

人工智能行业专题（13）

OpenAI 的软硬件生态布局与进展

行业研究 · 海外市场专题

互联网 · 互联网 II

投资评级：优于大市（维持）

证券分析师：张伦可

0755-81982651

zhanglunke@guosen.com.cn

S0980521120004

证券分析师：刘子谭

liuzitan@guosen.com.cn

S0980525060001

- 硬件与算力生态上，OpenAI的选择代表AI叙事主线的认可，一方面用巨额订单提振合作硬件厂商收入预期和股价，另一方面通过股权、协议等与硬件公司深度绑定，形成“风险共担、收益共享”的闭环、强化市场对AI的信心。OpenAI以“星际之门”为核心推进算力基建，联合软银、Oracle计划四年投5000亿美元建10GW AI基础设施。硬件合作上，英伟达投1000亿锁定未来10GW NV芯片供应；AMD以1.6亿股认股权证（行权价0.01美元）换OpenAI未来部署6GW AMD GPU；同时OpenAI与博通达成100亿美元3nm ASIC订单，计划2026年量产。
- 从算力需求上，我们测算OpenAI训练与推理今年均有超十倍的增长。训练端每个大版本模型训练所需计算量均有近10倍提升。推理端ChatGPT的Tokens消耗呈现“用户规模扩张+使用深度提升+单次交互复杂度暴增”的三重驱动下的爆炸式增长，测算ChatGPT 2025年Tokens消耗约为6425万亿（同比去年增长超16倍）。对应OpenAI 2025年收入预计达130亿美元（2024年37亿）、2030年预计达2000亿。
- 软件生态上，OpenAI ChatGPT正在“工具型应用”转向“Agent平台”，并积极抢夺信息获取（谷歌）、电商（亚马逊）与社媒（META）场景。2025年9月ChatGPT推出三大功能：Pulse（用户个人助理、主动推荐个性化简报）、即时结账（联合Etsy/Shopify）、Sora App（AI原生内容的社媒平台）。对比2023年插件时期，**ChatGPT的转型反映了在十亿级用户的高频交互、反馈下对用户数据和决策场景的深度嵌入**，更体现了OpenAI的技术革新、交互体验提升和商业变现的扩展。
- **ChatGPT正重塑用户行为、成为重要入口，其核心价值是决策与执行，不仅可获得生产力溢价、更可赚取用户决策/交易中间环节价值。**
 - ① **从用户规模来看：**ChatGPT用户规模在AI应用中显著领先，2025年7月WAU超7亿（覆盖全球10%成年人）、MAU近10亿。付费版（Plus）留存率80%+，用户日均时长近20分钟（在搜索工具中仅次于Chrome浏览器），2025年7月用户日均发送消息超25亿条（同比增长超5倍）。
 - ② **从用户行为来看：**非工作场景占比73%，深度介入用户生活、Agent属性凸显，用户付费意愿高。ChatGPT的主要作用是“辅助决策”（询问类消息占49%），尤其适配知识密集型工作提升生产力。通过介入“询问（占比49%）”、“执行（占比40%）”环节，带来降低决策成本、加速决策周期的效率价值，后续可进一步主导信息分发、获取和交易决策环节，作为中介平台赚取利润。
 - ③ **从用户分布来看：**跨人群使用差异显著、低收入国家增长快、不同教育/职业人群均能适配，说明产品通用性强、用户增长空间极大。
- 模型性能方面，虽前沿大模型差距缩小，GPT系列始终维持6-12个月技术领先，且推出性价比版本用于商业化。GPT-5（high）与Codex（high）智能化评分位列第一。参数均衡，GPT-5延迟0.78s，开源版gpt-oss-120B输入价0.04\$/M、极具性价比。
- **风险提示：**AI技术进展不及预期风险、行业竞争加剧风险、宏观经济波动风险、AI幻觉问题影响等。

- 【 01 】 OpenAI 硬件和算力生态建设
- 【 02 】 OpenAI 软件生态野心与进步
- 【 03 】 ChatGPT 功能、用户与商业化分析
- 【 04 】 模型性能对比

OpenAI 算力建设：星际之门以及硬件厂商合作计划

- 星际之门(Stargate) 项目：2025 年1月21日，特朗普与OpenAI 正式宣布启动，新成立Stargate公司，软银、Oracle、OpenAI、阿联酋MGX等计划在未来四年内投资5000亿美元并获取Stargate股权，目标建设10GW电力容量的AI基础设施。相当于全球现有AI 数据中心总算力的五分之一。首期1000亿美元已到账或签署托管协议，后三期靠绿色债券滚动和战略股东跟投。
- 25年9月22日，英伟达承诺向OpenAI 的基础设施投资1000亿美元，锁定未来10GW数据中心的芯片供应权。
- 25年10月6日，OpenAI 将部署6GW AMD GPU，AMD向OpenAI 授予最多1.6 亿股普通股的认股权证。行权价0.01美元，若完全行使，OpenAI 将持有AMD约10%的股份。认股权证的解锁与算力部署进度、AMD 股价目标（如达600美元）等挂钩，形成长期战略绑定。

表：星际之门合作方与参与方式

资金来源	金额 (亿美元)	性质	备注
软银愿景基金	1500	股权+可转债	牵头方，分5期认购普通股
阿联酋MGX+ADQ	750	股权	MGX 600 亿、ADQ 150 亿，换取 15% 优先股
甲骨文	400	股权 + 设备抵股	200 亿现金+200 亿云基础设施设备折价入股，获取3000亿美元数据中心承建权，超5.5GW
英伟达	200	设备抵股	以 GB200 芯片组、NVLink 交换机折价入股。不包括25年9月英伟达承诺向OpenAI投资1000 亿美元，这是在 "星际之门" 项目框架之外的独立合作。
CoreWeave	224	合作协议	协议由Coreweave提供高性能 AI 基础设施
美国及各州政府	300	补贴 + 税收减免	能源部贷款担保150 亿、各州税收减免 150 亿
项目债 (绿色债券)	1500	债务	由软银、高盛、摩根大通承销，期限10年
其他战略投资者	250	股权	Arm、富士康、台积电、苹果、AWS 等小额跟投
总计	5000	--	--

数据来源：公司官网、华尔街新闻、维基百科、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

表：OpenAI 后续数据中心建设计划

时间	地点	负责方与容量	建设说明
2025-2026	德克萨斯州阿比林市	Oracle负责，目前0.7GW，后续可达1.4GW	两栋建筑完工，每栋约60,000个GB200 NVL72 服务器
2025-2026	俄亥俄州洛兹敦、德克萨斯州米拉姆县	软银负责，未来18个月内可出售1.5GW	Vera Rubin 芯片
2026下半年	未知	1GW，与AMD合作	AMD MI450 GPU
2026-	阿联酋的Stargate UAE、挪威的 Stargate Norway等	阿联酋2026年上线200MW，后续计划规模5GW	星际之门阿联酋
2026-2027	德克萨斯州沙克尔福德县、新墨西哥州多纳阿纳县、美国中西部	Oracle负责，超4GW	
总计	未来四年内Stargate计划容量达到近10吉瓦（GW），投资超5000亿美元；英伟达投资1000亿美元锁定未来10GW数据中心的芯片供应权。AMD授权最多1.6亿股（AMD的10%）绑定未来6GW芯片供应权。		

数据来源：公司官网、华尔街新闻、彭博新闻、国信证券经济研究所整理

星际之门计划



初期投资400亿美元（包括现金与设施折价），获取3000亿美元数据中心承建权。



软银初期投资1500亿美元并负责项目的财务规划和资金筹集。孙正义持股90%的Arm提供芯片技术支持。软银旗下的可再生能源公司SB Energy提供至少1.5GW数据中心能源建设。



与OpenAI 签订总价值 224 亿美元的协议，提供高性能 AI 基础设施。



截止25年中旬，微软总投资约为130亿美元，持有约OpenAI 28%股权。微软获得知识产权共享、收入分成（20%以内）、API 独家经营权。OpenAI享有Azure优先供给以及30% Azure折扣。



25年9月4日，博通宣布第四个定制AI芯片的主要客户（OpenAI）承诺了价值 100 亿美元的订单。通过博通开发 3nm ASIC，计划 2026 年量产，目标降低训练成本 30%。



NVIDIA

25 年9月22 日，NVIDIA承诺投资OpenAI 1000亿美元，锁定未来10GW数据中心的芯片供应权，通过GPU销售回收资金。既锁定 OpenAI长期采购，分享OpenAI未来增长红利，又形成“资本-硬件-生态”闭环，巩固其在 AI 芯片领域的垄断地位。



25 年10月6日，OpenAI通过承诺部署6GW AMD GPU 算力换取1.6亿AMD股权。前提AMD股价达标且完成算力部署，形成“算力换股权”的新型合作模式。AMD 通过OpenAI旗舰客户背书获取AI入场券与AI算力市场份额，OpenAI降低对NV的单一依赖且可获得约10%AMD股权。

OpenAI 的选择代表AI叙事主线的认可, 合作后硬件公司均大涨



- 巨额订单提振合作硬件厂商收入预期。**OpenAI 部署至少10GW的英伟达算力资源，按1GW对应约500亿美元总支出、英伟达占比60-70%份额计算，合作预计为英伟达带来约4000亿美元的多年期收入确认； 与Oracle签署3000亿美元算力采购合同，覆盖未来五年的云计算服务, 相当于甲骨文2025年营收的3倍。
- OpenAI的技术认证提升硬件合作方生态地位。**OpenAI对硬件的选择具有行业标杆效应，其合作协议相当于为硬件公司颁发了“AI算力入场券”，显著提升市场对其技术竞争力的认可度。同时OpenAI通过股权、认股权证等资本工具与硬件公司深度绑定，形成“风险共担、收益共享”的闭环，进一步强化市场信心。

表：OpenAI 硬件合作方股价涨幅与说明

合作方	宣布合作后涨幅	说明
软银	20%+	宣布星际之门“Stargate”的项目后，在日本市场一度飙升12%，创下自 2024 年8 月以来最大单日涨幅。另外，在 OpenAI 与AMD 达成合作协议后，软银股价上涨近 3%。
甲骨文	90%+	受“星际之门”项目影响，甲骨文当日收盘大涨超7%。9月30日财报后，由于星际之门剩余履约义务（RPO）暴增至4550亿美元，甲骨文股价单日狂飙36%，市值激增 2500 亿美元，创下自1992 年以来的最大单日涨幅。
CoreWeave	100%+	2025年5月，CoreWeave与OpenAI达成一项价值高达 40 亿美元的协议，CoreWeave 股价一度上涨最高达 8%。
ARM	5%+	星际之门计划公布之后，ARM 两天暴涨21%。
电力设备等	——	星际之门计划公布之后，美股核电板块平均涨幅25%，电力设备vst等暴涨。
博通	15%+	2025年9月，博通与OpenAI 合作研发AI 加速卡芯片的消息公布后，博通股价飙升9.4%，公司市值因此增加 1350 亿美元
英伟达	30%+	星际之门计划公布后，英伟达大涨 6.8%。2025 年 9 月 22 日，英伟达宣布将向 OpenAI 投资最多 1000 亿美元，合作消息一经公布，英伟达股价应声上涨近4%，市值突破 4.5 万亿美元。
AMD	24%	2025 年 10 月 6 日，OpenAI 宣布与 AMD 达成为期四年的深度合作协议，美东时间周一盘前，AMD 股价飙升近收涨 24%，报 203.71 美元，创下自 2016 年以来最大单日涨幅。

数据来源：Wind、华尔街新闻、国信证券经济研究所整理

OpenAI 算力需求：测算训练与推理均有年化超十倍的增长

- 测算每个大版本模型训练所需计算量均有近10倍提升。** 尽管模型参数量和训练语料增长变慢，但计算量增长并非简单依赖“参数量×语料量”，如GPT-5 采用的动态稀疏 MoE 架构（1024个专家模块）虽总参数量低于 GPT-4.5（64个专家），但路由机制和专家协同计算更复杂。同时，GPT-5 首次实现大规模多模态训练（整合文本、图像、视频、3D点云），计算复杂度呈指数级提升。另外，GPT-5采用H100 构建超大规模集群，分布式训练的通信开销呈超线性增长。
- ChatGPT的Tokens消耗呈现“用户规模扩张+使用深度提升+单次交互复杂度暴增” 的三重驱动下的爆炸式增长。** 从2023到2025年，测算用户tokens消耗增长超60倍、2025年达到年度6425万亿（对比谷歌25年7月token消耗量980万亿）。尤其今年，伴随着深度推理、多模态功能与Agent功能的上线，多轮对话、复杂任务请求增多，单次交互Tokens消耗实现质变增长。DAU增长反映了大模型从“小众技术”向“全民应用”的跨越，单次tokens消耗量显示大模型正在承载更重、更复杂的用户需求。

表：OpenAI 大模型训练计算量变化

	GPT-4	GPT-4.5	GPT-5
模型参数量	1.8万亿，采用混合专家 MoE 架构，16 个专家模块，每个 1110 亿参数	约5.7万亿，动态稀疏MOE 架构，约为64个专家、每个约 900 亿参数	1.8 万亿
训练语料量	13万亿	120 万亿 tokens，包含合成数据（由推理模型 o1 生成）	130 万亿 tokens
所需计算量	约 2.15×10^{25} FLOPS	约 2.15×10^{26} FLOPS	约 2.15×10^{27} FLOPS
YoY		1000%	1000%
所需卡资源	A100约2.5万张	H100约3-5万张	H100约35万张
训练时长	90-100 天，总训练成本约 6300 万美元	148-247 天	203-225 天，训练成本预估超5亿美元

表：OpenAI ChatGPT Tokens消耗量测算（T/年）

	2023	2024	2025
MAU（年末） 百万	200	600	1000
AI DAU 百万	30	90	336
DAU时长 分钟/日	7.5	10.6	18.5
DAU日使用次数：次	10	9.5	9.34
假设DAU平均消耗tokens：千/次	1.0	1.2	4.7
DAU日消耗tokens：千	10	11	44.7
年度用户tokens消耗折算：万亿/T	103	366	6,425
YOY		256%	1653%

OpenAI 财务情况：收入与估值增长迅速，但面临现金流压力



- OpenAI 25H1实现43亿美元营收，较2024全年高出约16%。近期月营收已超10亿美元，预计可如期达成130 亿美元的全年营收目标。
- 公司目标年底ChatGPT周活跃用户从3月的5亿增至10亿。这将伴随着巨大的算力/现金开销，因此OpenAI积极通过向云服务商、硬件芯片厂商以股权/合作换服务器等方式，降低现金消耗（2025年服务器租赁预计支出160亿美元）。
- 据The Information报道公司预计2030年营收达2000亿美元，预计2029年前累计现金消耗1150亿美元，2030年起开始产生正向现金流。近期软银、Thrive Capital 等投资者以高价参与 OpenAI 交易，5000亿美元的估值相当于谷歌市值的 1/6（谷歌2024年净利润为1000亿美元），反映资本对AI技术壁垒与规模化产品的企业支付的高估值溢价。

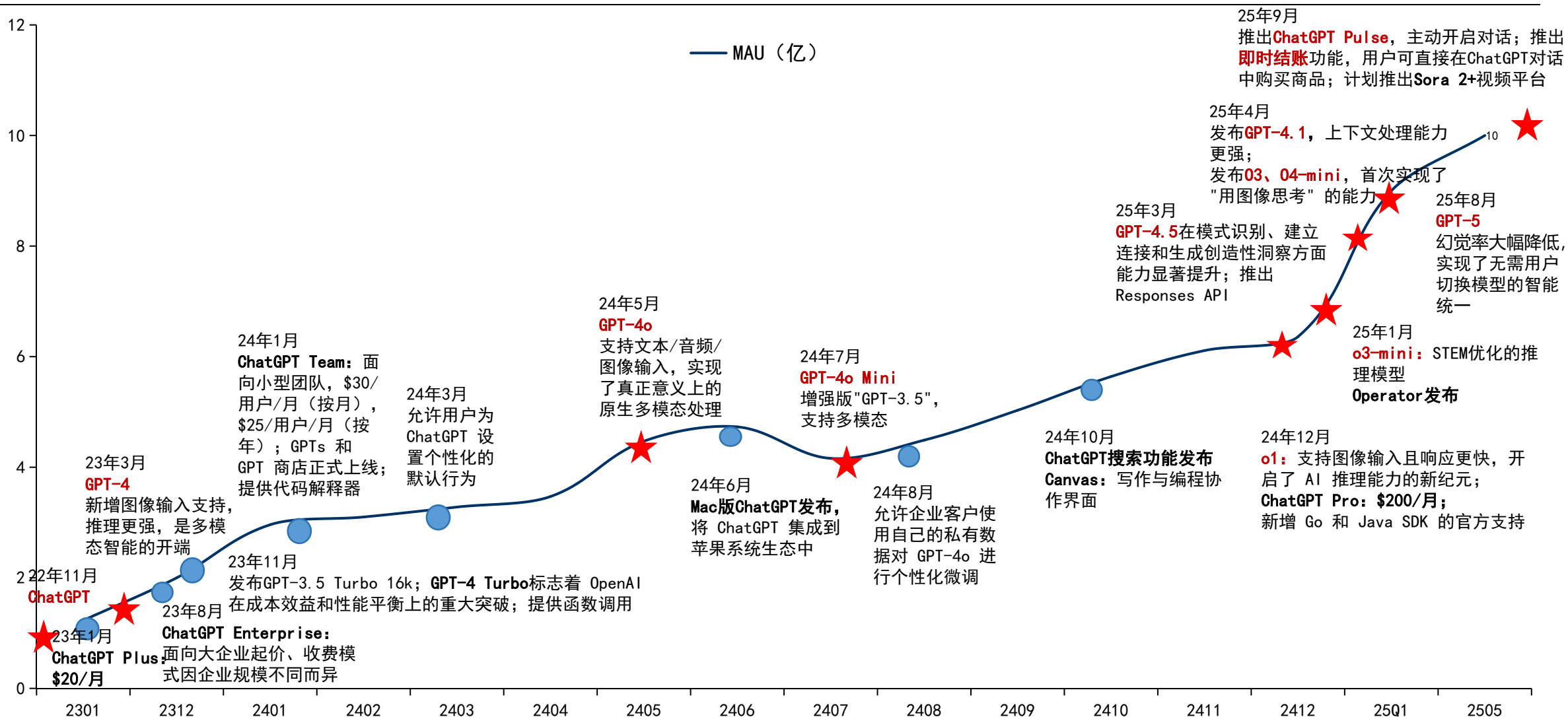
表：OpenAI公司财务数据变化

财务 (亿美元)	2024 全年	2025 上半年	2025 全年预测
收入	约 37	43	130
运营成本	服务器租赁近40亿，数据及训练成本30亿	25亿美元收入成本	服务器租赁预计支出160亿美元
- 研发费用	约67	67	约150
- 销售营销费用	10	20	约 40
- 管理费用	6	-	-
股权激励	约 25	近 25	60
运营利润/亏损	约-35	-78	
净亏损	约50	135，一半亏损来自可转化股权重估，涉及向投资者发布的10亿美金股权	
账面现金及证券（期末）	-	175（二季度末）	
估值	2600	5000	

- 【 01 】 OpenAI 硬件和算力生态建设
- 【 02 】 OpenAI 软件生态野心与进步
- 【 03 】 ChatGPT 功能、用户与商业化分析
- 【 04 】 模型性能对比

OpenAI ChatGPT发展梳理：模型与功能迭代并行，MAU达10亿

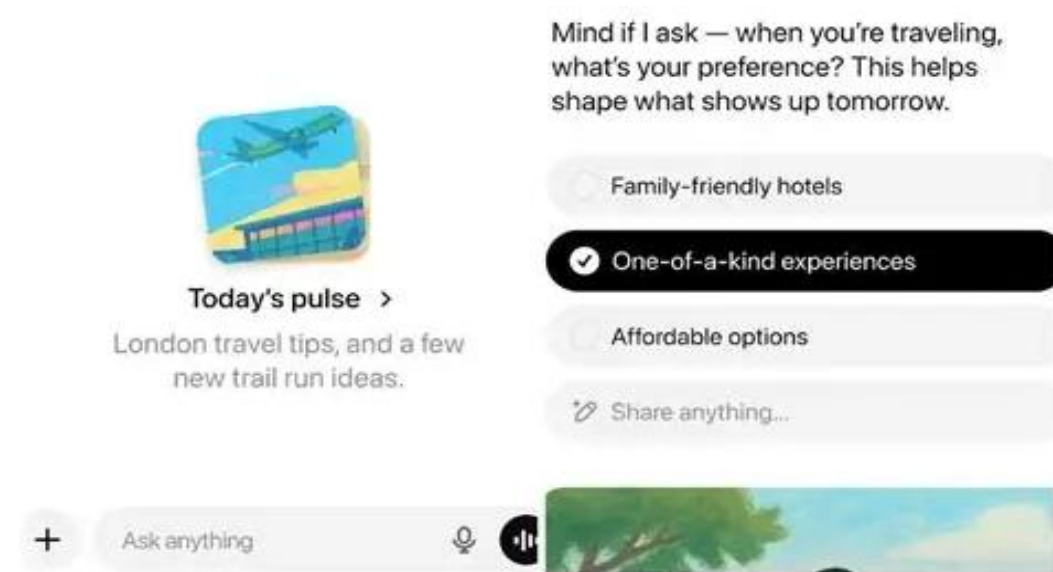
图：ChatGPT月活变化



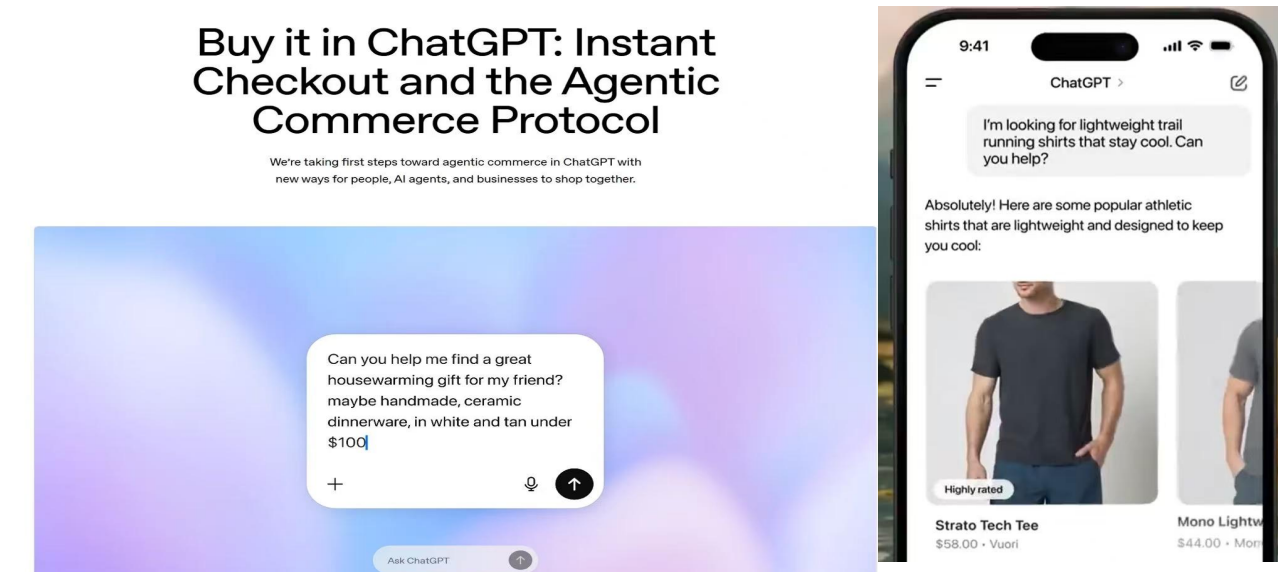
数据来源：ChatGPT官网、Xsignal、国信证券经济研究所整理

- 2025年9月26日，推出ChatGPT Pulse，从被动的问答工具到能主动思考、预测用户需求的私人助理（Agent初级形态）。可主动推送个性化简报（根据用户聊天记录、反馈及连接的应用如日历和邮件等信息），可通过策划 按钮高度定制化信息，支持外部应用集成。目前面向移动端 Pro用户推出。
- 2025年9月29日，OpenAI 推出即时结账功能，用户可直接在ChatGPT对话中购买商品。该功能与Etsy和Shopify合作，面向美国ChatGPT用户开放，超过百万商家将加入平台。基于与Stripe合作的代理商务协议，用户可无缝完成单件商品购买，无需跳转外部网站。订单处理、支付操作及商品履约全流程均由商家通过其现有系统完成，ChatGPT扮演“智能中介”角色。

图：ChatGPT即时结账功能演示



图：ChatGPT即时结账功能演示



- **2025年9月30日，OpenAI推出Sora2与Sora视频社交平台。** 不仅比早期模型更具物理准确性、真实性和可控性，用户可上传短视频/音频，将自己的形象(外貌+声音)插入任何 Sora生成的场景中。

- ① **更符合物理规律：**通过内置物理引擎，Sora 2 能模拟流体力学、刚体碰撞、生物力学等复杂运动。
- ② **首次实现视觉-语音-环境音的原生协同：**生成与视频同步的语音、音效和环境音景。
- ③ **跨镜头叙事的一致性：**通过世界状态跟踪技术，Sora 2能在多镜头切换时保持细节连贯。

Sora以AI原生内容为核心，传统社交媒体以“分享真实生活”为主，Sora则转向“共创虚拟体验”。 构建了创作-交互-社区三位一体的新型社交生态，具备类似抖音的信息客串功能及邀请制推广，聚焦创作而非消费。

图：Sora 2演示



Prompt: figure skater performs a triple axel with a cat on her head

With Sora 2, we are jumping straight to what we think may be the GPT-3.5 moment for video. Sora 2 can do things that are exceptionally difficult—and in some instances outright impossible—for prior video generation models: Olympic gymnastics routines, backflips on a paddleboard that accurately model the dynamics of buoyancy and rigidity, and triple axels while a cat holds on for dear life.

资料来源：公司官网，国信证券经济研究所整理

图：Sora 2演示



Prompt: Vikings Go To War — North Sea Launch (10.0s, Winter cool daylight / early medieval)...

As a general purpose video-audio generation system, it is capable of creating sophisticated background soundscapes, speech, and sound effects with a high degree of realism.

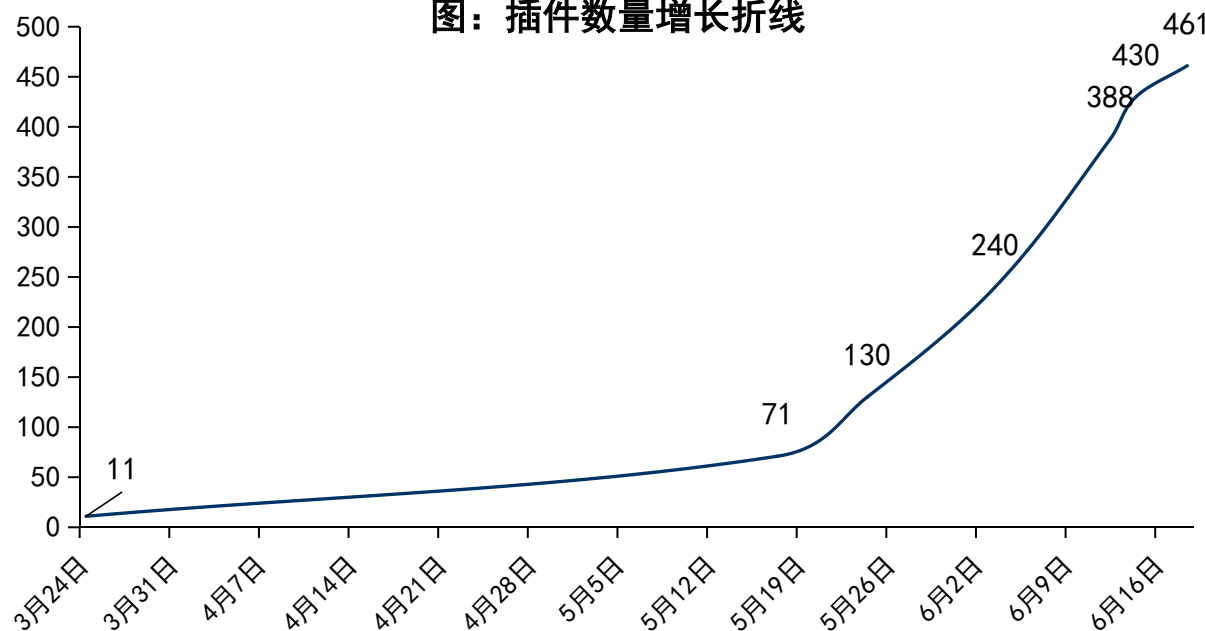
资料来源：公司官网，国信证券经济研究所整理

- **OpenAI转型平台型AI建立在十亿级用户的高频交互反馈下对用户数据和决策场景的深度嵌入。**有望重塑用户与 AI 互动的方式，相比2023年插件时期（ChatGPT plugins）体现了OpenAI的技术革新、交互体验提升和商业变现扩展。
- **技术架构性能的全面升级：从 插件堆砌 到 模型性能与协议的底层革新**
 - ① **大模型多模态能力大幅提升：**插件时期以文本交互为主，即使支持图像输入也需单独调用视觉插件；
 - ② **大模型智能性与决策辅助能力深化：**插件时期仅能基于关键词推荐商品，而当前系统可结合用户历史偏好、实时库存及价格波动生成决策建议，通过“即时结账”功能实现完整的购物交易链路的全流程闭环。
 - ③ **开源协议推动插件生态扩张：**插件时期各插件需独立开发 API 接口，商家集成成本高且兼容性差。与 Stripe 合作开发智能体电商协议（Agentic Commerce Protocol，简称 ACP）的核心技术，为“即时结账”功能提供支持，实现一次集成，全域适配。
- **信息交互与响应逻辑的颠覆：从功能拼凑到场景原生的设计升级**
 - ① **从被动响应到主动服务：**需用户主动触发特定插件（如手动调用“购物插件”搜索商品），且功能触发依赖明确指令，类似“工具调用器”。Pulse采用“预测式服务”模式，通过分析聊天记录、日历等数据主动生成信息卡片。
 - ② **面向移动端的交互更新：**多数插件为桌面端设计，如ChatGPT Box虽支持移动浏览器，但界面缩放混乱，操作需多步跳转。随着APP端用户增长，Pulse卡片式设计适配单手操作，便于用户单手操作和快速浏览。
- **商业变现进化：从工具订阅到价值分成的变现**
 - **从单纯订阅到撮合交易分成的定价变化：**插件时期仅依赖 Plus 订阅费（20 美元/月），用户付费率不足 10%；当前形成“订阅+佣金”双引擎：增值订阅Pulse仅限 Pro用户使用（40 美元/月），推动订阅收入增长45%；交易佣金每笔交易收取 0.5%-1.2% 佣金。

OpenAI 插件生态回顾：包括办公、本地服务等八大类

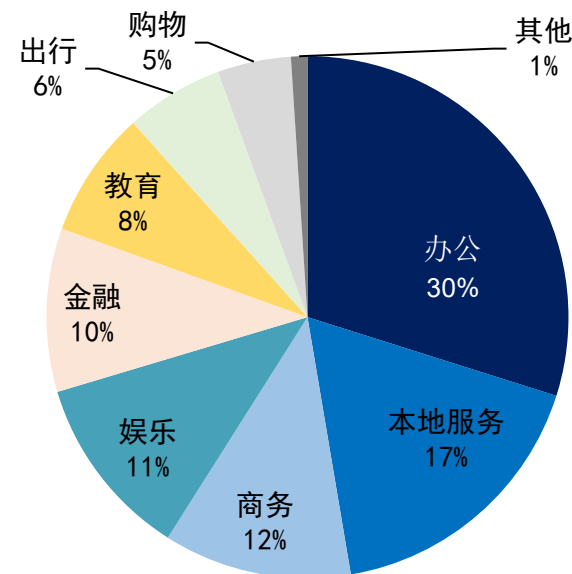
- 外界给予厚望的插件应用：自从OpenAI在23年3月23日发布了ChatGPT plugins以来, 外界对它在OpenAI生态打造、商业层面的潜力充满期待，一度认为这是OpenAI的“App Store 时刻”。5月，OpenAI 宣布向Plus订阅用户全量开放。
- 数量：ChatGPT的插件数量增长迅速，发布初期仅11个，23年6月18日已经从11个增长至461个。
- 功能类型：对于ChatGPT的插件类型进行区分，按照功能划分主要包括办公、本地服务、商务、娱乐、金融、教育、出行、购物等八大类。这些插件皆由第三方插件供应商提供，其中既有大型平台，也有小微创企。在所有的插件类型中以办公类、本地服务、娱乐类、商务类插件为主。

图：插件数量增长折线



资料来源：ChatGPT Plugin Store，国信证券经济研究所整理

图：插件类型分布



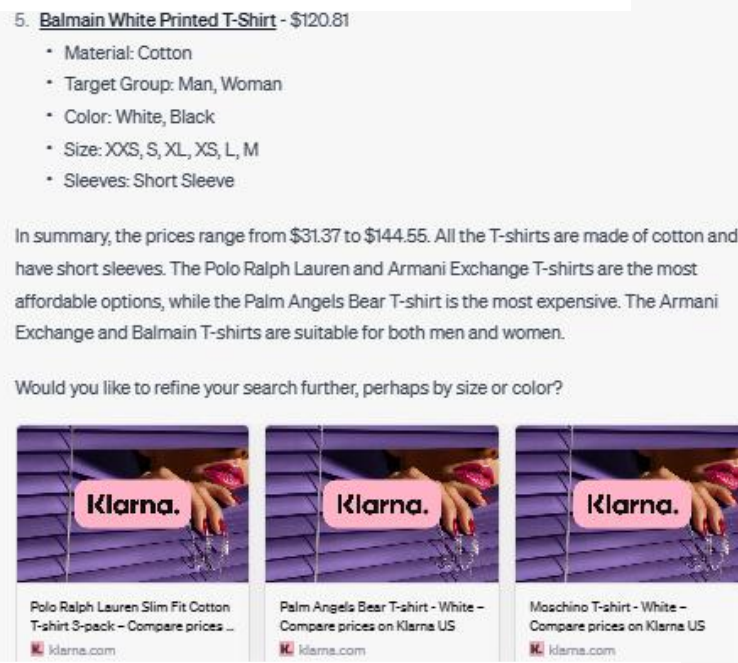
资料来源：ChatGPT Plugin Store，国信证券经济研究所整理

OpenAI 插件的缺点：信息交互、翻译、调用环节都有效率的折损

信息交互问题

1. **信息交互问题**：自然语言的形式限制了信息交流的效率和准确性。
2. **信息展示问题**：部分Plugins使用体验不如直接使用服务官网/APP。特别是在处理信息获取类任务和非标准化需求场景下，没有足够的选择引导用户决策；

图：ChatGPT显示商品链接界面

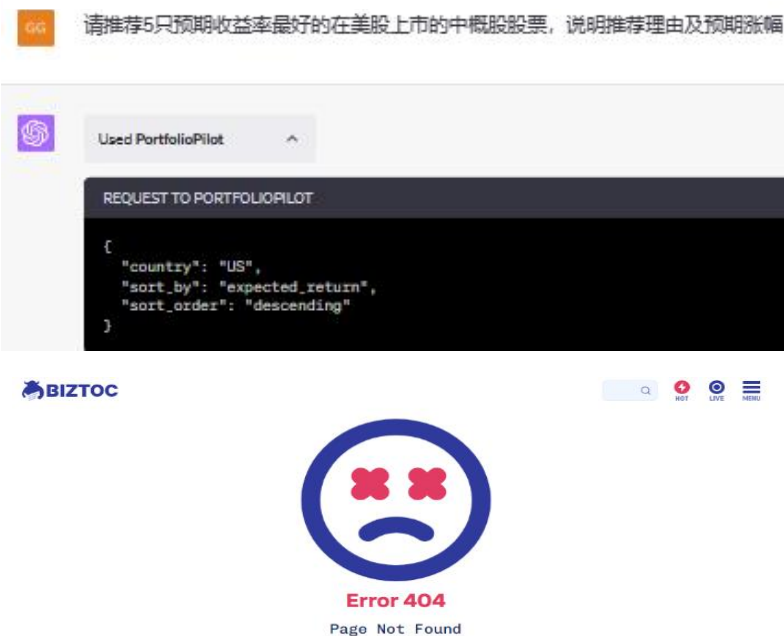


资料来源：ChatGPT，国信证券经济研究所整理

ChatGPT智能性欠缺

1. **ChatGPT前处理问题**：同时也存在ChatGPT在提炼关键、补齐信息、翻译、拆解任务过程中出错。
2. **ChatGPT后处理问题**：插件反馈并不能满足用户的需求，ChatGPT可能产生幻觉，可能生成虚假或错误的链接。
3. 模型本身性能如记忆问题、幻觉问题。

图：ChatGPT信息翻译出错、链接出错



资料来源：ChatGPT，BIZTOC，国信证券经济研究所整理

插件调用问题

1. **GPT对插件不够了解**：ChatGPT调用插件都是依赖开发者提供的文档，可能会出现错误调用影响正常调用。例如，插件不支持中文输入，但 ChatGPT输入中文信息导致错误。
2. **用户对插件不了解**：随着插件越来越多，用户选择合适的插件时遇到困难；

图：ChatGPT输入插件中文信息



资料来源：ChatGPT，国信证券经济研究所整理

- 【 01 】 OpenAI 硬件和算力生态建设
- 【 02 】 OpenAI 软件生态野心与进步
- 【 03 】 ChatGPT 功能、用户与商业化分析
- 【 04 】 模型性能对比

- **2025年7月用户日均发送消息超25亿条（周180亿条），相比去年同期增长超5倍。**根据OpenAI、杜克大学、哈佛大学联合开展并发布的《How People Use ChatGPT》，2025年7月ChatGPT周活跃用户（WAU）超7亿，覆盖全球约10% 成年人口，是史上扩散最快的消费级技术之一，且老用户活跃度持续提升。
- **“实用指导”、“信息查询”、“写作”是三大核心主题（共占近80%对话）。**其中实用指导占比最高、信息查询增速最快，写作是工作场景主要用途（占2025年6月工作消息的40%）。编程仅占4.2%，陪伴/情感相关仅2.3%。
- **用户与地区分析：**年轻用户为主，18-25岁用户贡献46%的消息。消费级用户涵盖Free、Plus、Pro 三档套餐，核心增长来自非北美市场（中低收入国家增速高于高收入国家）。

表：ChatGPT对话主题场景

核心话题	占比（2025 年 6 月）	关键场景举例	工作场景特殊性
实用指导（Practical Guidance）	~29%（占比最高）	定制健身计划、创意构思、辅导教学（10.2% 消息属此类）	占工作消息 24%，第二大工作场景
信息查询（Seeking Information）	24%（增速最快）	查马拉松资格标准、食谱、时事信息	占工作消息 13.5%
写作相关（Writing）	24%（占比下降）	编辑邮件、翻译、总结文档	占工作消息 40%，第一大工作场景（2/3 是“修改文本”而非“新建文本”）
多媒体（Multimedia）	7.3%	生成图像（2025 年4月功能开放后激增）	占工作消息 7.6%
技术帮助（Technical Help）	5.1%	编程（仅占总消息4.2%）、数学计算	计算机行业工作消息占 37%，其他行业仅 8%
自我表达（Self-Expression）	5.3%	闲聊、角色扮演（0.4%）、情感反思（1.9%）	占工作消息仅 1.1%

资料来源：Aaron Chatterji，《How People Use ChatGPT》，Working Paper 34255-D01 10.3386/w34255，国信证券经济研究所整理

ChatGPT用户行为分析：产品商业化价值量与用户增长空间极大



- 非工作用途潜力巨大，深度介入用户生活：非工作消息占比超70%且增长更快（占比从2024年6月的53% 升至2025年6月的73%），AI对“家庭生产”和“个人生活”的影响不亚于工作场景。根据Collis的研究，美国用户2024年因使用AI的“消费者剩余”（非工作场景的决策支持）至少达970亿美元，用户愿意为AI辅助决策支付每月98美元。
- 核心价值是决策支持与执行、Agent属性具象化，不仅可获得生产力溢价、更可赚取用户决策/交易中间环节价值：ChatGPT的主要作用是“辅助决策”（询问类消息占49%），尤其适配知识密集型工作（决策质量直接影响生产力），获取生产力溢价。通过介入“询问（49%）”、“执行（40%）”环节，凸显降低决策成本、加速决策周期的效率价值，后续可进一步主导信息分发、获取和交易决策环节，作为中介平台赚取利润。
- 产品通用性强、用户增长空间极大，使用场景差异化高、与人群高度适配：跨人群差异显著（如女性爱写作、男性倾向 技术帮助和多媒体），且工作场景使用与职业核心任务强匹配（如管理商务侧重写作、计算机侧重技术帮助），体现通用性产品与技术的“场景嵌入性”。低收入国家使用增长快、不同教育/职业人群均能适配，说明AI将极强的普惠场景渗透能力。

表：ChatGPT对话主题场景

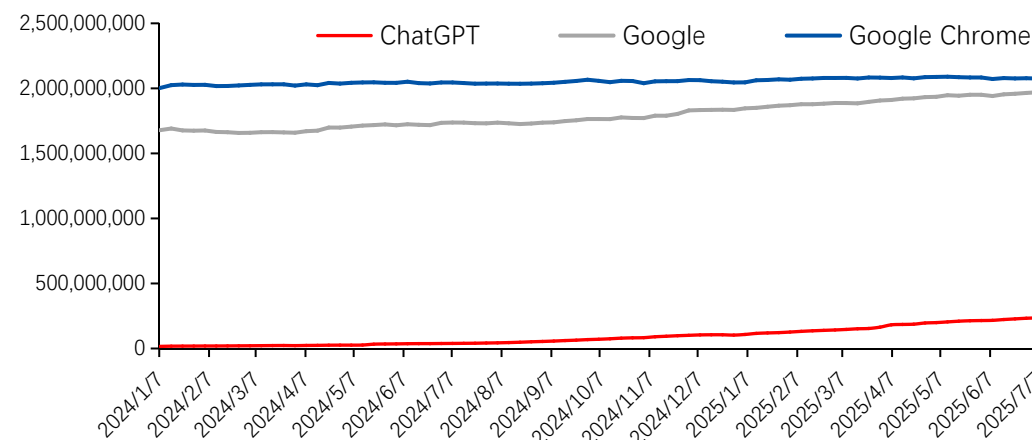
意图类型	整体占比	核心定义	与主题关联	工作场景占比	增长趋势 (2024.7-2025.7)
询问	49%	求信息/建议以辅助决策（如“如何制定季度预算”）	多对应“实用指导 / 信息查询”	35%	从42% 升至 51.6%
执行	40%	求任务输出（如“重写邮件”、“生成代码”）	多对应“写作/多媒体/技术帮助”	56%	从42% 降至 34.6%
表达	11%	无信息/任务需求（如“今天心情很好”）	多对应“自我表达”	9%	从 8% 升至 13.8%

资料来源：Aaron Chatterji, 《How People Use ChatGPT》，Working Paper 34255-D01 10.3386/w34255，国信证券经济研究所整理

热门AI APP DAU变化：ChatGPT用户增长在AI应用中一骑绝尘

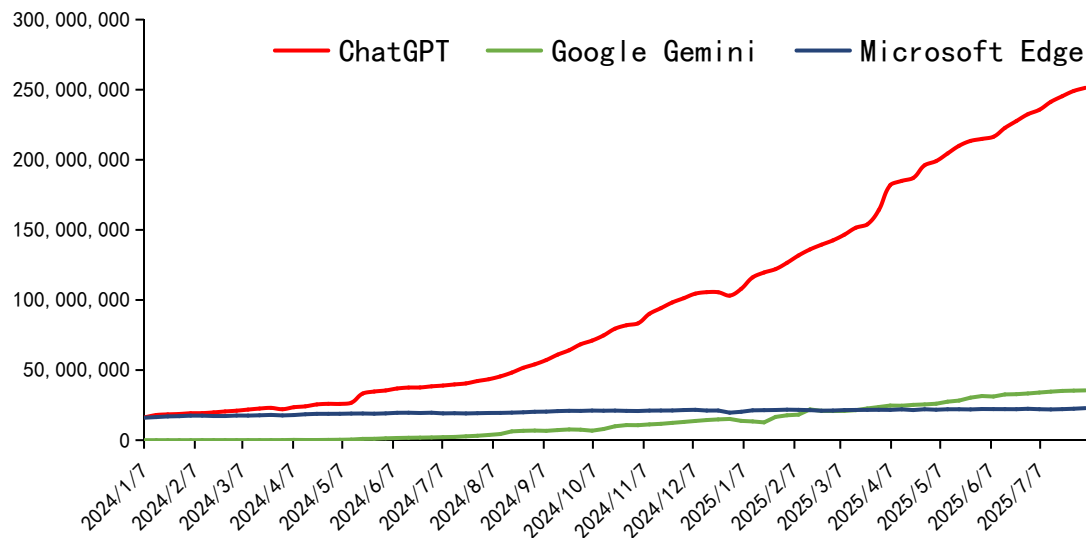
- 传统信息获取工具活跃用户数未见下滑：传统互联网巨头（Google、Microsoft）的核心产品（搜索、浏览器）维持稳定，如Google Chrome长期稳定在20亿上下，用户/热度的“基本盘”极其稳固。
- ChatGPT在通用AI领域DAU显著领先：自发布以来呈现爆发式增长，持续陡峭上升至近2.5亿日活。传统科技厂商的AI新产品（Gemini、Copilot等）处于“稳扎稳打或缓慢渗透”阶段；另外有新兴AI工具（DeepSeek、Grok等）通过特定优势（如技术迭代、场景创新）短期内引爆热度，反映AI赛道创新活跃、用户注意力快速迁移的特点。

图：AI应用DAU变化



资料来源：Similarweb，国信证券经济研究所整理

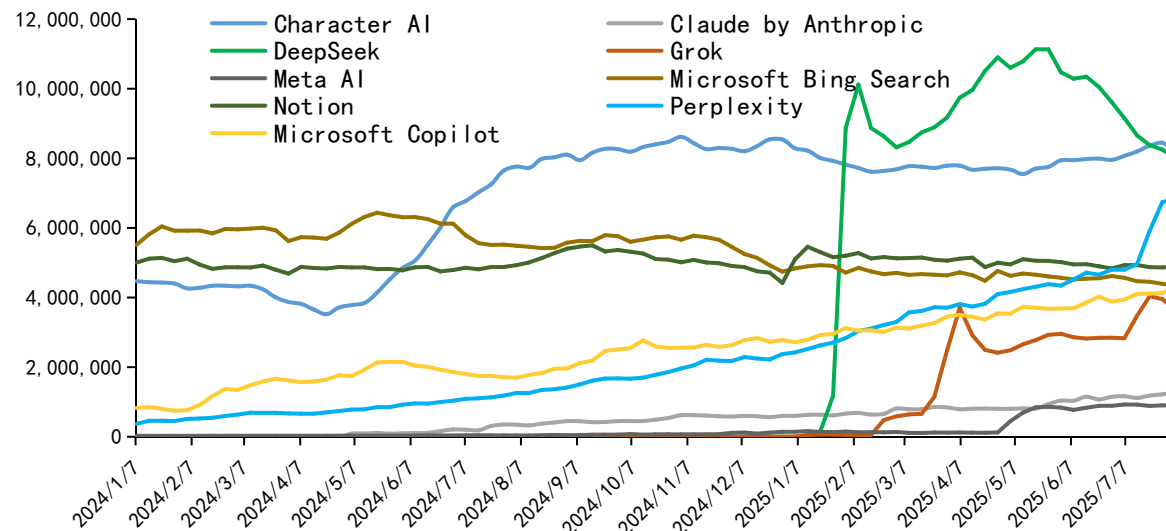
图：AI应用DAU变化



资料来源：Similarweb，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

图：AI应用DAU变化

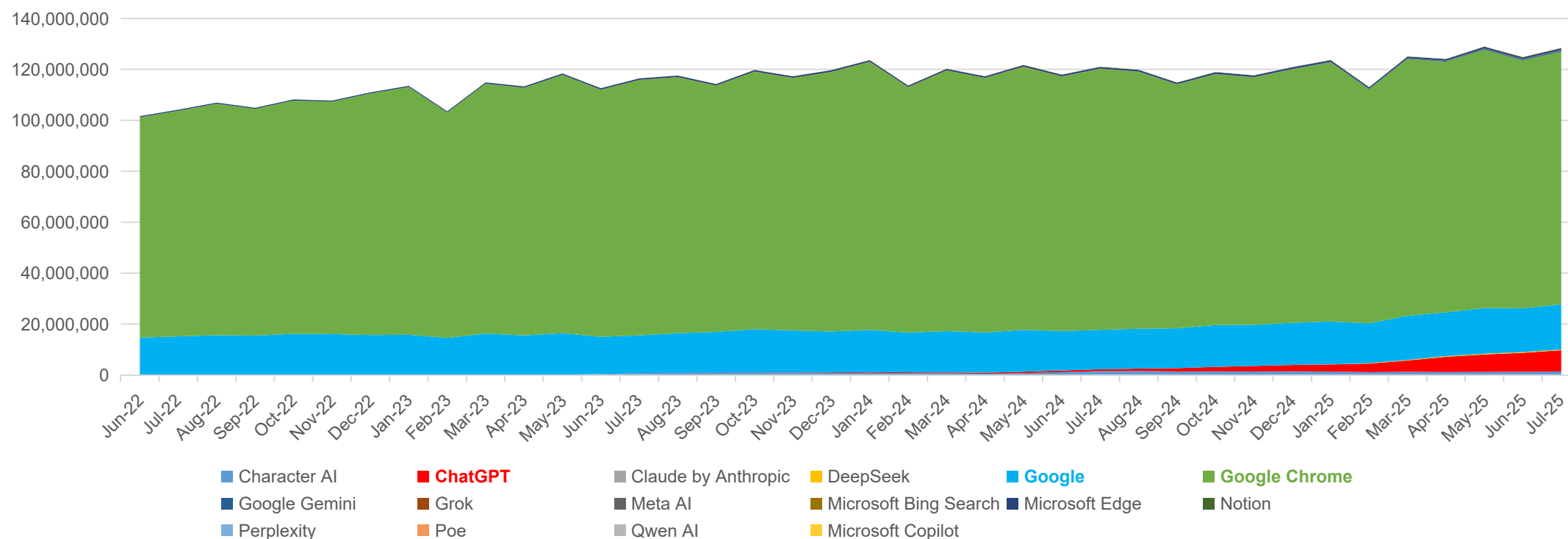


资料来源：Similarweb，国信证券经济研究所整理

用户时长角度，ChatGPT呈现对信息获取场景的快速拓展

- 绿色区（Google Chrome浏览器）和蓝色区（Google搜索）用户整体时长目前并未显著下滑。占据几乎90%以上的面积，作为互联网基础设施级应用，仍是用户信息获取的核心，伴随AI改造不断提升体验，垄断用户的APP时长、粘性极强。
- ChatGPT（红色区）用户使用时长快速增长，反映AI对用户行为的重塑、成为用户的“新增重要入口”。
- 信息获取环节马太效应显著，竞争高度集中。无论是传统互联网（Google系）还是新兴AI领域（ChatGPT），头部APP对用户时间的垄断性都较强，其他AI应用（如META AI、Claude等）在用户时长中占比极低，被ChatGPT挤压。

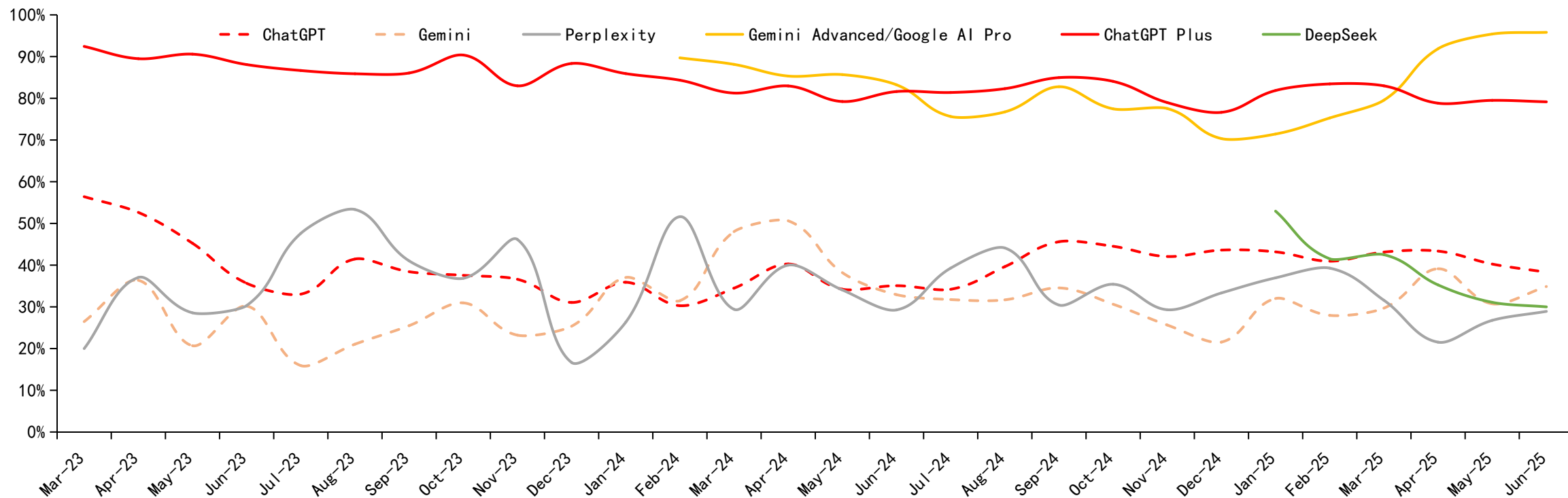
图：各AI应用APP用户月使用时长变化（百万秒/月）



热门AI APP用户留存：头部模型厂商付费版本显著更优

- 付费版AI应用（ChatGPT Plus、Gemini Advanced）的用户留存显著高于免费版（ChatGPT、Gemini）。用户使用AI应用逻辑已从“初期技术新鲜感”转向“服务深度、生产效率、场景专属度”，付费版本“更优体验、专属功能”下用户留存可达90%、忠诚度极强。
- 新兴AI应用距离头部模型厂商产品用户留存差距较大，且波动大、稳定性不足。DeepSeek上线后，2025年初留存达在50%以上，但随后降至30%，表明长期黏性不足。Perplexity的留存在20% - 40%区间，留存率多次出现峰值（如2023年9月、2024年3月），但快速回落，反映其AI搜索类服务的用户留存受功能迭代节奏、用户新鲜感影响，长期粘性不足。

图：AI应用APP用户留存



资料来源：Similarweb，国信证券经济研究所整理

热门AI APP商业化：ChatGPT呈现用户规模与变现效率的综合领先



- ChatGPT是“C 端订阅+B端 API+广告”多渠道协同，不仅C端用户规模大、付费意愿强，又凭借模型技术优势赚取B端企业服务的钱。相比其他厂商或依赖B端（Anthropic）、或依赖生态（谷歌）、或处于早期（Perplexity），商业化的“独立性与多元化”均不如OpenAI/ChatGPT。

表：海外重磅AI应用产品数据汇总

产品	OpenAI/ChatGPT	Anthropic/Claude	谷歌/Gemini	Perplexity
年化收入	2504: 50亿美元+ (+225%) (ChatGPT订阅: 70%; API业务: 30%) 2507: ARR达120亿美元	2024: 10亿美元 (+900%) (第三方API: 60%-70%、直接API: 10-20%、Claude订阅: 15%) 2508: ARR 超50亿美元	2024: 约20亿美元 (估算)	2024: ARR 约5000万美元 (+900%) 预计26年底超6.5亿美元 2506: ARR 1.5亿美元
DAU	2024: 3860W 2502: 1.23亿 2508: 约1.8亿	2023: 约500W 2024: 800W 2505: 1847W	2024: 超1500W 2503: 3500W	
MAU	2404: 1.8亿 2501: 3.5亿 2503: 5亿 (预计年底前超10亿)	2023年底: 约5000W 2024年底: 1.89亿 2506: Web 1550W	2024: 超8000万 2503: 3.5亿 2507: 财报会MAU 4.5亿	2401: 1000W 2507: 5041W
变现模式	订阅制 (ChatGPT Plus\$20/月)、企业API (按token计费)、广告分成 (嵌入Bing搜索)。 订阅用户超1100万 (Plus+企业)，API收入占比提升至30%	企业 API (占比70%)、C 端订阅 ClaudePro \$20/月)、第三方渠道 (亚马逊Marketplace)； API调用量增长300%，企业客户超2000家 (如Notion、Quora)	云服务授权 (API)、广告整合 (Search Generative Experience) 开发者API订阅、企业定制模型 (如Salesforce合作)	免费版每天仅提供 5 次深度搜索，“个人Pro 订阅每月20美元、企业Pro40 美元/月 + API调用”收费模式，计划通过广告变现 (50美元 /CPM)
底层模型	GPT-3.5/GPT-4/GPT-5 (内测)	Claude3.5 Sonnet/Opus	Gemini 1.0/Ultra/Pro/Nano	自研的Sonar模型，还有Claude 3.5 Sonnet、GPT-4o等多个模型
主要合作伙伴	微软Azure	亚马逊AWS	谷歌GCP	亚马逊AWS
估值	25年4月融资400亿美元，估值3000亿美元。 9月为员工发售约103 亿美元股票，此次股票发售估值将达 5000 亿美元。	25年3月完成35亿美元 E轮融资，投后估值达 615亿美元。25年9与130亿美元 F 轮融资，投后估值达到 1830 亿美元。		25年5月估值140亿美元，7月新筹资1亿美元，估值增至180亿美元

数据来源：公司官网、Xsignal、Similarweb、Sacra、The Information、TechCrunch、DemandSage、国信证券经济研究所整理

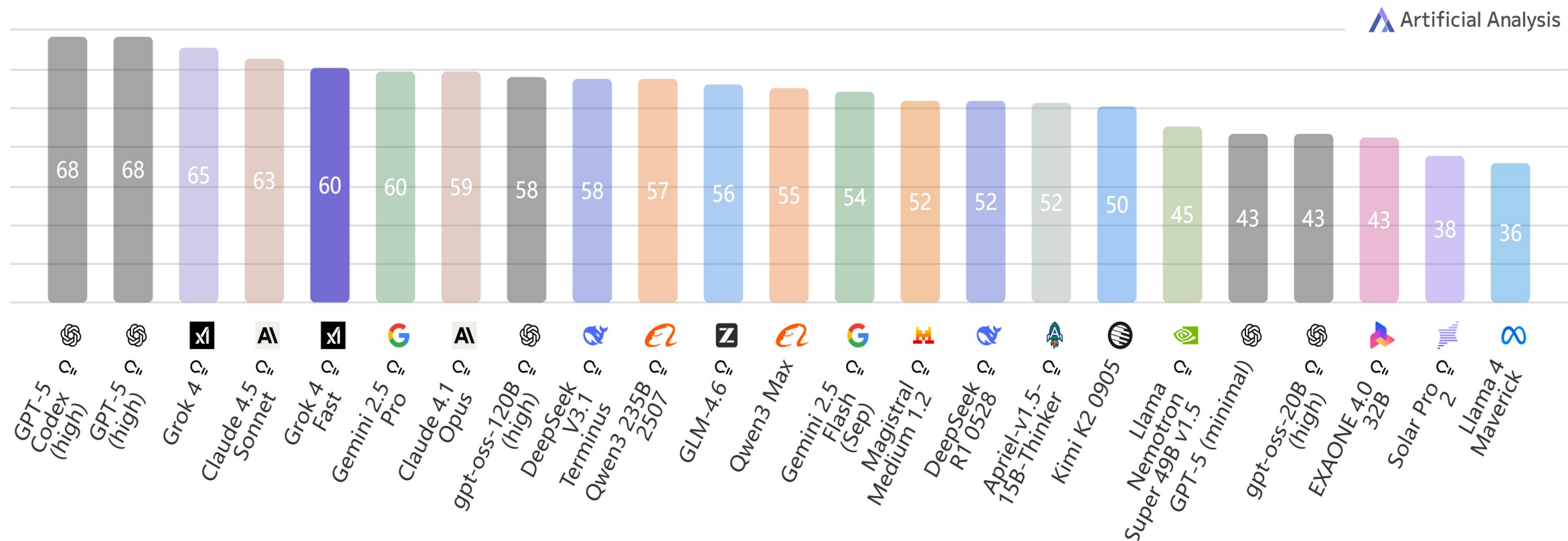
请务必阅读正文之后的免责声明及其项下所有内容

- 【 01 】 OpenAI 硬件和算力生态建设
- 【 02 】 OpenAI 软件生态野心与进步
- 【 03 】 ChatGPT 功能、用户与商业化分析
- 【 04 】 模型性能对比

AI模型综合能力评分：GPT-5凭借最高分领先

- AI大模型的智能评分GPT-5凭借最高分领先。但ChatGPT系列综合能力更是显著领先，包括技术指标的多维度领跑（上下文窗口、最大输出、相应延迟，参考后文表格）、成本与商业性的平衡、快速迭代的技术活力，构建了持续的行业领先优势。

图：AI模型综合能力评分比较



资料来源：artificialanalysis，国信证券经济研究所整理 *人工智能分析智能指数：涵盖智能多个维度的组合指标-比较模型智能程度的最简单方法。3.0 版于 2025 年 9 月发布，包括：MMLU-Pro、GPQA Diamond、Humanity's Last Exam、LiveCodeBench、SciCode、AIME 2025、IFBench、AA-LCR、Terminal-Bench Hard、 τ^2 -Bench Telecom

ChatGPT系列各维度参数领先，且推出性价比版本用于商业化



表：全球智能领先的大模型参数

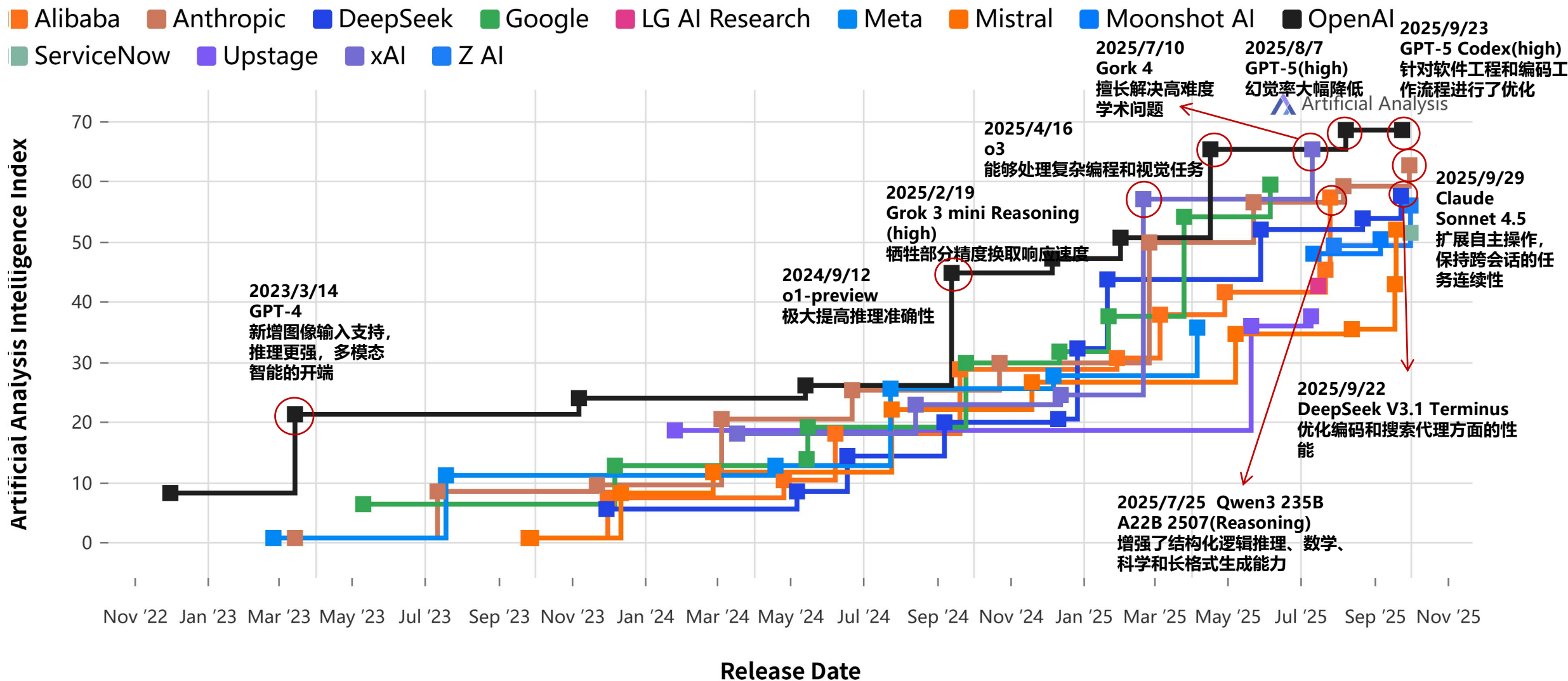
排名	模型商	模型名称	发布时间	综合评分	上下文 (Tokens)	最大输出	输入价格 (\$/M)	输出价格 (\$/M)	延迟 (s)	模型可用平台
1	OpenAI	GPT-5 Codex(high)	2025/9/23	68	400k	128K	1.25	10	7.99	OpenAI
2	OpenAI	GPT-5(high)	2025/8/7	68	128K	16.4K	1.25	10	0.78	OpenAI、 Azure
3	xAI	Gork 4	2025/7/9	65	256K	256K	3	15	15.83	xAI、 Azure
4	Anthropic	Claude Sonnet 4.5	2025/9/29	63	1M	64K	3	15	1.63	Google、 Amazon、 Anthropic
5	xAI	Gork 4 Fast	2025/9/19	60	2M	30K	0.2	0.5	3.07	xAI
6	Google	Gemini 2.5 Pro	2025/7/17	60	1.05M	65.5K	1.25	10	2.71	Google
7	Anthropic	Claude 4.1 Opus	2025/8/5	59	200K	32K	15	75	1.90	Google、 Amazon、 Anthropic
8	OpenAI	gpt-oss-120B(high)	2025/8/5	58	131.1K	131.1K	0.04	0.4	2.09	开源，多数主流云平台 (如 AWS、 Azure、 GCP)
9	DeepSeek	DeepSeek V3.1 Terminus	2025/9/22	58	163.8K	163.8K	0.23	0.90	0.67	开源，多数主流云平台 (如 AWS、 Azure、 GCP)
10	阿里	Qwen3 235B A22B 2507	2025/7/25	57	262.1K	262.1K	0.11	0.60	2.57	开源，阿里云、 Hugging Face等可下载
11	智谱	GLM-4.6	2025/9/30	56	202.8K	202.8K	0.50	1.75	1.15	开源，多数主流云平台 (如 阿里云)
12	阿里	Qwen3 Max	2025/9/23	55	256K	32.8K	1.20	6	1.58	阿里云
13	Google	Gemini 2.5 Flash(Sep)	2025/9/25	54	1.05M	65.5K	0.30	2.50	0.45	Google
14	Mistral	Magistral Medium 1.2	2025/9/18	52	128K		2	5	0.36	Mistral
15	DeepSeek	DeepSeek R1 0528	2025/5/28	52	164K	164K	0.40	1.75	0.99	开源，多数主流云平台 (如 AWS、 Azure、 GCP)

资料来源：各公司官网，artificialanalysis，openrouter，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

虽前沿大模型差距缩小，OpenAI 仍长期维持智能性水平领先

图：AI 模型综合能力评分变化



资料来源: artificialanalysis, 国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

ChatGPT系列：兼具高性能与高性价比版本

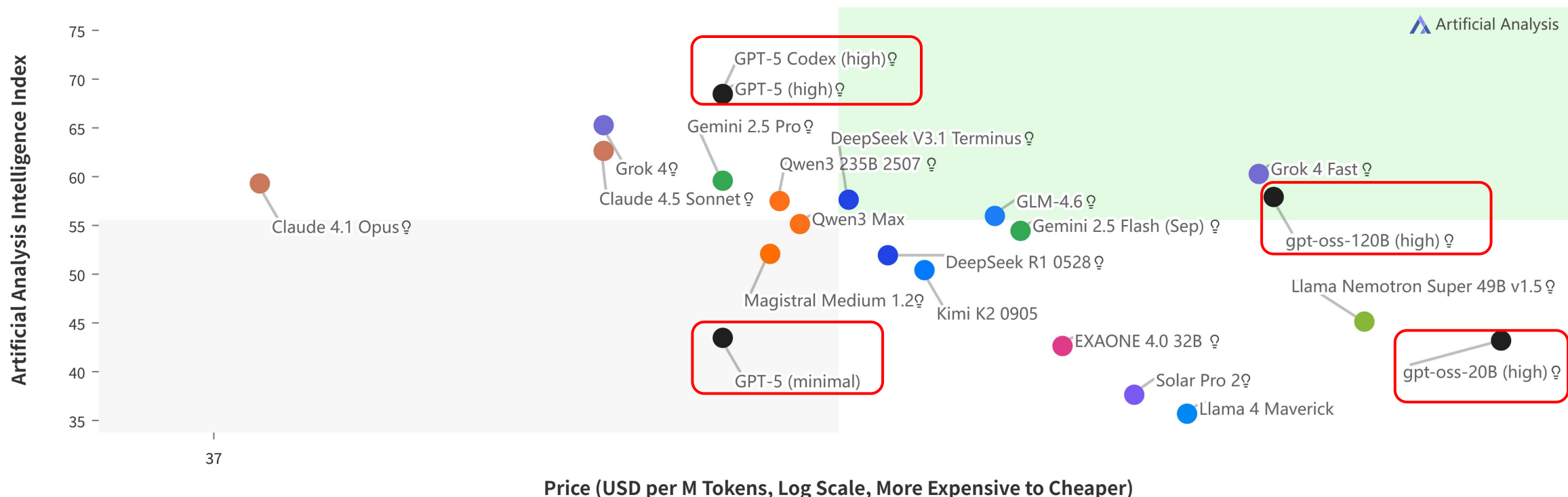
- 浅绿色区域是图表标注的“最具吸引力象限”，同时满足“智能水平高”和“价格便宜”，是“性价比最优”的区间，如xAI的Grok4 Fast、开源类的gpt-oss-120B（high）等。

图：AI模型智能性与价格分布（从左到右，价格从贵到便宜）

Artificial Analysis Intelligence Index; Price: USD per 1M Tokens; Inspired by prior analysis by Swyx

Most attractive quadrant

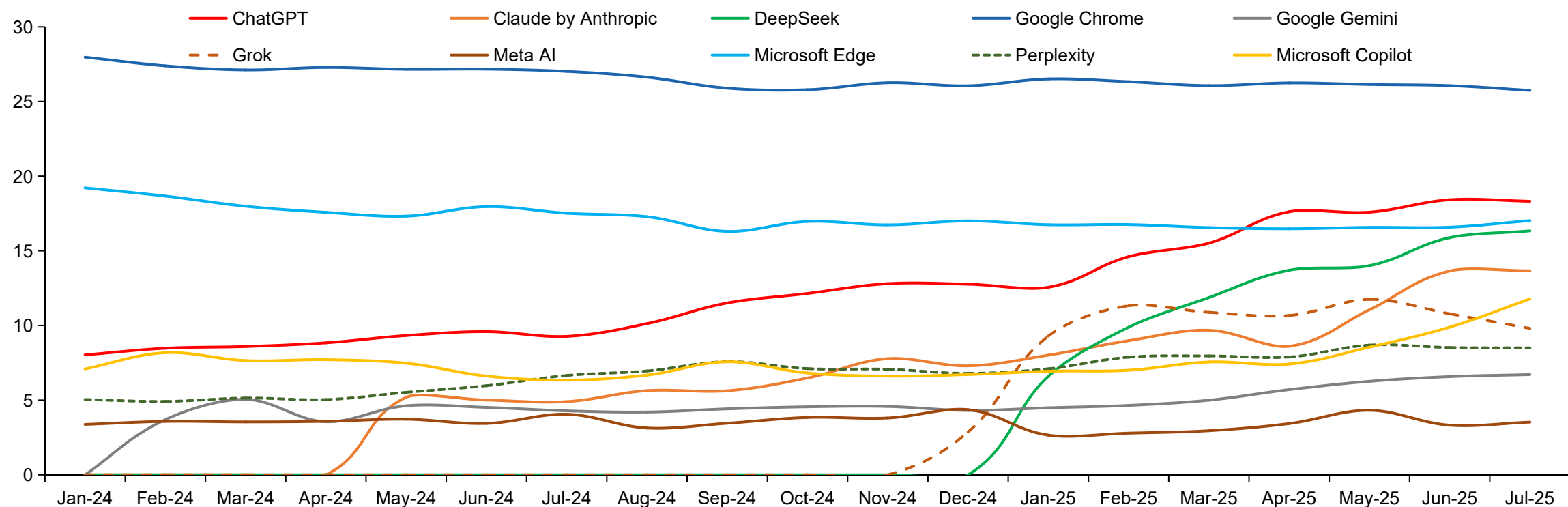
Alibaba Anthropic DeepSeek Google LG AI Research Meta Mistral Moonshot AI NVIDIA OpenAI Upstage xAI Z AI



热门AI APP用户时长：与底层大模型性能正相关

- 成熟浏览器（Chrome、Edge）凭借用户习惯和生态优势，用户时长较高、但有小幅下滑趋势。
- AI APP中用户时长与模型能力显著正相关。** ChatGPT持续发布迭代版本，凭借高性能模型 用户时长持续上升、已接近日均20分钟，一直处于AI应用首位。DeepSeek伴随V3和R1模型更新25年实现用户时长爆发式增长。Grok 2024年底发布Grok2 后用户时长才开始显著增长。Claude by Anthropic，25年初发布Claude 3.7 sonnet后用户时长增速显著提升。

图：AI应用APP用户时长



资料来源：Similarweb，国信证券经济研究所整理

第一，宏观经济波动。若宏观经济波动，公司业务、产业变革及新技术的落地节奏或将受到影响。

第二，下游需求不及预期。若下游AI需求不及预期，相关的AI研发投入增长或慢于预期，致使行业增长不及预期。

第三，核心技术水平升级不及预期的风险。AI大模型研发进度落后，AIGC相关产业技术壁垒较高，核心技术难以突破，影响整体进度。

第四，AI快速迭代、平权化下竞争加剧，影响云业务利润率。

国信证券投资评级			
投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后6到12个月内的相对市场表现，也即报告发布日后的6到12个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A股市场以沪深300指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数（HSI.HI）作为基准；美国市场以标普500指数（SPX.GI）或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数10%以上

分析师承诺

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。本报告仅供我公司客户使用，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、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意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。



国信证券
GUOSEN SECURITIES

国信证券经济研究所

深圳

深圳市福田区福华一路125号国信金融大厦36层

邮编：518046 总机：0755-82130833

上海

上海浦东民生路1199弄证大五道口广场1号楼12楼

邮编：200135

北京

北京西城区金融大街兴盛街6号国信证券9层

邮编：100032