AI 技术每日分析：Anthropic 据报以 34 亿美元锁定 1.7 吉瓦燃气发电设备

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 11 月 9 日

摘要

11 月 8 日，《金融时报》披露 Anthropic 通过约 34 亿美元安排取得 1.7 吉瓦燃气发电设备的订单与工程能力，AI 模型公司的竞争由 GPU 扩展至电力基础设施；开源开发者工具 Qwen Code Desktop 发布 0.27.0 版本，围绕 GitHub 交互、并行智能体与代码评审完善工程可靠性；视频模型社区关注 HyperFlow H3 Flash，以 13.7B 级压缩模型和少步采样降低视频与同步音频生成的硬件门槛；llama.cpp 继续更新本地推理适配；安全研究界则重新审视编码智能体许可记录中的凭证泄露风险。本期关注这些消息背后的供电成本、软件可靠性与工程治理。

Contents

1. Anthropic 拟以 34 亿美元锁定 1.7 吉瓦发电设备：AI 算力开始直接争夺电力交付能力 **2**
2. Qwen Code Desktop 发布 0.27.0：编码智能体继续补齐代码审查和会话可靠性 **2**
3. HyperFlow H3 Flash 将联合视频与音频生成压缩到 14B 量级 **3**
4. llama.cpp 持续更新端侧推理支持，开源生态保持高频迭代 **3**
5. 开发者安全组织分析 Claude Code 许可文件中的凭证暴露风险 **4**
趋势观察 **4**
参考资料 **5**

1. Anthropic 拟以 34 亿美元锁定 1.7 吉瓦发电设备：AI 算力开始直接争夺电力交付能力

11 月 8 日，英国《金融时报》报道，Anthropic 在 9 月末通过一项约 34 亿美元的交易，承接美国得克萨斯州设备供应商 OnePWR Solutions 此前订购的燃气发电机组，涉及约 1.7 吉瓦重型机组，以及相关工程安装与建设能力。路透社当天转述称，交易并未包括对 OnePWR 公司的整体收购，也没有员工转移。双方截至报道时尚未回应，路透社未能独立确认协议条款，因此该交易规模与细节目前仍以《金融时报》报道为依据。

这笔设备安排指向 AI 数据中心最现实的限制：电力能否如期交付。训练集群和推理平台采购 GPU 后，机房能否按时接电、能否拿到变压设备与冷却系统，成为影响实际投产时间的关键。大型燃气机组的交期、燃料合同、环保许可和供电稳定性都属于 AI 项目需要管理的工程条件。技术公司直接锁定设备订单，是在为供电时间表留出确定性。对模型服务采购方而言，算力供给的可持续性，也将成为评估服务商的重要变量。

2. Qwen Code Desktop 发布 0.27.0：编码智能体继续补齐代码审查和会话可靠性

QwenLM 开源项目于 11 月 8 日发布 Qwen Code Desktop v0.27.0，GitHub 发布记录列出大量更新，包括 GitHub 消息读取失败后的恢复机制、将 Pull Request 评审讨论提供给沙箱化检查流程、优化受管智能体的会话租约与工作树锁，以及强化部分工具调用及校验逻辑。这次更新着重处理编程智能体长期运行时的故障恢复、并行操作与评审衔接。

在企业研发环境中，Agent 经常需要同时读取多个仓库、创建分支、执行测试、提交修改，并等待人工审查。某个消息读取失败或子会话未能正确释放工作树锁，都可能使长任务中断或让审查结论失真。Qwen Code

这次修的，正是软件团队日常会碰到的故障。代码生成速度之外，任务状态能否保持、评审能否接续，也会影响团队是否愿意长期使用。

3. HyperFlow H3 Flash 将联合视频与音频生成压缩到 14B 量级

Video Rebirth 维护的 HyperFlow H3 Flash 项目近期开放了模型权重、代码与 ComfyUI 工作流，11 月 8 日继续在开发者社区传播。官方模型卡介绍，该项目基于 MiniMax-H3 构建压缩版本，核心视频音频生成模型规模约 13.7B 或 13.9B 参数，相比约 31B 的基线模型明显缩小；采用 4、6 或 8 步采样，默认 6 步，可输出 5.2 至 15 秒、24 帧每秒的视频及同步 32kHz 立体声音频。项目提供 INT8 与 GGUF 方案，Q4_K_M 版本模型文件约 9.1GB。开发团队报告，5.2 秒视频在单张 RTX 5090 上完成端到端生成约需 92 秒，该数值来自其指定硬件与测试流程。

模型规模和采样步数都降下来，本地生成视频时的显存与耗时压力随之减轻。模型文件变小，并不代表完整推理流程只需相同大小的显存：文本编码器、视频和音频解码器、临时缓存以及系统内存仍需另算。开发者还需要量化格式、推理代码和可复现工作流，才能在不同 GPU 上核算画质、音画一致性与成本。

4. llama.cpp 持续更新端侧推理支持，开源生态保持高频迭代

ggml-org/llama.cpp 的公开发布记录显示，11 月 8 日出现了多个预发布版本，围绕本地推理模型适配、后端计算与工具调用进行了连续维护。llama.cpp 负责让不同团队发布的模型权重能在 PC、服务器和边缘硬件上运行。新模型上线后的模型结构适配、量化配置、位置编码检查和后端内核优化，往往决定开源权重能否快速进入用户系统。

目前各个预发布构建的稳定性仍需要开发者按硬件环境验证。模型权重之外，转换工具、推理引擎、量化格式与应用框架都需要同步适配。对需要本地部署和成本控制的机构而言，能否稳定管理版本、复现测试并保留回退路径，直接影响开源模型的使用价值。

5. 开发者安全组织分析 Claude Code 许可文件中的凭证暴露风险

Cloud Security Alliance 旗下研究团队 11 月 8 日发布研究说明，梳理应用安全厂商 Apiiro 此前发现的 Claude Code 本地设置文件凭证暴露问题。Apiiro 称，在 104185 份公开的 ‘.claude/settings.local.json‘文件中，1122 份至少包含一项凭证，发现 1379 项不同秘密信息。问题与“以后不再询问”许可选项保存原始 Shell 命令有关：若命令中直接写入密码或密钥，审批记录可能把敏感字符串留在项目文件中，随后被错误提交至公开仓库。上述统计来自 Apiiro 抽样研究，CSA 没有独立重新复现实测。

便利的许可选项一旦把明文命令写进仓库，就会牵涉企业凭证管理。企业需要从流程上减少明文密钥进入命令行的机会，对已经暴露的密钥执行吊销与轮换，并检查历史提交和分支副本。随着编程智能体获得更多外部系统权限，权限配置文件和工具调用日志正成为安全审计的重点对象。

趋势观察

本期涉及发电设备订单、编码智能体、开源视频模型和本地推理工具。AI 规模化应用正在促使开发商重新评估供电、运行稳定性、硬件利用率与安全治理。采购方最终要看工程交付能否复现，以及成本能否算清。

参考资料

1. Financial Times | Anthropic strikes $3.4bn deal to generate gas-fired power for data centres | 2026-11-08 | 用于确认 1.7GW 设备采购报道、交易结构与时间
   https://www.ft.com/content/41bf5c6a-9735-4f1a-a8e3-2bcccad58ba0

2. Reuters | Anthropic secures $3.4 billion generator deal for AI data centers, FT reports | 2026-11-08 | 用于交叉核验、确认独立核查尚未完成
   https://www.reuters.com/business/energy/anthropic-secures-34-billion-generator-deal-ai-data-centers-ft-reports-2026-11-08/

3. QwenLM/GitHub | Qwen Code Desktop v0.27.0 | 2026-11-08 | 用于编码智能体桌面版更新、PR 审查与运行可靠性
   https://github.com/QwenLM/qwen-code/releases/tag/desktop-v0.27.0

4. Video Rebirth/GitHub | HyperFlow H3 Flash | 项目近期更新, 2026-11-08 社区传播 | 用于权重、推理工作流与模型规模
   https://github.com/Video-Rebirth/hyperflow-h3-flash

5. VideoRebirthAI/Hugging Face | hyperflow-h3-flash model card | 2026-11 月可访问版本 | 用于参数、采样步骤、显存和性能测试环境
   https://huggingface.co/VideoRebirthAI/hyperflow-h3-flash

6. ggml-org/GitHub | llama.cpp releases | 2026-11-08 | 用于本地推理引擎版本更新

https://github.com/ggml-org/llama.cpp/releases

7. Cloud Security Alliance AI Safety Initiative | Approved Once, Leaked Forever: Secrets in Claude Code Settings Files | 2026-11-08 | 用于编程智能体的许可与凭证安全
   https://labs.cloudsecurityalliance.org/research/csa-research-note-claude-code-settings-secret-leak-20261108/

8. GitHub Engineering | Inside GitHub's new Git storage architecture | 2026-11-05，背景 | 用于大规模智能体并行开发的基础设施背景
   https://github.blog/engineering/architecture-optimization/inside-githubs-new-git-storage-architecture/

机构介绍与关注

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。

关注高促会新质生产力工委会公众号

关注工业智能算网平台

https://gyznsw.cn