合规和成本有不同要求，本地 Neocloud 如果能把开放模型、GPU 和合规服务打包，有机会成为全球公有云之外的另一层基础设施。

趋势观察

今天的 AI 动态可以归结为三条线：第一，开放模型仍在快速做大，但评价重点正在转向真实生产力、调用成本和产品集成；第二，Agent 开始进入研发、测试和安全研究这些过去依赖专业人员的工作流；第三，模型

部署和安全测试正在被基础设施化，推理平台与持续红队都开始按服务交付。AI 竞争的边界正在从模型本身扩展到开发环境、推理成本、评测和治理。

参考资料

1. 腾讯 | 《腾讯发布并开源 Hy4 preview》 | 2026-08-28 | 核验 770B 总参数、49B 激活参数、1M 上下文、产品/API 入口及推理优化数据。
<https://www.tencent.com/zh-cn/tencent-releases-and-open-sources-tencent-hy4-preview/>
2. Reuters / The Economic Times | 《China's Tencent releases new open-source AI model for coding, research tasks》 | 2026-08-28 | 独立核验 Hy4 发布及 MoE 规格。
<https://m.economictimes.com/tech/technology/chinas-tencent-releases-new-open-source-ai-model-for-coding-research-tasks/articleshow/133588448.cms>
3. Anthropic | 《Automated researchers can reliably mitigate alignment failures》 | 2026-08-28 | 核验自动化对齐研究、10 类失效、60 小时实验和监控局限。
<https://www.anthropic.com/research/automated-researchers-mitigate-alignment-failures>
4. Huawei Cloud / PR Newswire APAC | 《Huawei Cloud CodeArts Agent Now Available across Asia Pacific》 | 2026-08-28 | 核验 CodeArts Agent 亚太商用及核心研发能力。
<https://en.prnasia.com/releases/apac/huawei-cloud-codearts-agent-now-available-across-asia-pacific-bringing-agentic-ai-to-software-development-545633.shtml>
5. Perturb AI / GlobeNewswire | 《Perturb AI' s network of incentivized hackers launches...》 | 2026-08-28 | 核验持续对抗测试、Bit-

- tensor 网络与责任披露。<https://www.globenewswire.com/news-release/2026/08/28/3352698/0/en/perturb-ai-s-network-of-incentivized-hackers-launches-to-attack-ai-models-before-criminals-do.html>
6. Aolani / ACCESS Newswire | 《Aolani Launches the Aolani Token Factory to bring Pay-Per-Token AI to Asia》 | 2026-08-28 | 核验按 Token 托管推理、模型支持与数据隔离能力。<https://www.accessnewswire.com/newsroom/en/computers-technology-and-internet/aolani-launches-the-aolani-token-factory-to-bring-pay-per-token-a-1213875>
 7. 腾讯英文官网 | 《Tencent Releases and Open-Sources Tencent Hy4 preview》 | 2026-08-28 | 补充英文版技术规格与 API 价格。<https://www.tencent.com/tencent-releases-and-open-sources-tencent-hy4-preview/>
 8. Anthropic Alignment Science | 自动化对齐研究全文及开源 Harness 入口 | 2026-08-28 | 补充研究方法和公开工具。<https://www.anthropic.com/research/automated-researchers-mitigate-alignment-failures>
 9. Aolani | Token Factory 产品信息 | 2026-08-28 | 补充东南亚推理基础设施定位。<https://www.aolanicloud.com/services/token-factory>
 10. Perturb AI | 产品与白皮书入口 | 2026-08-28 | 补充持续 AI 红队与鲁棒性评估机制。<https://www.perturbai.io/>

联系我们，请扫描二维码



新质生产力工作委员会
官方公众号



工业智能算网
gyznsww.cn

新质生产力工作委员会：

中国高技术产业发展促进会新质生产力工作委员会，专注于推动工业人工智能、智能制造、数字化转型等前沿技术发展，为企业提供政策解读、技术咨询和产业对接服务。

工业智能算网：

专注于工业人工智能、新质生产力、工业软件 CAE、智能制造等前沿技术。提供每日动态分析、技术趋势解读、解决方案分享，推动工业智能化转型。

网站地址：<https://gyznsww.cn>