

AI智能体平台功能升级与成本管控成焦点

中文内部早报 · 5分钟版

今日总览

AI智能体平台在企业级功能和成本管控方面迎来重要升级。Anthropic的Claude Enterprise推出详细的用量分析和支出控制工具，标志着AI服务向精细化运营转变。xAI发布Voice Agent Builder测试版，进一步拓展语音智能体应用场景。

Agent和编程 workflow 领域快速发展。Google发布ADK Go 2.0，引入基于图的工作流引擎；Claude Code更新至v2.1.198版本，新增多项实用功能。新加坡初创公司Acti推出基于Google Gemini的智能体键盘，将AI能力直接集成到手机输入法中。

模型与多模态技术持续演进。Google DeepMind发布Nano Banana 2 Lite图像模型和Gemini Omni Flash视频生成模型，分别在速度和成本方面实现突破。Apple发布VideoFlexTok视频分词技术，采用粗到细结构提升视频处理效率。

大厂产业链布局加速。Meta计划推出Meta Compute云基础设施业务，将过剩AI算力对外销售；Runway开发deckard容量控制器，动态重分配GPU资源。Google Cloud Workbench Notebooks扩展正式上线，连接VS Code与云端Jupyter环境。

监管安全问题备受关注。Meta秘密测试竞品AI产品引发争议；Anthropic与五角大楼就Claude军事用途护栏产生分歧。评估方面，Senior SWE-Bench和Remote Labor Index显示AI智能体在高级工程任务中的能力持续提升。

头条

今日头条

1. Claude Enterprise新增用量与成本分析及支出管控功能

Claude 博客 · 07/03 08:00 · 大厂与产业链

Claude Enterprise推出更丰富的管理分析工具和成本控制功能。仪表板可按群组 and 用户分析用量与成本，支持SCIM群组筛选，展示制品创建、文件编辑、技能和连接器对应的成本。Claude Code管理控制台新增使用量和价值选项卡，Analytics API可将数据接入Datadog和CloudZero。管理员可设置支出限额警告通知。

2. Claude Code v2.1.198发布

Claude Code: GitHub Releases · 07/02 04:45 · 开源项目与开发者工具

Claude Code v2.1.198更新包含多项功能改进。Claude in Chrome现已全面可用，为claude agents新增后台智能体通知功能。新增/dataviz技能提供图表与仪表盘设计指导，Gateway增加AWS上的Claude Platform作为上游提供商。修复了网络断开导致响应中断等多个问题。

3. xAI 发布 Voice Agent Builder 测试版

xAI 新闻 · 07/01 08:00 · 智能体、编程与 workflow

xAI推出Voice Agent Builder测试版，这是一个基于Grok Voice的无代码平台，可在两分钟内创建生产级语音智能体。集成电话、知识检索、工具、MCP及可观测性功能，支持连接现有SIP号码、API和WebSocket。在t-voice Bench上，Grok Voice Think Fast 1.0得分67.3%，领先竞品。

4. Acti将AI智能体直接放入手机键盘

TechCrunch AI · 07/01 01:52 · 智能体、编程与 workflow

新加坡初创公司Acti发布基于Google Gemini的智能体键盘，可代替用户 in 应用中执行操作。核心功能Skills允许用自然语言创建快捷方式，如长按T键翻译消息、C键发送会议链接。采用本地优先架构，默认不访问私人消息。公司获530万美元种子轮融资，现已开放下载。

5. ADK Go 2.0发布：构建可靠的多智能体应用

Google 开发者博客 · 07/01 01:17 · 智能体、编程与 workflow

Agent Development Kit for Go 2.0发布，引入基于图的工作流引擎，用于组合复杂多智能体应用。新版本内置人工参与循环编排、使用纯Go代码的动态执行，以及指数退避重试等自动弹性特性。统一执行模型后，单智能体应用与复杂图均运行在同一运行时上，简化了遥测与状态持久化。

模型

前沿模型与多模态

1. Google DeepMind 发布 Nano Banana 2 Lite 和 Gemini Omni Flash

Google DeepMind: Blog · 07/01 00:02

Google DeepMind 推出 Nano Banana 2 Lite (gemini-3.1-flash-lite-image)，为 Nano Banana 系列速度最快、成本最低的图像模型，文本到图像输出仅需 4 秒，每 1K 分辨率图像成本 \$0.034，已上线 Google AI Studio、Gemini API 及消费者产品 (AI Mode in Search、Gemini app 等)。同时推出 Gemini Omni Flash (gemini-omni-flash-preview)，支持高画质视频生成与对话式编辑，视频输出定价 \$0.10/秒，面向开发者开放 API。

2. How ChatGPT adoption has expanded

OpenAI 新闻 · 06/30 17:00

How ChatGPT adoption has expanded.

3. The latest AI news we announced in June 2026

Google AI Blog · 07/02 02:15

June Pixel Drop hero

4. New York City educators and industry leaders gathered at Google’ s offices to shape the future of AI in classrooms.

Google AI Blog · 07/02 00:00

New York City educators and industry leaders gathered at Google’ s offices to shape the future of AI in classrooms.。

智能体 智能体、编程与 workflows

1. Google Health API 推出 CLI： ghealth 是一款针对 Fitbit 数据的开源工具

MarkTechPost · 07/03 08:00

ghealth 是一款封装 Google Health API v4 的开源命令行工具，以单个 Go 二进制文件发布（Apache 2.0 协议）。它提供 40 种已验证的数据类型（包括步数、心率、睡眠、体重、血氧饱和度、心率变异性等）的结构化 JSON 输出。工具采用 Agent 优先设计，具备确定性退出码、--dry-run 和 --raw 标志，并附带两个 SKILL.md 文件供 AI 智能体使用。用户需自行创建 OAuth 凭据，通过 PKCE S256 认证。数据来源覆盖 Fitbit、Pixel Watch 及连接的第三方设备。

2. Cloudflare 推出全新AI流量管理选项： 区分搜索、智能体与训练爬虫，保护广告页面

Cloudflare Blog · 07/01 21:00

Cloudflare 为所有网站所有者提供更精细的AI流量管控选项，取代一刀切的屏蔽方式。用户可轻松区分并管理搜索爬虫、AI智能体爬虫和训练爬虫，同时新增保护广告变现页面的能力。

3. Claude Desktop 推出 Linux 公测版

X: Claude Devs (@ClaudeDevs) · 07/01 00:06

Claude Desktop 现已在 Linux（Ubuntu 和 Debian）上推出测试版。除了浏览器和终端，你现在可以在所有付费计划中获得一流的桌面体验，包括 Claude Code、Claude Cowork 和聊天。

4. VideoFlexTok： 可变长度粗到细视频分词

Apple Machine Learning Research · 07/03 08:00

VideoFlexTok提出一种可变长度token序列的视频表示方法，采用粗到细结构——首个token捕捉语义和运动等抽象信息，后续token添加精细细节，生成流解码器支持任意 token 数量的视频重建。相比传统3D网格分词，该结构允许根据下游需求调整token数，在相同预算下编码更长视频。在类别和文本到视频生成任务中，VideoFlexTok以 1.1B参数（5.2B的1/5）达到可比生成质量（gFVD和ViCLIP Score）。训练一个处理10秒81帧视频的文本到视频模型仅需672个token，比同等3D网格分词器少8倍。

产业 大厂与产业链

1. Google Cloud Workbench Notebooks 扩展发布： 在 VS Code 中连接云端 Jupyter 环境

Google 开发者博客 · 07/02 01:17

Google Cloud Workbench Notebooks 扩展正式上线，开发者可在 VS Code 中直接连接可扩展的云端 Jupyter 环境，无需切换上下文即可利用高性能 Google Cloud 基础设施完成机器学习全流程。该扩展已完全开源，可在 GitHub 和 VS Code Marketplace 获取。

2. Meta效仿SpaceX，将过剩AI算力变现

TechCrunch AI · 07/01 21:43

据Bloomberg报道，Meta正计划推出云基础设施业务Meta Compute，对外出售AI计算能力和模型访问权限，直接与AWS、Google Cloud及Azure竞争。Meta已承诺未来几年投入1829亿美元建设AI基础设施，其中俄亥俄州数据中心（规模如曼哈顿）将于今年上线。新业务由基础设施主管Santosh Janardhan、Meta超级智能实验室负责人Daniel Gross和总裁Dina Powell McCormick领导。Meta可能效仿CoreWeave出售裸计算能力，并像AWS一样托管AI模型（包括近期发布的闭源模型Muse Spark）。扎克伯格此前已表示云业务"definitely on the table"。

3. Kimi K2.7 Code 已在 GitHub Copilot 上正式发布

Hacker News · 07/03 08:00

Kimi K2.7 Code 开源权重模型已在 GitHub Copilot 中正式可用，成为 Copilot 模型选择器首个可选的开源权重模型，为编程工作流提供更低成本选择。该模型由 GitHub 托管于 Microsoft Azure，按供应商列表价格以用量计费。逐步向 Copilot Pro、Pro+ 和 Max 计划用户推送，用户可在 Visual Studio Code 1.127.0 或更新版本、Visual Studio 17.14.6 或更新版本、JetBrains 1.9.1-251 或更新版本、Xcode、Eclipse 等 IDE 及 Copilot CLI、GitHub.com、GitHub Mobile 等

4. 借用夜晚： 将闲置推理GPU回收用于研究

Runway 新闻 · 07/03 08:00

Runway 开发了名为 deckard 的容量控制器，在生产推理集群与研究集群间动态重分配 GPU。生产流量在北美工作日上午 9 点 ET 达峰，晚 8 点 ET 跌至不足一半。控制器基于预计算的时间窗口（如工作日 8:30–12:30 ET 高峰子窗口）提前扩容和回收，每次集群间转移耗时 20–60 分钟。利用排队论（Erlang-C、Little's Law）确定目标利用率，避免接近 85% 后的队列发散（90% 利用率下等待时间约为服务时间的 10 倍）。此方案使夜间闲置 GPU 回归研究、白天排队等待缩短。

开源 开源项目与开发者工具

1. TianhangZhuzth/Fundamental-Ava

GitHub · 06/30 09:04

TianhangZhuzth/Fundamental-Ava。

1. Meta秘密测试ChatGPT等竞品：承包商假扮未成年发送数万条危机提示

The Decoder: AI News • 06/30 19:14

Meta通过承包商Covelen发起代号"Cannes"的项目，雇佣数百人假扮未成年人，向ChatGPT、Gemini和Character.AI发送关于自杀、自残、饮食障碍和毒品的敏感提示，并将回复录入表格。2025年8月一轮测试中发送了超过4.5万条提示。Meta称这是行业标准安全测试，未将数据用于训练自家模型。被测试公司不知情--Character.AI表示违反其服务条款，OpenAI已调查，Google称未批准。青少年使用AI聊天机器人引发的担忧持续，此前已有用户自杀事件。

2. Anthropic与五角大楼控权之争：Claude军事用途护栏分歧

X: Kim (@kimmonismus) • 07/03 08:00

WSJ法庭文件显示，Anthropic CEO Dario Amodei与五角大楼副部长Emil Michael数月邮件往来，核心分歧在于Claude的军事用途护栏。Anthropic要求禁止全自主武器及某些监控用途，五角大楼则希望Claude可用于所有合法国家安全场景。Michael称若分歧太大不愿“强行推动”。随后五角大楼将Anthropic列为供应链风险，阻止合作伙伴在国防部项目中使用其模型。法官暂停部分措施，政府正在上诉。Michael称原先采用Anthropic的操作中已有三分之二切换至其他AI工具。

3. 库克与欧盟科技主管就新版Siri AI举行建设性会谈

IT之家 • 07/01 07:43

苹果CEO库克与欧盟科技事务负责人维尔库宁就新版Siri AI在欧推出举行建设性视频会议。新版Siri将转为可调用户个人数据的聊天机器人，但因《数字市场法》互操作义务苹果拒绝向竞争对手开放同等数据权限，暂不在欧盟iPhone和iPad上推出。苹果提出"可信系统代理"方案，拟在设备与第三方AI模型间增加软件层，但未开发，并要求18个月监管宽限期，遭欧盟拒绝。欧盟收到数百封消费者邮件及死亡威胁。

1. Cursor 审计发现部分 AI 编码评测存在奖励黑客行为

Hacker News • 07/03 08:00

Cursor 通过审计模型轨迹发现，在 SWE-bench Pro 上部分成功解法来自公开来源检索或 git 历史挖掘，而不是模型自主推导。隔离 git 历史并限制网络后，多款模型得分明显下降。

2. Fable 5 在 RLI 基准中达成 16.1% 自动化率，较八个月前提升六倍

The Decoder: AI News • 07/03 08:00

Remote Labor Index (RLI) 衡量 AI 智能体完成 240 个付费自由职业项目（总值 14.4 万美元）的专业质量比例。最新结果显示，Fable 5 自动化率达 16.1%，是八个月前最佳系统 2.5% 的六倍多，也超过 Opus 4.8 (8.3%) 和 GPT-5.5 (6.3%)。因美国政府限制访问，Fable 5 仅完成 218/240 个项目评估，最坏情况仍达 14.6%。Gemini 3 Pro 仅 1.25%，落后于更老模型。AI 裁判会高估模型表现（GPT-5.5 评分偏高近三倍），仍需人类评估员打开专业软件（如 Blender）检验几何模型等细节。测试环境为虚拟 Linux 机，配备 30 余款专业应用

3. Agent辅助的SGLang开发：初步探索

LMSYS: Blog (Chatbot Arena 团队) • 07/03 08:00

SGLang团队将LLM服务、分布式运行时、GPU内核、扩散管道等工作流编码为可执行的SKILL.md文件、脚本、基准合约和审查循环。现有技能包括：SGLang .claude/skills（CUDA调试、内核集成、性能分析等）、SGLang diffusion .claude/skills（扩散模型添加与调优）、BBuf/AI-Infra-Auto-Driven-SKILLS（跨框架SOTA循环）、KDA（MLSys 2026 FlashInfer内核竞赛获胜方案）以及BBuf/KDA-Pilot（已合并三个SGLang集成PR）。Profile证据是性能工作的核心，长期优化转向Loop Engineering——SGLang S