下载日志已记录，仅供内部参考

计算机

报告日期：2026 年 09 月 27 日

Opus 5.5 能力与成本同步优化

—— 2026.09.18—2026.09.26 大模型商业化周报

投资要点

- **本周大模型商业化沿三条主线推进：Anthropic 发布 Opus 5.5，强化长任务执行与 3D 场景构建能力，并降低调用成本；AI 进入科学研究与医疗筛查的具体流程；个人 Agent 向眼镜、桌面终端及本地运行延伸。**
- **Opus 5.5 强化代码与 3D 场景构建，长任务能力与成本改善同步推进。**
  Claude Opus 5.5 支持 100 万 Token 上下文，标准 API 输入、输出价格降至每百万 Token 4、20 美元。案例方面，Clio 反馈模型在一项跨 6 个代码库的大型工程任务中连续工作 18 小时， Optiver 反馈模型在编程任务中耗时及交互轮次约减半；Forward Future 展示包含行人和交通的旧金山 3D 场景。模型能够在更长任务中调用工具、修改和验证代码，应用范围由局部代码生成拓展至跨文件开发和交互原型制作。
- **AI for Science 持续推进，应用覆盖生物研究、理论物理计算和医疗筛查。**
  Anthropic 组织约 950 个智能体在 21 小时内筛查 DNA 数据，识别新酶系统线索并交由实验室验证；Claude Science 使用 Fable 5.1 完成特定理论物理计算。阿里达摩院食管癌筛查模型在 3 个国家 80,612 名患者中验证。通用模型承担检索、编程和分析，专业研究平台连接计算工具与验证流程，AI 正进入具体科研环节。
- **个人 Agent 增加硬件入口，本地模型补充执行能力。**
  Meta 在 Connect 宣布 Muse 将接入 AI 眼镜，并公布口袋设备 Muse Charm；阿里推出 Qwen Intelligence 手机方案和 899 元的桌面机器人 QwenNote Eva。Google Antigravity SDK 支持本地模型，官方代码修复演示中 97.2% 的 Token 在本机运行。硬件减少输入与切换操作，本地执行减少云端调用并保留本机数据，两者共同扩大 Agent 的日常使用场景。
- **风险提示**
  技术迭代与产品落地不及预期；科研及客户案例推广不及预期；行业竞争加剧；数据安全与合规风险。

行业评级：看好(维持)

分析师：刘雯蜀
执业证书号：S1230523020002
liuwenshu03@stocke.com.cn

分析师：郑毅
执业证书号：S1230524070002
zhengyi@stocke.com.cn

相关报告

1 《AI 安全需求和格局有望重塑》 2026.09.22

2 《Jev 探索模型新分工，Muse 拓展个人 Agent 生态》 2026.09.22

3 《开放模型与 Harness 重塑成本结构》 2026.09.19

下载日志已记录，仅供内部参考

正文目录

1 本周模型发布、商业化数据总结 4
2 行业数据 5
3 本周重点事件 6
3.1 Opus 5.5 强化长任务执行，代码开发与 3D 场景构建同步拓展 6
3.2 AI for Science 进入研究执行，专业工具与验证流程共同形成交付能力 7
3.3 个人 Agent 向硬件入口延伸，本地模型补充低成本执行能力 9
4 各厂商其他动态 11
4.1 Anthropic 11
4.2 OpenAI 12
4.3 Gemini 13
4.4 字节跳动 14
4.5 阿里巴巴 15
4.6 智谱 16
4.7 MiniMax 16
4.8 Kimi 17
4.9 DeepSeek 17
5 风险提示 17

下载日志已记录，仅供内部参考

图表目录

图 1： AI 同等能力成本下降速度与其他技术的历史比较 5
图 2： 美国成年人几乎每天使用 AI 的比例上升 5
图 3： Opus 5.5 在命令行任务中的能力与成本表现 6
图 4： Claude Code 的任务时长与上下文规模增加（2026 年 3—9 月） 7
图 5： Opus 5.5 参与构建包含行人与交通的旧金山 3D 场景 7
图 6： Claude 从 DNA 序列中识别重复模式 8
图 7： Claude Science 参与的九圈散射振幅计算示意 8
图 8： DAMO EAGLE 的模型开发、外部验证及筛查流程 9
图 9： Meta 扩展 AI 眼镜产品形态 10
图 10： Meta 以 Muse 为核心扩展个人 Agent 的使用入口 10
图 11： Antigravity SDK 接入 LiteRT，支持本地模型执行 10
图 12： QwenNote Eva 连接个人设备与办公任务 11
图 13： Anthropic 前 1% 客户的支出份额由 25% 升至 46% 11
图 14： GPT-6 Sol 与 Luna 的任务表现及单次任务成本 12
图 15： Proaction 车队管理平台界面 13
图 16： Ringg 健康险续保智能体工作流 13
图 17： Gemini 3.8 Flash 与 Flash-Lite TTS 的语音质量评测 13
图 18： Gemini 语音表演示例：通过指令控制角色的紧张情绪 14
图 19： Gemini Live Avatar 数字形象对话演示 14
图 20： Dramagic 将剧本、素材、分镜与预览串联为制作流程 14
图 21： Qwen-Image-2.1 的官方图像生成与编辑评测 15
图 22： Qwen3.8-Max 自我改进实验流程与结果 16
图 23： MiniMax-H3 在 8 张 MI355X 上的生成耗时对比 16

表 1： 本周模型发布情况 4
表 2： 本周商业化数据总结 4

下载日志已记录，仅供内部参考

1 本周模型发布、商业化数据总结

表1： 本周模型发布情况

发布方	模型	类型/形态	核心信息
Anthropic	Claude Opus 5.5	通用 / 编程	100 万 Token 上下文；标准输入 / 输出 4 / 20 美元，缓存读取 0.20 美元；Fast 模式 8 / 40 美元。
OpenAI	GPT-6 Sol / Luna	通用 / 编程	短上下文标准输入 / 输出分别为 2 / 10 美元、0.10 / 0.50 美元；缓存读取分别为 0.20、0.01 美元。
Google	Gemini 3.8 Flash TTS / Flash-Lite TTS	语音生成	文本输入均为 0.50 美元，音频输出分别为 9、6 美元；优惠截至 2026 年末，随后价格翻倍。
字节跳动	Seedance 2.5 Draft	视频草稿	480P 草稿与 1080P 成片分步生成。
阿里巴巴	Qwen-Image-2.1	开放权重图像模型	原生透明图层、最多 10 张参考图。
阿里巴巴	Qwen Audio-3.1 系列	实时语音 / 识别 / 合成	Realtime-Plus 文本输入 / 输出 5 / 40 元、音频 40 / 150 元；识别、语音合成各型号价格见正文。
阿里巴巴	Qwen3.8-Omni-Flash-Realtime	实时全模态	文本、图像、视频输入 1.5 元，音频输入 6 元；文本 / 音频输出 4.5 / 12 元。

资料来源：各公司官方发布、API 文档，浙商证券研究所

除注明按秒、按张外，价格均为每百万 Token；输入与输出分别计费

表2： 本周商业化数据总结

主体	收入及 ARR	毛利、利润及成本	用户及使用量	收入结构、定价及渠道等
Anthropic	据《纽约时报》，年化收入运行速率超过 1,000 亿美元	官方称 Opus 5.5 典型任务总成本较 Opus 5 低约 40%	3—9 月 Claude Code 每次提示运行时长增至 3.3 倍	与 Akamai 签订 7 年 116 亿美元 CPU 云协议，另有扩展选项
OpenAI	—	客户 Ringg 部分实时任务更换模型后成本约降 90%	Ringg 每月处理超过 700 万次接通电话	ChatGPT 广告覆盖增至 60 多个市场
Google	—	本地模型减少云端 API 调用；演示中 97.2% 的 Token 在本机运行	Gemini Live Avatar 支持 97 种语言	Live Avatar 在 Gemini Enterprise 提供；语音生成按量计费
字节跳动	—	视频先出低分辨率草稿，再生成成片	—	BytePlus 推出覆盖剧本、分镜及预览的 Dramagic
阿里巴巴	—	语音识别、合成和全模态接口分别计价	DAMO EAGLE 在 3 国 80,612 名患者中验证	QwenNote Eva 售价 899 元（产品不含屏幕，用户需接入手机作为屏幕）；真武 V900 超节点计划明年一季度上市
智谱	—	—	—	ZCode 移除本地仓库快照上传链路；MaaS 可申请数据内容不留存
MiniMax	—	第三方 H3 推理测试较同硬件基线最高加速 26.7 倍	—	测试采用 8 张 MI355X，不含最终 MP4 编码
Kimi	—	—	Code Desktop 上线 Windows 和 macOS	K3 上线 Bedrock；全球跨区标准输入 / 输出 3 / 15 美元每百万 Token

下载日志已记录，仅供内部参考

表2： 本周商业化数据总结

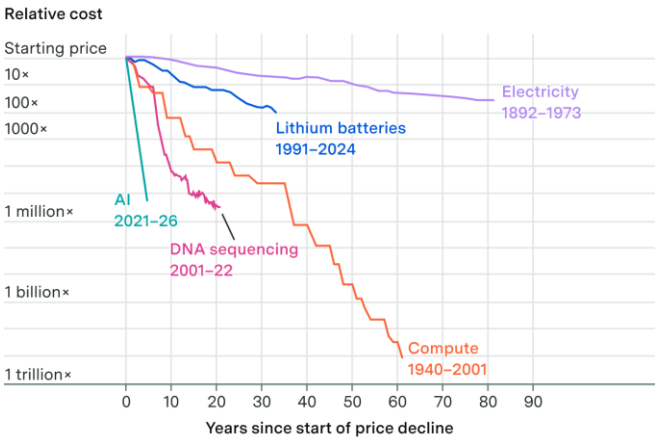
主体	收入及 ARR	毛利、利润及成本	用户及使用量	收入结构、定价及渠道等
DeepSeek	据 The Information，年化营收超过 10 亿美元	调整空闲时段费率的适用日期	DSec 日运行约 300 万个沙盒	据 The Information，推进约 75 亿美元第二轮融资，目标估值约 5,000 亿元
Meta	—	—	Muse 拓展眼镜入口及业务连接器	Ray-Ban Meta Audio 349 美元起、Gen 3 449 美元起；Muse Charm 已公布

资料来源：公司公告及文档、Axios、The New York Times、The Information、Nature Medicine、Nunchux，浙商证券研究所

2 行业数据

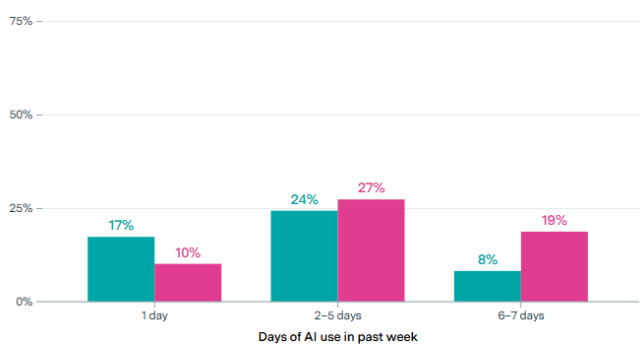
同等能力下的 AI 任务成本持续下降。Epoch AI 对模型在不同任务中达到相同能力水平所需的成本进行比较，估计自 2023 年以来平均每季度下降约 47%。该指标同时考虑模型表现与完成任务的开销，并非单纯比较每 Token 报价。更低的任务成本，使文档处理、代码编写等高频应用能够在同样预算下覆盖更多工作。

图1： AI 同等能力成本下降速度与其他技术的历史比较



资料来源：Epoch AI，浙商证券研究所

图2： 美国成年人几乎每天使用 AI 的比例上升



资料来源：Epoch AI，Ipsos，浙商证券研究所

横轴为受访者过去一周使用 AI 的天数，绿色代表 2026 年 3 月，粉色代表 2026 年 8 月

AI 正从尝试性使用转向更高频的日常使用。Epoch AI 与 Ipsos 的美国成年人调查显示，几乎每天使用 AI 的受访者占比由 2026 年 3 月的 8% 升至 8 月的 19%。成本下降扩大可用任务范围，使用频率上升则反映 AI 工具更深地进入日常工作和生活；两组数据分别体现供给成本与使用习惯的变化。

下载日志已记录，仅供内部参考

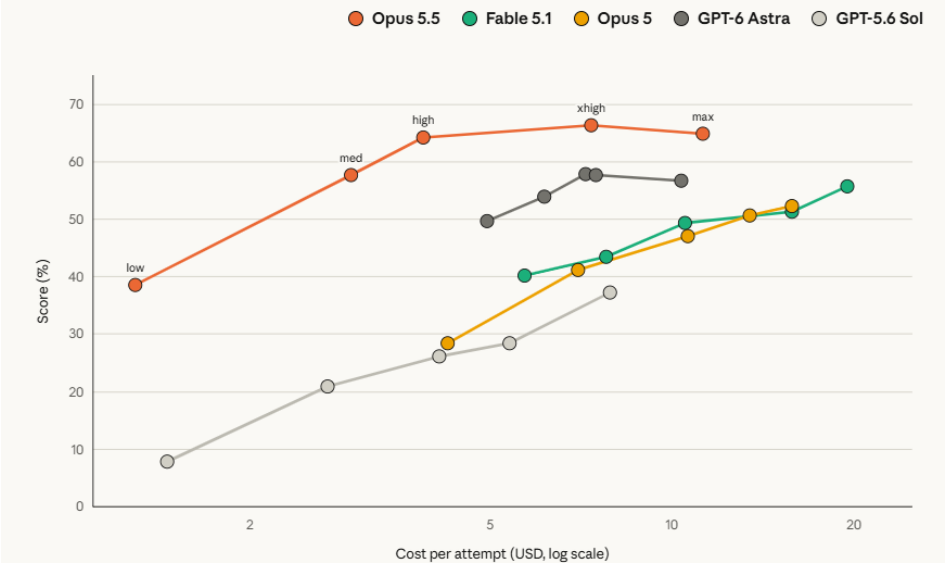
3 本周重点事件

3.1 Opus 5.5 强化长任务执行，代码开发与 3D 场景构建同步拓展

长任务执行能力提升与调用降价同步推进，Opus 5.5 的更新重点从单次回答转向持续完成工作。9 月 22 日，Anthropic 发布 Claude Opus 5.5，默认支持 100 万 Token 上下文和最多 12.8 万 Token 输出。上下文是模型一次处理任务时能够读取的材料范围，更大的容量有助于同时处理代码、文档和执行记录。标准 API 输入、输出价格分别为每百万 Token 4、20 美元，缓存读取为 0.20 美元，较 Opus 5 分别下降 20%、20% 和 60%；Fast 模式输入、输出为 8、40 美元。

成本改善来自单价下降与任务执行效率提升的共同作用。Anthropic 估计，Opus 5.5 在典型任务中的总成本较 Opus 5 低约 40%；其 Terminal-Bench 4.0 测试展示了命令行专业任务的成功率与成本关系。第三方 Artificial Analysis 智能指数中，Opus 5.5 得分为 58，高于 GPT-6 Astra 与 Fable 5.1 的 53。价格、任务完成质量和执行过程共同影响实际使用成本，前沿模型的竞争已延伸至完成一项工作的整体效率。

图3： Opus 5.5 在命令行任务中的能力与成本表现



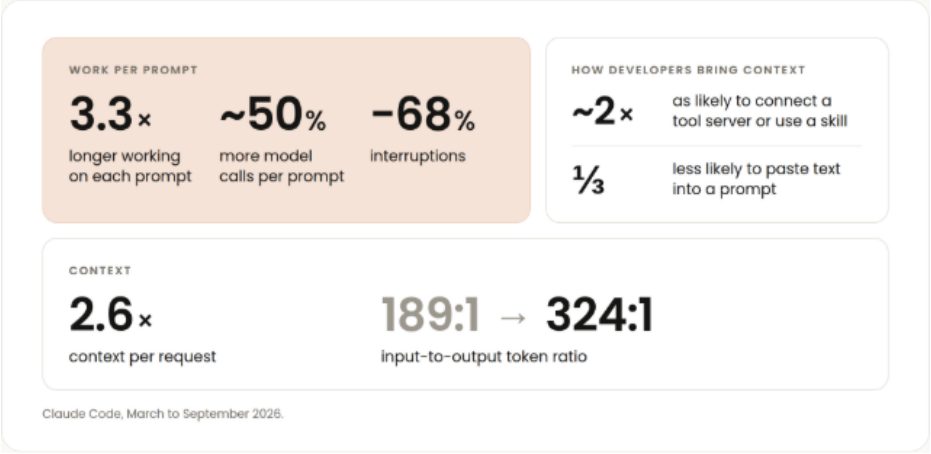
资料来源：Anthropic，浙商证券研究所

跨代码库开发案例体现了模型持续推进复杂工程任务的能力。Anthropic 发布材料引用客户反馈：法律软件公司 Clio 的任务涉及 6 个代码库，Opus 5.5 可连续工作 18 小时；交易公司 Optiver 的编程测试中，完成任务的时间和交互轮次约减半，成本下降 40%—50%。这类任务需要在多个文件与项目之间理解依赖关系、修改代码并检查结果，模型的作用已超过生成一段代码，进一步承担工程任务中的连续执行。

开发者的使用方式也在向更长、更少中断的任务迁移。Anthropic 披露，3 月至 9 月，Claude Code 每次提示对应的运行时长增至 3.3 倍，中断减少 68%，每次请求的上下文增至 2.6 倍。更长的材料输入与更持续的执行过程，使缓存价格、上下文管理和验证能力对使用成本的影响增加，也解释了本次更新为何同时强调长任务和缓存降价。

下载日志已记录，仅供内部参考

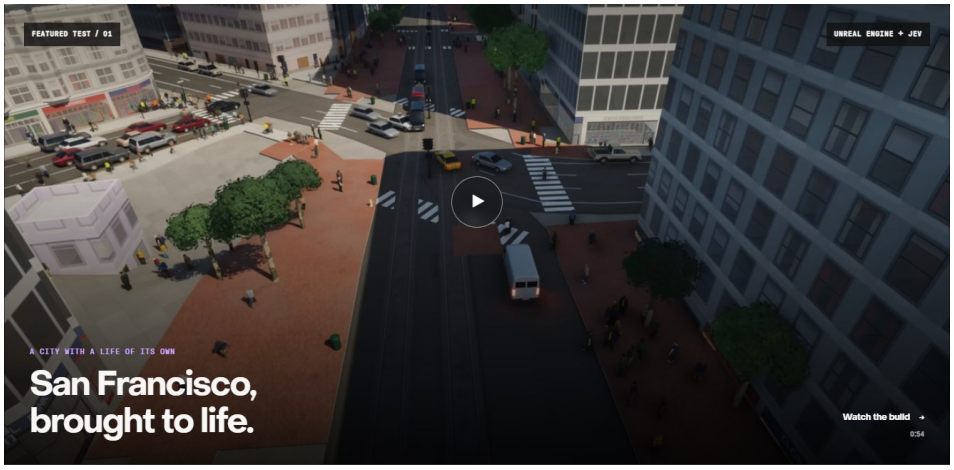
图4： Claude Code 的任务时长与上下文规模增加（2026 年 3—9 月）



资料来源：Anthropic，浙商证券研究所

3D 案例将代码能力转化为可直接体验的交互场景。Forward Future 使用 Opus 5.5，结合 Unreal Engine 三维引擎与 Jev 工具，搭建包含行人、宠物、汽车和交通的旧金山场景。其展示重点是可运行、可交互的城市环境，涉及场景搭建和行为逻辑，而非单张图片生成。此类作品体现了模型调用开发工具、编写代码和反复调整效果的能力，可用于游戏或交互产品的原型验证。

图5： Opus 5.5 参与构建包含行人与交通的旧金山 3D 场景



资料来源：Forward Future，浙商证券研究所

3.2 AI for Science 进入研究执行，专业工具与验证流程共同形成交付能力

本周 AI for Science 进展已落实到数据筛选、科学计算和医学影像分析等具体环节。通用模型可以读取文献、编写程序、调用计算工具并整理候选结果，研究人员负责问题定义、专业核验与实验。由此形成的价值是提高分析和执行效率，研究工具的商业化也从通用对话延伸至面向科学家的计算平台与专业系统。

生命科学研究通过并行智能体扩大候选筛选范围，再由实验室检验研究线索。Anthropic 设立生命科学团队与实验室，并披露约 950 个 Claude 智能体在 21 小时内分析 DNA 数据，识别具有重复序列特征的 ART 酶系统。ART 属于与重复序列相关的逆转

下载日志已记录，仅供内部参考

录酶系统，逆转录酶可以将 RNA 转写为 DNA。该项目将大规模数据检索和异常识别交给模型，再由科学家开展实验，已发现候选系统，但其主要生物功能仍在研究。

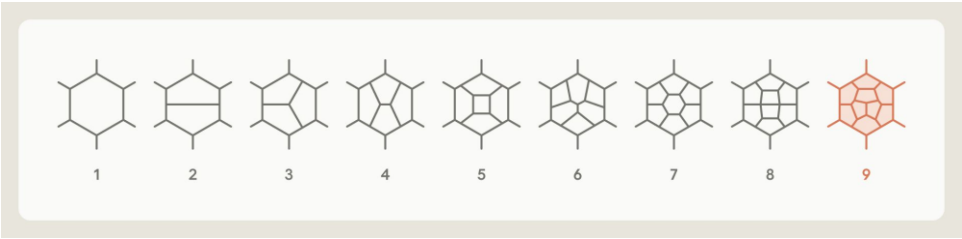
图6： Claude 从 DNA 序列中识别重复模式



资料来源：Anthropic，浙商证券研究所

科学计算案例说明，模型可以在明确约束下持续组织程序与计算资源。Claude Science 是结合模型、工具和执行规则的科研平台。研究者使用 Fable 5.1 完成特定理论中六粒子散射振幅的“九圈”计算：散射振幅用于描述粒子相互作用的概率，“圈数”增加意味着计算更复杂、近似更精细。该结果由专业物理学家核验，使用的是已有计算方法；其意义在于将复杂编程和持续计算组织完成，而非提出新的物理定律。

图7： Claude Science 参与的九圈散射振幅计算示意

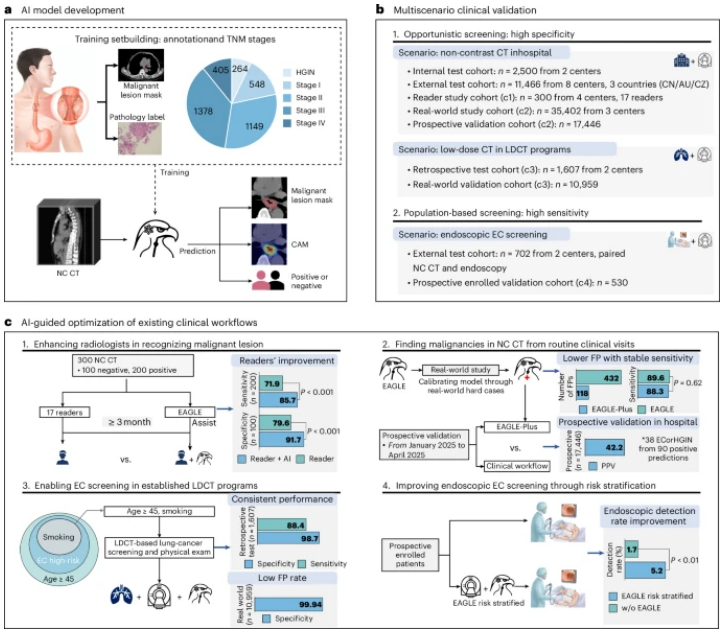


资料来源：Anthropic，浙商证券研究所

医疗筛查将专业模型嵌入已有检查流程，减少额外采集数据的负担。阿里达摩院与医疗机构研发的 DAMO EAGLE 利用平扫 CT 识别食管恶性病变，论文发表于《Nature Medicine》，在 3 个国家的 12 个中心、80,612 名患者中验证。系统先定位食管，再判断和标记可疑病变，可用于从已完成 CT 检查的人群中筛出高风险患者，进一步开展内镜检查。该路径以现有影像和临床验证为基础，将模型结果接入筛查流程。

下载日志已记录，仅供内部参考

图8： DAMO EAGLE 的模型开发、外部验证及筛查流程



资料来源：Nature Medicine，阿里达摩院等研究团队，浙商证券研究所

公共卫生应用则首先改善信息整理效率。WHO 非洲区域团队使用 Claude 汇总刚果（金）各卫生区资料、对比每日变化，将疫情形势报告整理时间由一整天缩短至不足一小时；相关合作还覆盖疫苗资料比较与基因组分析。上述案例表明，AI 既能参与前沿计算，也能承担科研和医疗工作中反复发生的数据整理任务；两类需求分别对应专业研究平台和机构工作流程的产品化。

3.3 个人 Agent 向硬件入口延伸，本地模型补充低成本执行能力

个人 Agent 正从接收指令的聊天界面，走向能够调用应用并持续执行任务的个人工具。Agent 即能够围绕目标使用工具、推进任务的智能体。手机、眼镜和桌面设备使其更容易获取用户当下的语音、画面与工作材料，本地模型则允许部分任务在设备上完成。入口扩展与推理部署是两个不同层面的进展，共同减少从提出需求到完成操作之间的步骤。

Meta 以 Muse 为核心扩展个人 Agent 的使用入口。Muse 于 9 月上旬推出，在独立云端虚拟机中运行，可使用浏览器和已连接应用推进任务，并在购买、发邮件等敏感操作前征求用户许可。本周 Connect 宣布，Muse 将在未来数月接入 AI 眼镜，用户可让它围绕眼前物品采取行动；同时扩展购物、支付及办公连接器。Meta 还公布口袋设备 Muse Charm，用于与 Muse 实时语音交互，更多产品细节将在年内披露。

眼镜产品通过轻量化和续航改善，增加语音助手的日常可用性。Ray-Ban Meta Audio 重约 43 克，单次充电最长使用 12 小时，349 美元起，计划 10 月 13 日发货；Ray-Ban Meta Gen 3 增加自定义操作按钮，最长续航 9 小时，449 美元起。Meta 同时公布约 100 克的 Meta VR Glasses，提供大屏办公与沉浸内容体验，计划于 2027 年春季以 1,299.99 美元上市。Muse 的云端执行与上述硬件入口相连接，不等同于整套 Agent 均在眼镜本地运行。

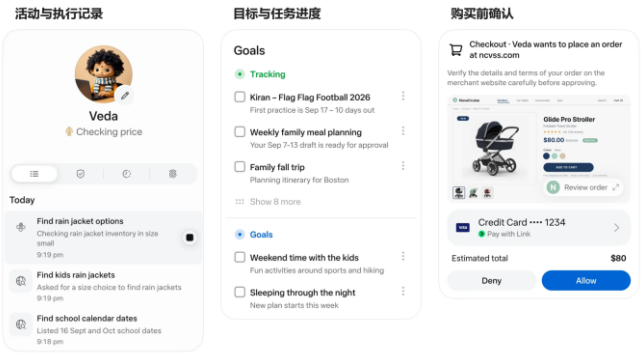
下载日志已记录，仅供内部参考

图9： Meta 扩展 AI 眼镜产品形态



资料来源：Meta，浙商证券研究所

图10： Meta 以 Muse 为核心扩展个人 Agent 的使用入口

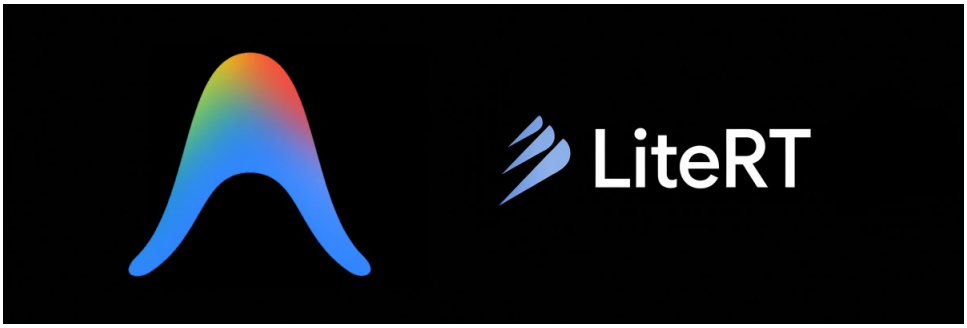


资料来源：Meta，浙商证券研究所

实时数字形象补充了个人 Agent 的交互方式。Meta 发布 Muse Realtime Avatar，可将参考图转化为实时对话形象，使生成的声音、唇形与身体动作同步。其系统将语音生成与视频生成连接起来，用户能够一边交谈、一边看到持续响应的数字角色。本次公布的是交互技术与产品展示，尚无独立公开 API 价格。

Google 将云端规划与本地执行结合，减少代码任务中的云端调用。Antigravity SDK 是用于构建智能体的软件开发工具包，本周加入本地模型支持，首批针对 Gemma 4 26B A4B 和 LiteRT 运行框架优化，可离线执行任务。官方演示中，云端 Gemini 根据文件名和任务说明拆解工作，本地 Gemma 负责复现漏洞、修改代码和测试，97.2% 的 Token 在本机运行，源代码不上传云端。本地模式减少 API 费用与网络依赖，混合模式则保留云端模型的规划能力。

图11： Antigravity SDK 接入 LiteRT，支持本地模型执行



资料来源：Google，浙商证券研究所

阿里同时拓展手机系统和桌面办公入口。Qwen Intelligence 为手机提供系统级智能体执行框架，支持跨应用任务，荣耀 Magic9 系列将首发搭载；QwenNote Eva 桌面机器人连接钉钉、手机和电脑，支持会议记录、日程、消息及文档任务，售价 899 元（产品不含屏幕，用户需接入手机作为屏幕）。前者把任务执行能力接入操作系统，后者通过常驻桌面终端连接已有办公应用，均围绕减少应用切换和重复输入展开。

下载日志已记录，仅供内部参考

图12： QwenNote Eva 连接个人设备与办公任务



资料来源：千问办公，IT 之家，浙商证券研究所

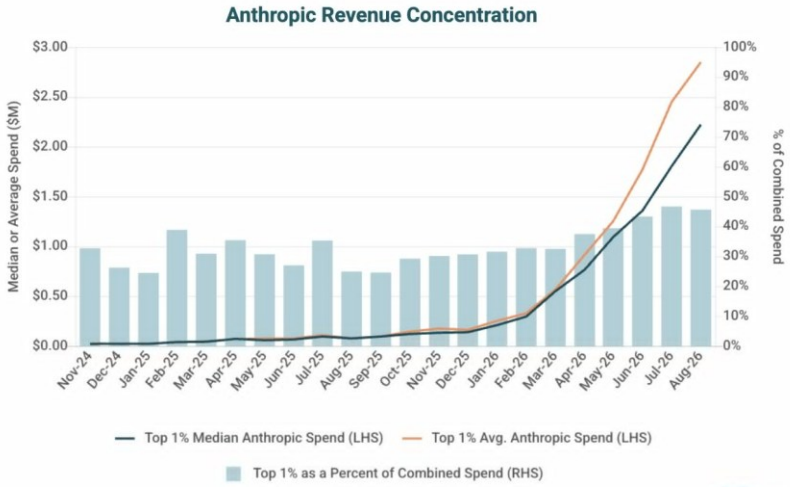
桌面软件也在降低智能体的使用门槛。Kimi Code Desktop 面向 Windows 和 macOS，将任务规划、并行执行、终端、浏览器和代码审阅集中在同一界面；DeepSeek Harness 桌面预览版提供 Windows 与 Mac 安装包。可穿戴设备负责随时发起需求，桌面软件集中承接复杂任务，本地推理补充执行方式，个人 Agent 的竞争因此同时涉及模型、应用连接和终端体验。

4 各厂商其他动态

4.1 Anthropic

- **据 The New York Times 报道，Anthropic 年化收入运行速率超过 1,000 亿美元。**
- **Anthropic 客户支出进一步向头部集中。**YipitData 数据显示，2026 年 8 月，支出最高的 1% 客户贡献了样本客户总支出的 46%，高于 2025 年 8 月的 25%。

图13： Anthropic 前 1% 客户的支出份额由 25% 升至 46%



资料来源：YipitData，浙商证券研究所

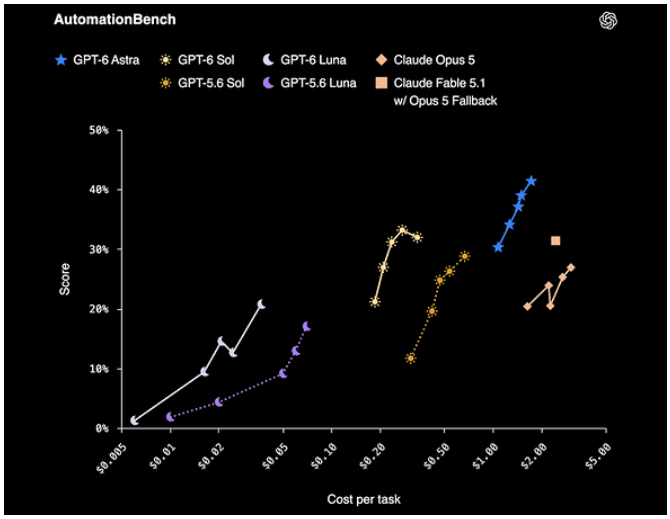
下载日志已记录，仅供内部参考

- **Akamai 宣布与 Anthropic 签订 7 年、116 亿美元算力协议：**提供 CPU 云计算能力以支持智能体业务需求，并包含最高额外 90 亿美元的扩展选项。
- **Anthropic 与 Accenture 建立驻场模型评估合作：**Faculty 团队将参与模型测试、红队评估与安全机制检查，双方预计未来 5 年各投入至少 10 亿美元建设相关能力。
- **Claude Marketplace 上线：**将超过 2,000 个插件与连接器、基于 Claude 的软件及服务合作伙伴集中展示，企业可将部分已承诺的 Anthropic 支出用于采购上架软件。
- **Claude Tag 支持在 Slack 频道中调用个人连接器：**可读取用户本人有权访问的日历、文档和 CRM 信息，回复可先由用户审阅再发到频道；功能先向 Team 开放，Enterprise 后续跟进。
- **Claude 扩展企业会话审计接口。**Microsoft 365 中 Excel、PowerPoint、Word、Outlook 会话的 Compliance API 退出测试，Chrome 会话记录接口进入企业测试。企业可通过接口获取获授权的会话记录，用于内部审计和合规检查。
- **缓存诊断功能正式开放。**开发者可检查请求是否有效使用缓存，并排查未命中的原因，以减少重复处理相同材料的开销。

4.2 OpenAI

- **OpenAI 发布 GPT-6 Sol 和 GPT-6 Luna：**向 ChatGPT Work、Codex 和 API 开放，覆盖不同成本要求的日常工作；标准短上下文 API 输入、输出价格分别为每百万 Token 2、10 美元，以及 0.10、0.50 美元；长上下文另行计价；缓存命中输入价格分别为每百万 Token 0.20、0.01 美元。

图14： GPT-6 Sol 与 Luna 的任务表现及单次任务成本



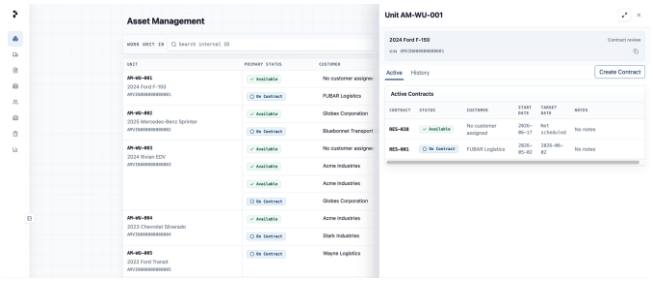
资料来源：OpenAI，浙商证券研究所

- **OpenAI 更新 GPT-6 提示缓存：**符合条件的共享前缀在 30 分钟窗口内可复用，并增加命中率看板、缓存失效诊断、显式断点和预热工具；调整任务中的推理强度也可保留缓存。

下载日志已记录，仅供内部参考

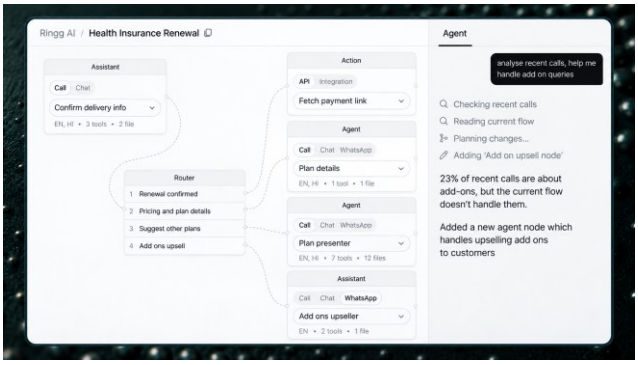
- **ChatGPT 语音支持插件与工作任务：**网页、iOS 和 Android 可在语音对话中调用已连接的应用；在 Work 中可制作文档、演示和表格，结束语音后未完成的任务可继续以文字方式处理。
- **OpenAI 披露 Proaction 使用 Codex 制作客户定制演示：**客户称，每月减少约 40—60 小时工程工作，另为创始人节省约 33 小时；从初次接触进入方案开发阶段的商机比例提高约 50%—60%。
- **OpenAI 披露 Ringg 的语音客服案例：**平台每月处理超过 700 万次接通电话，智能体可独立解决最高约 65% 的常规咨询；部分实时任务从 GPT-4.1 转用 GPT-5.6 后，模型成本约下降 90%。以上为客户案例口径。

图15： Proaction 车队管理平台界面



资料来源：OpenAI、Proaction，浙商证券研究所

图16： Ringg 健康险续保智能体工作流



资料来源：OpenAI、Ringg，浙商证券研究所

- **ChatGPT Finances 接入 Experian 信用报告和 VantageScore 3.0：**向美国 Plus、Pro 用户分批开放，提供月度更新与信用变化提醒。
- **ChatGPT 广告扩展至东南亚 6 国和中国台湾：**全球覆盖市场增至 60 多个；广告面向 Free 和 Go 用户，符合条件的企业可使用自助投放工具。

4.3 Gemini

- **Google 发布 Gemini 3.8 Flash TTS 和 Flash-Lite TTS：**前者支持用自然语言设计角色声音和逐句控制表演，覆盖超过 100 种语言及方言、超过 2,000 种预设声音；经授权可用 30 秒音频复现声音，Flash-Lite 面向高频、低成本语音生成；API 每百万 Token 文本输入价格均为 0.50 美元，音频输出分别为 9、6 美元；该优惠截至 2026 年 12 月 31 日，随后输入与输出价格均翻倍。

图17： Gemini 3.8 Flash 与 Flash-Lite TTS 的语音质量评测

Text-to-Speech Quality Benchmark
Hume AI

	Gemini 3.8 Flash TTS	Gemini 3.8 Flash-Lite TTS	Gemini 3.1 Flash TTS	Eleven Labs v3	Eleven Labs v3 conversational	Cartesia Sonic 3.6	OpenAI gpt-4o-mini-tts	Inworld TTS-2
Overall Reliability x Expressiveness	**0.920**	0.914	0.783	0.706	0.769	0.840	0.740	0.576
Human-like variation Lower = flatter or wilder than a human across turns	4.58	4.51	3.95	**5.00**	4.97	3.40	4.22	3.76
Multispeaker	**4.14**	4.10	3.60	3.85	--	--	--	--
Style Tag Control Single Tag	**4.34**	4.32	4.31	3.89	3.97	3.37	--	4.15

Methodology: deepmind.google/models/evals-methodology/gemini-3-8-tts

资料来源：Google、Hume AI，浙商证券研究所

下载日志已记录，仅供内部参考

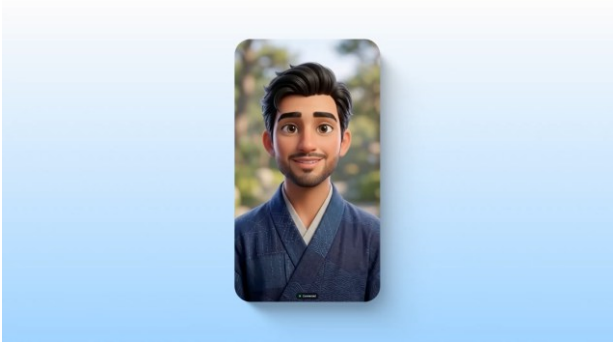
- **Google 展示 Gemini 3.8 Flash TTS 的双人音频剧与情绪切换：**在“The Heist”示例中，两名角色可在对方说话时插入应答；另一示例把低声讲述、突然大喊与喘息后的正常说话放进连续表演中，展示播客和戏剧配音的细粒度控制。

图18： Gemini 语音表演示例：通过指令控制角色的紧张情绪



资料来源：Google，浙商证券研究所

图19： Gemini Live Avatar 数字形象对话演示



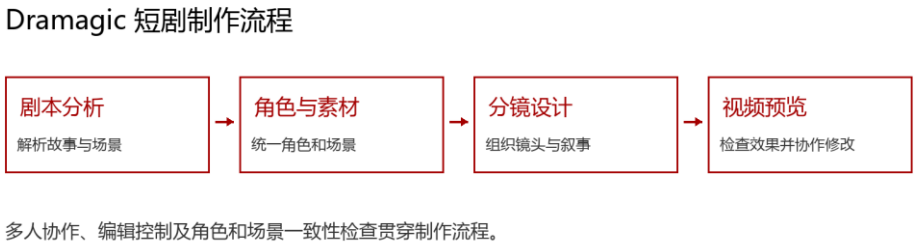
资料来源：Google，浙商证券研究所

- **Gemini 3.8 Live 新增 Live Avatar：**可根据摄像头或共享屏幕内容，以同步的语音和数字形象回应，支持 97 种语言及对话中调用工具；在 Gemini Enterprise 提供，自定义形象需申请相应权限；底层 Gemini 3.8 Live 标准 API 的每百万 Token 音频输入、输出价格为 3、12 美元，文本输入、输出为 0.75、4.50 美元，图像或视频输入为 1 美元。
- **Google 公布 Private AI Compute 的持久记忆方案：**将上下文以加密形式存于云端，密钥由用户设备持有，以支持跨设备连续工作；本周披露技术架构与验证机制。
- **Google Beam 扩展至 6 个国家：**通过 18 家渠道合作伙伴提供企业沉浸式视频协作，并与 Industrious 建立共享使用网络。

4.4 字节跳动

- **Seedance 2.5 新增 Draft 模式：**先用 480P 生成草稿以验证镜头与创意，满意后再生成 1080P 成片，降低反复试错时的生成开销。
- **BytePlus 推出 Dramagic 短剧制作平台：**覆盖剧本分析、角色设计、分镜和视频预览，支持多人协作与一致性检查，面向完整制作流程提供工具。

图20： Dramagic 将剧本、素材、分镜与预览串联为制作流程



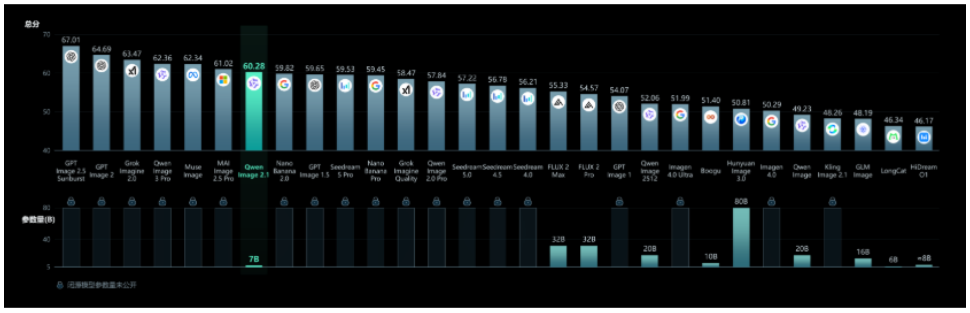
资料来源：BytePlus，浙商证券研究所

下载日志已记录，仅供内部参考

4.5 阿里巴巴

- **Qwen 发布 Qwen-Image-2.1 开放权重模型：**视觉生成部分为 70 亿参数，统一图像生成与编辑，原生支持透明图层，最多可使用 10 张参考图；仓库采用 Qwen Research License。

图21： Qwen-Image-2.1 的官方图像生成与编辑评测

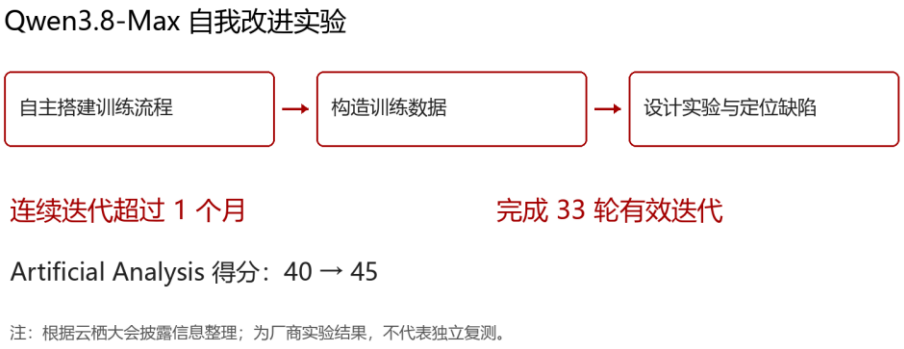


资料来源：Qwen，浙商证券研究所

- **Qwen Audio-3.1 Realtime-Plus 支持实时双工语音对话：**提供工具调用、联网搜索和声音复刻。北京地域每百万 Token 的文本、音频输入价格分别为 5、40 元，文本、音频输出分别为 40、150 元；Realtime 论文中的任务成功率由前代 78.4% 提升至 82.0%，测试采用语音转文字的半双工设置。
- **Qwen Audio-3.1 语音识别系列按输入、输出 Token 分别计费：**北京地域 ASR-Flash 与 Filetrans 的输入、输出价格分别为每百万 Token 0.8、2.7 元；Streaming 与 Message 分别为 6、4.5 元，分别对应流式及消息式识别场景。
- **Qwen Audio-3.1 TTS-Flash 提供语音合成：**北京地域文本输入、音频输出价格分别为每百万 Token 1.5、12 元；国际站新加坡自 9 月 22 日起改为按 Token 计费，输入、输出分别为 0.23、1.87 美元。
- **Qwen Audio-3.1 TTS-Next 支持统一音频生成：**可将对白、音效和环境声组合生成完整音频，播客单次输出最长 4 分钟，其他场景最长 2 分钟；北京地域每百万 Token 输入、输出价格为 6、12 元。
- **Qwen3.8-Omni-Flash-Realtime 更新实时交互能力：**新增多通道音频、视频聚合与远程 MCP 工具，支持多种实时接入方式；北京地域每百万 Token 的文本／图片／视频输入为 1.5 元、音频输入为 6 元，文本输出为 4.5 元、音频输出为 12 元；生成语音时，音频及对应文本分别计费。
- **阿里在云栖大会披露 Qwen3.8-Max 的自我改进实验：**公司称模型在无人参与的流程中持续迭代超过 1 个月、完成 33 轮有效迭代，并用于推理框架和芯片设计优化；相关结果为厂商披露的实验数据。

下载日志已记录，仅供内部参考

图22： Qwen3.8-Max 自我改进实验流程与结果



资料来源：阿里云栖大会、亿邦动力，浙商证券研究所

- **阿里披露 AI 基础设施扩张目标：**吴泳铭在云栖大会表示，将继续投入 AI 芯片、模型和云，目标到 2032 年阿里云运营的全球数据中心规模超过 20 GW。
- **阿里平头哥发布真武 V900：**官方称算力达到 M890 的 3 倍，支持 216 GB 显存；搭载该芯片的新一代磐久超节点服务器计划于 2027 年第一季度上市。

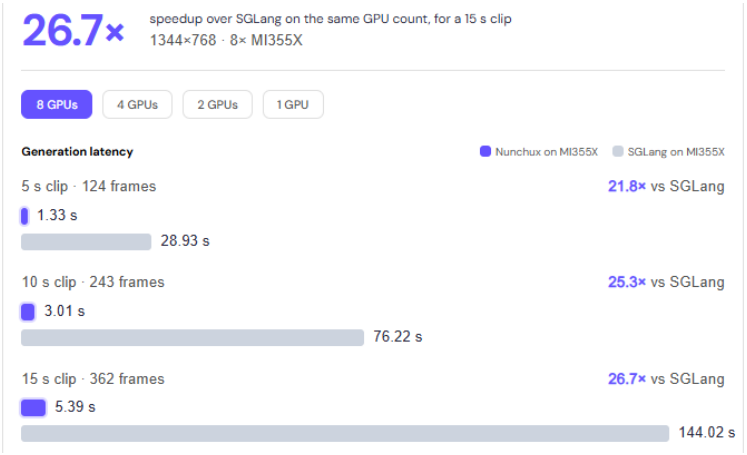
4.6 智谱

- **智谱就 ZCode 仓库数据上传问题道歉并开源代码：**修复版本移除 Repo Wiki 及本地仓库快照上传链路，公布外部安全审查结果；公司称相关数据无留存、未用于模型训练。
- **智谱 MaaS 宣布提供数据内容不留存机制：**用户可申请不静态保存输入和输出，实际开通以平台确认结果为准；Batch API、File API 以及法定留存和违规调查等情形不在一般承诺范围内。

4.7 MiniMax

- **Nunchux 优化 MiniMax-H3 在 AMD MI355X 上的推理：**在 8 张 GPU 上生成约 5 秒视频耗时 1.33 秒，15 秒视频耗时 5.39 秒；相较同硬件 SGLang 基线最高加速 26.7 倍。计时不含最终 MP4 编码，属于开发方测试。

图23： MiniMax-H3 在 8 张 MI355X 上的生成耗时对比



资料来源：Nunchux；开发方测试，不含最终 MP4 编码，浙商证券研究所

下载日志已记录，仅供内部参考

4.8 Kimi

- **Kimi K3 在 Amazon Bedrock 正式开放：**扩展企业云端调用渠道，支持原生视觉、长上下文与显式提示缓存；Bedrock 标准层全球跨区推理的每百万 Token 输入、输出价格为 3、15 美元，缓存读取为 0.30 美元；美国跨区推理对应价格为 3.30、16.50、0.33 美元。
- **Kimi Work 更新文档与移动协作：**支持选中 PPT、Excel、Word 和 Markdown 内容精确修改，新增 Office 文件版本历史、手机预览与上传附件，以及继承当前上下文的侧边对话。

4.9 DeepSeek

- **DeepSeek 调整空闲时段计费覆盖范围：**调休上班的周末和中国法定节假日全天按空闲时段价格计费，具体费率仍以对应模型价格表为准。
- **DeepSeek 公布 DSec 智能体训练基础设施论文：**统一管理不同隔离强度的沙盒环境，披露每秒创建超过 5,000 个沙盒、日运行约 300 万个和峰值约 38 万并发等生产规模，并介绍资源复用与故障隔离。
- **据 The Information 报道，DeepSeek 年化营收超过 10 亿美元：**公司推进约 75 亿美元第二轮融资，目标估值约 5,000 亿元人民币，并计划于 10 月底前完成；年化营收是运行速率口径，融资安排尚需公司确认。

5 风险提示

技术迭代与产品落地不及预期。长任务执行、三维生成及个人智能体对准确性、稳定性和工具调用能力要求较高，模型评测与演示效果可能无法在复杂业务环境中持续复现，影响产品推广和用户采用。

科研及客户案例的推广不及预期。科研实验与企业案例通常基于特定数据、任务和部署条件，扩展至其他场景仍需验证；实验结果的可重复性、专业人员复核要求及系统集成成本，可能影响规模化应用进度。

行业竞争加剧及价格下降影响盈利。模型价格下行与研发、训练及推理投入增加，可能压缩厂商利润空间；客户支出集中度上升，也可能使收入表现更容易受到大客户采购和使用量变化的影响。

数据安全与合规风险。个人智能体、语音及数字形象应用涉及个人信息、企业数据和声音肖像授权，权限管理、数据泄露、知识产权及监管要求变化，可能影响产品上线和商业推广。

下载日志已记录，仅供内部参考

股票投资评级说明

以报告日后的 6 个月内，证券相对于沪深 300 指数的涨跌幅为标准，定义如下：
1.买　入 ： 相对于沪深 300 指数表现 + 20％以上；
2.增　持 ： 相对于沪深 300 指数表现 + 10％ ~ + 20％；
3.中　性 ： 相对于沪深 300 指数表现 - 10％ ~ + 10％之间波动；
4.减　持 ： 相对于沪深 300 指数表现 - 10％以下。

行业的投资评级：

以报告日后的 6 个月内，行业指数相对于沪深 300 指数的涨跌幅为标准，定义如下：
1.看　好 ： 行业指数相对于沪深 300 指数表现 + 10%以上；
2.中　性 ： 行业指数相对于沪深 300 指数表现 - 10% ~ + 10%以上；
3.看　淡 ： 行业指数相对于沪深 300 指数表现 - 10%以下。
我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重。
建议：投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者不应仅仅依靠投资评级来推断结论。

法律声明及风险提示

本报告由浙商证券股份有限公司（已具备中国证监会批复的证券投资咨询业务资格，经营许可证编号为：Z39833000）制作。本报告中的信息均来源于我们认为可靠的已公开资料，但浙商证券股份有限公司及其关联机构（以下统称“本公司”）对这些信息的真实性、准确性及完整性不作任何保证，也不保证所包含的信息和建议不发生任何变更。本公司没有将变更的信息和建议向报告所有接收者进行更新的义务。
本报告仅供本公司的客户作参考之用。本公司不会因接收人收到本报告而视其为本公司的当然客户。
本报告仅反映报告作者的出具日的观点和判断，在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议，投资者应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求。对依据或者使用本报告所造成的一切后果，本公司及/或其关联人员均不承担任何法律责任。
本公司的交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。本公司没有将此意见及建议向报告所有接收者进行更新的义务。本公司的资产管理公司、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。
本报告版权均归本公司所有，未经本公司事先书面授权，任何机构或个人不得以任何形式复制、发布、传播本报告的全部或部分内容。经授权刊载、转发本报告或者摘要的，应当注明本报告发布人和发布日期，并提示使用本报告的风险。未经授权或未按要求刊载、转发本报告的，应当承担相应的法律责任。本公司将保留向其追究法律责任的权利。

浙商证券研究所

上海总部地址：杨高南路 729 号陆家嘴世纪金融广场 1 号楼 25 层
北京地址：北京市东城区朝阳门北大街 8 号富华大厦 E 座 4 层
深圳地址：广东省深圳市福田区广电金融中心 33 层
上海总部邮政编码：200127
上海总部电话：(8621) 80108518
上海总部传真：(8621) 80106010
浙商证券研究所：https://www.stocke.com.cn