AI 技术每日分析

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 6 月 25 日

摘要

今日 AI 技术主线集中在推理算力、企业智能体基础设施、Agent 身份治理、开源模型压缩与 AI 科研工具化。OpenAI 与 Broadcom 发布 Jalapeño 推理芯片，表明前沿模型竞争正在向自研硬件和垂直基础设施延伸；AWS 继续补齐企业 Agent 所需的上下文、知识图谱、代码安全与治理能力；Linux Foundation 拟推出 Agent Name Service，将 AI Agent 身份认证问题纳入开放标准；Multiverse Computing 发布 Pulsar 16B，显示中等规模开源推理模型正在通过压缩和优化进入更多企业环境；SandboxAQ 推出 GPCR 虚拟筛选方案，代表 AI for Science 从演示走向专门工作流。

Contents

一、OpenAI 与 Broadcom 发布 Jalapeño，推理芯片成为模型公司的基础设施战场

OpenAI 于 6 月 24 日展示其与 Broadcom 共同设计的首款自研 AI 推理芯片 Jalapeño。Reuters 报道，该芯片面向 ChatGPT 等应用的推理任务，计划在 2026 年底前部署；Broadcom CEO Hock Tan 称其性能可与 NVIDIA Blackwell 和 Google TPU 相比较。OpenAI 还披露，Jalapeño 样片已在实验室运行，并与 GPT-5.3-Codex-Spark 模型达到目标功耗与性能。

这条新闻的核心不是“又一颗 AI 芯片”，而是模型公司开始把推理成本、模型服务质量和硬件路线纳入自身控制。训练决定模型上限，推理决定商业可持续性。随着对话、编程、搜索和 Agent 调用频次上升，推理成本会直接影响毛利率、响应速度和产品价格。OpenAI 自研推理芯片，意味着大模型公司正在从“模型供应商”转向“模型—芯片—系统—云服务”的垂直基础设施竞争。

二、AWS 强化企业 Agent 栈，竞争焦点转向上下文、治理和交付安全

AWS 在纽约峰会公布多项 Agent 能力，包括 AWS Context、AWS Continuum、Kiro、AWS DevOps Agent、AWS Transform 以及 Amazon Bedrock AgentCore 增强。AWS 称，AWS Context 为企业数据构建知识图谱，让 Agent 知道应访问哪些信息，同时内置治理机制；AWS Continuum 则面向代码漏洞，提供持续发现、优先级判断、验证和修复能力。

这显示企业 Agent 落地的难点，已经从“能不能调用工具”转向“能不能持续、安全、可治理地完成工作”。企业真实场景里，Agent 需要访问多源数据、遵守权限边界、记录操作过程、避免越权和幻觉，并与 DevOps、客服、文档、知识库、代码仓库连接。AWS 的动作说明，云厂商正在把 Agent 做成企业级基础设施，而不是单一聊天产品。

三、Linux Foundation 拟推出 Agent Name Service，Agent 身份治理进入开放标准阶段

Linux Foundation 于 6 月 23 日宣布拟推出 Agent Name Service（ANS），试图基于现有 DNS 基础设施，为互联网上运行的 AI Agent 提供可信身份、验证和发现机制。ANS 目标包括识别 Agent 代表谁、具有什么权限、代码和运行历史是否保持真实未篡改，并避免依赖中心化或封闭注册表。

这条动态反映了 Agent 走向生产后的安全基础问题。未来企业可能会同时调用来自不同平台、不同组织、不同权限等级的 Agent。如果没有标准身份体系，企业无法判断某个 Agent 是否可信、是否来自授权主体、是否被篡改。ANS 把 Agent 身份问题放到 DNS 和开放标准层面，是 AI Agent 从“应用功能”走向“互联网基础协议”的重要信号。

四、Pulsar 16B 发布，中等规模开源推理模型继续向低成本部署演进

Multiverse Computing 于 6 月 23 日发布 Pulsar 16B。这是一个 16.15B 参数开源推理模型，基于 NVIDIA Nemotron 架构压缩优化而来，并在 Hugging Face 以 Apache 2.0 许可发布。发布材料称，Pulsar 16B 在约一半参数量下保持 30B 级别推理能力，并在 Blackwell GPU 上取得更高吞吐和更低首 Token 延迟。

对企业用户而言，这类模型的重要性在于“可部署性”。很多业务不需要最大模型，而需要在有限显存、单机、私有化环境或低延迟场景中稳定运行。Pulsar 16B 代表的趋势是：开源模型竞争不只是参数规模竞赛，还包括压缩、吞吐、长上下文保持、工具调用接口和部署成本优化。

五、SandboxAQ 推出 GPCR 虚拟筛选方案，AI for Science 转向专用工作流

SandboxAQ 于 6 月 24 日宣布推出面向 GPCR 药物发现的虚拟筛选方案，并由 NVIDIA BioNeMo Agent Toolkit 加速。该方案不仅预测分子是否与受体结合，还试图预测其是否激活或阻断受体活性。GPCR 是药物研发中的关键靶点家族，但结构状态和功能机制复杂，传统筛选成本较高。

这表明 AI for Science 正在从通用大模型问答转向专业科研工作流。药物发现并不只需要“找相似分子”，而要把结构、生物物理、药理机制、实验成本和候选筛选流程打通。未来科研 AI 的价值，不在于生成漂亮答案，而在于缩短实验循环、降低筛选成本，并把高维科学假设转化为可验证候选。

六、开源供应链中的 AI 编码 Agent 开始可量化

arXiv 6 月 23 日上线论文《Detecting AI Coding Agents in Open Source》，研究者对 World of Code 中 1.8 亿多个 Git 仓库进行多方法识别，发现单靠 bot 账号远远低估 AI 编码 Agent 活动；研究识别到 Claude Code 一次快照中 85 万余个相关提交，而 bot 账号方法只能找出约 3.3%。

这说明 AI 编码 Agent 已进入开源供应链，但其活动痕迹并不完全显性。未来开源项目需要新的披露规则、审计机制和代码 provenance 管理。企业采用 AI 编程工具，也不能只看效率提升，还要关注责任归属、许可证风险、代码审查压力和供应链安全。

参考文献

- Reuters ｜ OpenAI unveils custom chip it designed with Broadcom to boost its AI infrastructure ｜ 2026-06-24 ｜用于 Jalapeño 推理芯片分析。
- About Amazon ｜ AWS Summit New York 2026: New AI agent innovations ｜ 2026-06 ｜用于 AWS 企业 Agent 基础设施分析。
- Linux Foundation ｜ Linux Foundation Announces Intent to Launch Agent Name Service ｜ 2026-06-23 ｜用于 Agent 身份标准分析。
- AIwire ｜ Multiverse Computing Launches Pulsar 16B in Collaboration with NVIDIA ｜ 2026-06-23 ｜用于开源模型压缩与部署分析。
- AIwire ｜ SandboxAQ Launches Virtual Screening Solution for GPCR Drug Discovery ｜ 2026-06-24 ｜用于 AI for Science 分析。
- arXiv ｜ Detecting AI Coding Agents in Open Source: A Validated Multi-Method Census of 180 Million Repositories ｜ 2026-06-23 ｜用于 AI 编码 Agent 供应链分析。

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
gyznsw.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://gyznsw.cn