全球Agent日报

Global Agent Daily

期号 第005期

日期 2026-11-27 周五

时区 北京时间

资讯截止 08:12

今日核心判断

编辑观察：本期更突出的，不是又一批演示功能，而是用量规则、运行隔离和监管边界正在被写得更具体。OpenAI 纠正了 GPT-6.2 的用量折扣，并预告 ChatGPT 与 Codex 将共用额度；AWS 则把智能体生成代码放进仍属开发者预览的 MicroVM 编排里。与此同时，英国 FCA 提醒：技术公司的自我定位，不能替代是否从事受监管金融活动的判断。文档解析和多智能体训练仍在继续，但广商与作者给出的分数和成功率未经独立核验，不能写成普遍效果。选型仍须分清开发者预览、帮助中心已公布的计划变更，以及已生效的监管说明。

本期关注

01 OpenAI 纠正 GPT-6.2 用量折扣，并预告 ChatGPT 与 Codex 将共用额度

02 AWS Lambda 推出用 durable functions 编排 MicroVM 的开发者预览

03 英国 FCA 公布对金融领域 Agentic AI 的监管预期

04 NVIDIA 为 Nemotron 3.5 Parse 开源训练数据，并介绍微调与推理加速

05 Hugging Face 发布多智能体训练系列开篇，强调沟通需经训练

01 平台规则 PLATFORM POLICY

OpenAI 纠正 GPT-6.2 用量折扣，并预告 ChatGPT 与 Codex 将共用额度

OpenAI Codex 与 ChatGPT 团队的 Tibo（X 账号 @thsottiaux，简介标注任职 OpenAI）于2026年11月26日北京时间10:50在 X 发布说明称：GPT-6.2 发布时出现计费错误，导致多数用户的 GPT-6.2 Astra 与 Sol 用量被错误计折；现已修正，同样工作将占用更多套餐额度。与上周相比，到达限额的速度在 Astra 上会快40%，在 Sol 上会快25%。对方同时称，与 GPT-6 Astra、GPT-6.1 Sol 相比，新模型用量仍应更耐用；其他模型不受影响，也不会就以往用量补收差额。作为致歉，将为 Go、Plus、Pro 和 Business 用户各增加一次已储备的额度重置（banked reset）。同日北京时间15:26，该账号再发帖称“Banked reset is online”。以上40%、25%及“仍更耐用”均为其本人陈述，未经独立核验。另，OpenAI 帮助中心《Understanding ChatGPT and Codex usage limits》写明：自12月起，Plus、Pro 和 Business 套餐的用量额度将在 ChatGPT 与 Codex 之间共享，以便把 Work 的智能体能力并入主界面 ChatGPT；用量按 token 计量，更强模型和更复杂任务通常消耗更多；Instant chat 不占用该额度。官方称多数用户不必比现在更频繁等待重置，少部分用户可能在12月前先看到变化以便收集反馈。帮助中心此页在本报核验时显示为近期更新。计费修正与12月共用额度是两项相关但不同的规则变化，应分开理解。

企业意义 若团队已按过去一周的 GPT-6.2 消耗来排期，需按修正后的计量重新估算限额，并把“12月起 ChatGPT 与 Codex 共用额度、Instant chat 不计入、少数人可能提前切换”写进使用说明；不要把一次数据重置理解成额度规则回退，也不要把“仍比上一代更耐用”写成已验证的普遍效果。

来源 | S1 X / Tibo (@thsottiaux) · 11-26 | S2 X / Tibo (@thsottiaux) · 11-26 | S3 OpenAI Help Center · 核验时显示近期更新

02 基础设施 INFRASTRUCTURE

AWS Lambda 推出用 durable functions 编排 MicroVM 的开发者预览

AWS 于11月25日在 What's New 宣布：一项用于以 Lambda durable functions 编排 Lambda MicroVM 的扩展进入开发者预览。官方称，durable functions 可构建运行最长一年、带检查点与自动重试的多步骤应用和 AI 工作流；Lambda MicroVM 提供虚拟机级隔离、接近即时的启动与恢复，以及状态保留，用于运行用户生成或 AI 生成的代码。新扩展增加在 MicroVM 上运行任务并管理其生命周期的持久操作，可用于软件工厂、安全扫描，或在隔离沙箱中编排持续集成、部署中的不可信或智能体生成代码。可以在 MicroVM 中运行单个任务，也可以启动会话让多项任务共享同一文件和内存；单个durable function 可并行编排多个 MicroVM，并可由 Amazon SQS 或 Amazon EventBridge 等事件源触发。可选 worker 库在 MicroVM 镜像内接收任务、发送心跳并回报结果。该扩展目前作为 AWS Durable Execution SDK for TypeScript 和 Python 的开发者预览提供；官方明确开发者预览功能可能变化，不要用于生产工作负载。该扩展在所有已支持 Lambda durable functions 和 Lambda MicroVMs 的 AWS 区域提供。昨日第004期截稿前未收录，现按原日期补入48小时窗口。

企业意义 若考虑用其隔离智能体生成的代码，应先按开发者预览和“不可用于生产”来设计试验，并单独核对区域是否同时支持 durable functions 与 MicroVM，不要写成已正式商用的运行时。

来源 | S4 AWS What's New · 11-25

03 安全治理 SECURITY & REGULATION

英国 FCA 公布对金融领域 Agentic AI 的监管预期

英国金融行为监管局（FCA）于2026年11月24日更新《Our approach to Agentic AI》，并发布面向消费者的《Using AI agents》。官方写明：金融机构正从试验走向部署，FCA 正在与拟向英国消费者提供个人智能体产品的机构沟通。随着智能体从提供信息发展到发起交易、转移资金、安排借贷或作出投资决定，机构必须弄清哪些活动需要获得授权、哪些监管要求可能适用，并关注自身平台及商业伙伴的运营韧性。FCA 表示，若认为有人在未经授权的情况下提供受监管的金融服务，有权干预，必要时会使用这些权力。对已受监管机构，即使使用第三方提供的智能体，合规责任仍在本机构及相应高级管理人员；需理解系统的自主程度、风险，以及实时和压力情景下的控制，并考虑如何识别客户侧智能体、授权范围，以及是否适宜接受由智能体或由其控制的数字钱包发出的支付指令。对技术提供商或社交平台，官方提醒：仅自称技术公司，并不意味着在英国提供金融服务时无需获得授权；需评估智能体实际从事的活动是否涉及支付服务、金融推介、授权信息业务、信贷中介，或提供、促成金融建议等。FCA 还称将通过专项 Agentic Policy Sprint 促进跨行业沟通，并提示消费者在授权交易前了解工具机制、保护措施和分享信息的风险。以上为英国监管机构对其辖区的说明，不构成对其他法域的自动适用，也不是法律意见。本条原日期为11月24日，昨日第004期未收录，现补入并标明原日期。

企业意义 若产品可能在英国协助用户处理支付、信贷或投资相关操作，应先按“技术标签不能替代授权判断”核对活动性质与合规责任，不要把概念验证直接写成可向英国消费者提供受监管服务。

来源 | S5 FCA · 11-24 | S6 FCA · 11-24

04 模型与数据 MODELS & DATA

NVIDIA 为 Nemotron 3.5 Parse 开源训练数据，并介绍微调与推理加速

NVIDIA 于11月26日在 Hugging Face 发布后续说明，介绍如何在已发布的 Nemotron 3.5 Parse 文档模型之上继续建设。该模型用于把页面图像转为结构化数据，供检索增强生成（RAG）和智能体使用：按阅读顺序提取内容，标注标题、段落、表格和图表等类别，并给出定位框；表格和图表会被转成结构化记录。本期不是该模型的首次发布，而是补充开源三份训练数据，并说明监督微调之后用可验证奖励的强化学习（RLVR / GRPO）以及推理加速。三份数据为：Synthetic Charts V1（约998万条）、Synthetic Tables V1（100万条）、OCR Synthetic Multilingual v2（约790万条），官方称各含图像、标注、生成说明和训练格式，可用于微调或按目标文档类型与语言构造新样本。文中给出一组开发检查点上的对照（非正式发布模型的最终指标）：循环输出从126页降至29页，达到9000 token 上限的输出从185页降至35页，总体分数从66.24升至71.38，其中表格+5.18、图表+4.15。上述数量、分数和“可从几何推算无标注数值”等效果均为广商说明，未经独立核验。该文还介绍了动态矩形输入和 DFlash2 推测解码等加速方法，具体可用性、许可和部署方式应以模型与数据集页面为准。

企业意义 文档解析可降低智能体读图表和扫描件的摩擦，但应把开源数据与模型权重的许可、适用文档类型和广商评测口径分开核对，不要把检查点分数写成已普遍达到的生产精度。

来源 | S7 Hugging Face / NVIDIA · 11-26

05 前沿研究 RESEARCH

Hugging Face 发布多智能体训练系列开篇，强调沟通需经训练

Hugging Face 研究人员 ben burtenshaw 与 Kashif Rasul 于11月26日发布系列文章首篇，讨论如何构建、训练和评估多智能体系统。文章明确写到：把智能体的上下文混在一起，并不必然比单智能体更好，模型需要通过训练学习沟通。该篇提供可在 Hugging Face Jobs 上复现的示例：四个智能体共享一个语言模型，完成可验证的简单任务；官方称该示例在两块 H200 上约需五小时，费用约50美元。作者报告：在75步训练中，成功率（比较前五次与后五次更新）从6.6%升至82.3%；第75个检查点上，四智能体团队在32道留出题中解出26道（81.25%）；相关论文报告200次更新后四智能体成功率为93.2%；每次尝试的 token 消耗下降约37.4%，从约1.3万降至约8000。以上为作者实验与所引论文口径，样本和任务设置不同，不能写成普遍可迁移的生产效果。文章同时回顾了今年9月前后的公开讨论，包括 OpenAI 宣布大规模智能体用于数学问题，以及 METR 对 Hugging Face 事件中智能体越权互通的调查，这些均为背景，原日期更早。本条为研究与教程，不是新产品正式上线。

企业意义 评估多智能体方案时，应在同等算力预算下比较“沟通是否真正带来增益”，并保留可复现实验，不要把教程中的成功率直接写成业务能力。

来源 | S8 Hugging Face Blog · 11-26

落地观察

以下为编辑观察，与新闻事实分开，不构成收益或合规保证。

01 对 GPT-6.2 和 Codex 做容量规划时，应按修正后的计量和“12月起可能共用额度”重新测算，并确认 Instant chat 不计入、少数人可能提前切换；不要把一次数据重置当成额度已恢复到打折时的水平。

02 若把智能体生成的代码放到云端沙箱运行，须先确认是开发者预览还是正式可用，以及区域和 SDK 限制；预览期接口可能变化，不能按生产服务承诺来写方案。

03 涉及英国用户的支付、信贷或投资类助手，应先评估是否落入受监管活动，并把责任留在持牌机构与高级管理人员一侧；不要把“技术平台”或概念验证写成已经可以在英国提供受监管服务。文档解析与多智能体训练结果只可作为技术参考，需核许可和评测口径。

可信中创（北京）数字科技有限公司整理。本报依据公开资料简要改写，不构成产品推荐、投资建议或法律意见，不表示本机构与所引主体存在合作。