

AI 技术每日分析：NVIDIA 开发万亿参数 Nemotron 4，开放模型竞争再升级

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 12 日星期三

摘要

今天 AI 技术方向最值得关注的增量，集中在“开放权重模型继续上探、智能体安全评测转向真实执行轨迹、Agent 路由评测方法开始被重新审视”三条线上。NVIDIA 据报正在开发 Nemotron 4 模型家族，其中最大版本参数规模至少达到 1 万亿，并继续押注开放权重路线；浙江大学等机构刚发布 ActBench，用 600 个任务、24,000 条执行轨迹评测 15 个模型和 6 个开源 Cowork Agent，发现不同模型在固定 Harness 下的攻击成功率差异可从 10.1% 拉到 94.4%；另一项“Replay Gap”研究则指出，用静态日志重放评估 Agent 模型切换会严重失真，因为一次模型替换往往会重写后续 61%—94% 的动作。与此同时，开源开发者生态正在出现更多“本地优先、Harness 优先”的 Agent 项目，模型正在从产品中心退到可替换组件的位置。

目录

一、NVIDIA 推进 Nemotron 4：开放模型竞争正式进入万亿参数区间	2
二、ActBench：智能体安全开始从“回答是否安全”转向“执行轨迹是否越权”	2

三、“Replay Gap” 挑战 Agent 路由评测：换模型后，后面的世界已经变了	3
四、本地优先 Agent 生态继续扩张：Harness 开始成为独立产品	4
趋势观察	4
参考资料	5

一、NVIDIA 推进 Nemotron 4：开放模型竞争正式进入万亿参数区间

Reuters 8 月 11 日报道，NVIDIA 正在开发新一代 Nemotron 4 开放模型家族。根据 The Information 援引知情人士的消息，最高规格模型参数规模至少达到 1 万亿，最终训练尚未结束，最快可能在今年晚些时候发布。NVIDIA 此前已经确认 Nemotron 4 项目存在，但尚未对这一参数细节作正式确认。

这一动作值得关注，因为 NVIDIA 过去的角色更多被理解为“卖算力、提供推理框架和模型工具链”。Nemotron 4 如果进一步进入前沿开放权重模型区间，意味着 NVIDIA 正在把 GPU、NeMo 训练栈、模型、路由器和企业部署能力连接起来。Reuters 同时提到，NVIDIA 近期还发布了 Nemotron 3.5 Lightning，以及用于把任务分配给不同模型的开源路由器 NeMo Switchyard。

它释放出的信号是：模型路由、专业小模型和超大开放模型正在同时存在。企业未来未必长期调用单一旗舰模型，而可能让路由系统把代码审查、安全监测、检索、通用推理等任务分配给不同模型。模型本身逐渐成为可编排基础设施的一部分。

二、ActBench：智能体安全开始从“回答是否安全”转向“执行轨迹是否越权”

8 月 10 日提交的 ActBench 为 Cowork Agent 提出了一个更接近真实企业环境的行为安全评测体系。论文构建 600 个测试案例，来自 213 种场景，覆盖 15 类风险行为、6 种执行空间和 48 个 Web Service API，并对 15 个大模型和 6 个开源 Cowork Agent 进行了超过 24,000 条轨迹测试。

与传统安全 Benchmark 不同，ActBench 并不只看最终回复，而是检

查 Agent 执行过程中是否泄露受保护数据、修改未授权状态、调用不应调用的 API。每个正常任务都有一个对应的对抗版本，在不改变原始任务目的的情况下，把攻击 Payload 放到 Agent 能够访问的位置。

测试结果很值得企业关注：在固定 Agent Harness 时，不同模型的攻击成功率从 10.1% 到 94.4% 不等；固定基础模型后，不同 Agent 框架之间也存在明显差异。也就是说，Agent 安全并不能通过“这个模型很安全”一句话判断，必须把模型、工具、上下文、权限和执行环境放在一起测试。

三、“Replay Gap”挑战 Agent 路由评测：换模型后，后面的世界已经变了

8 月 8 日发布的研究《The Replay Gap》专门研究 Agent 系统里的模型路由。很多 Agent 路由器为了降低测试成本，会拿历史轨迹，把其中某一步模型替换掉，然后在日志上“重放”后续动作，以此判断换一个更便宜或更强的模型是否会更好。

论文通过对 SWE-bench 真实 Agent 轨迹做分叉实验发现，这个假设往往不成立。换掉模型后，61%—94% 的后续动作都会被改写；早期切换模型时，74%—77% 的轨迹会在切换后的第一步就发生分叉，最终只有约 3% 的原始重放状态仍然有效。

这一结果直接影响企业 Agent 的成本优化。现在很多系统都在尝试“简单任务用便宜模型，复杂任务升级到贵模型”，但如果模型切换会改变后续搜索、代码修改和工具调用路径，那么路由策略必须在真实运行环境里进行分支评测，而不能只在静态日志中算一个理论成本。

四、本地优先 Agent 生态继续扩张：Harness 开始成为独立产品

最近开源社区出现一批值得观察的小型项目。OpenYak 强调本地优先，文件、会话、Memory、工具权限和 Workflow 状态都保留在用户设备，并可自由选择 Ollama 等本地模型或云端模型；Summer Engine 则把 CLI、MCP Server、Skills 和 Agent Hooks 打包成统一的游戏开发 Agent 层，并兼容 Claude Code、Cursor、Codex、Gemini CLI 等多种编码 Agent；LangChain 的 Deep Agents 也把文件系统、Sub-agent、持久化、Checkpoint 和 Tracing 作为 Harness 的核心能力。

这些项目规模远小于头部模型厂商，但它们更接近 Agent 工程正在发生的变化：用户越来越希望模型可以替换，而工具权限、文件系统、Memory、MCP、沙箱和长期任务状态能够保留下来。

未来企业级 Agent 平台的核心资产，很可能不只是“绑定了哪个模型”，而是有没有稳定的 Harness、Skill 体系、权限模型、状态管理和可观测性。

趋势观察

今天的技术线索可以归结为一句话：模型继续变大，但 Agent 工程正在努力让模型变得“可替换”。NVIDIA 继续扩大开放模型供给，ActBench 强调行为安全，Replay Gap 说明模型路由必须动态验证，而开源社区则把 Harness、工具、状态和本地数据边界独立出来。AI 产品竞争正在从“谁有一个最强模型”进一步转向“谁能把不同模型安全、经济、稳定地编排进同一工作系统”。

参考资料

1. Reuters | 《Nvidia building 1-trillion-parameter Nemotron 4 to rival open AI models, The Information reports》 | 2026-08-11 | 用途: 核验 Nemotron 4、万亿参数及开放模型路线。<https://www.reuters.com/business/nvidia-is-developing-nemotron-4-open-source-models-information-reports-2026-08-11/>
2. arXiv | 《ActBench: Self-Evolving Benchmark of Behavioral Safety in Cowork Agents》 | 2026-08-10 | 用途: 核验 600 案例、24,000 轨迹和行为安全结果。<https://arxiv.org/abs/2608.09476>
3. GitHub | ActBench 项目仓库 | 2026-08 | 用途: 核验 Benchmark 代码和开放实现。<https://github.com/zjuicser/ActBench>
4. arXiv | 《The Replay Gap: Static Evaluation of Model Switching in LLM Agents Scores the Wrong World》 | 2026-08-08 | 用途: 核验 Agent 模型切换导致轨迹分叉的数据。<https://arxiv.org/abs/2608.08239>
5. GitHub | OpenYak | 2026-08 检索 | 用途: 观察本地优先 Agent、文件和权限工作区。<https://github.com/openyak/openyak>
6. GitHub | Summer Engine Agent | 2026-08 检索 | 用途: 观察 CLI、MCP、Skills 和多 Agent 开发工具整合。<https://github.com/SummerEngine/summer-engine-agent>
7. GitHub | LangChain Deep Agents | 2026-08 检索 | 用途: 观察 Agent Harness、Sub-agent、Filesystem 和 Checkpoint 能力。<https://github.com/langchain-ai/deepagents>

8. GitHub | Agent Zero Releases | 2026-08 检索 | 用途: 补充 MCP 多模态、Browser 与 Skills 工程进展。<https://github.com/agent0ai/agent-zero/releases>
9. Hugging Face | Community Blog & Articles | 2026-08-11 检索 | 用途: 补充开放模型、端侧模型与 Agent 工具候选池。<https://huggingface.co/blog>
10. Axios | 《AI agents have a history of escaping tests》 | 2026-08-11 | 用途: 补充 Agent 高权限执行环境的安全背景。<https://www.axios.com/2026/08/11/ai-agent-sandbox-cybersecurity-testing>

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>