下载日志已记录，仅供内部参考

国泰海通证券 GUOTAI HAITONG SECURITIES

AI 大模型行业深度报告：模型跃迁，应用生花

传播文化业

评级： 增持

姓名	电话	邮箱	登记编号
陈筱(分析师)	021-38675863	chenxiao@gtht.com	S0880515040003
陈星光(分析师)	021-23219104	chenxingguang@gtht.com	S0880525040043
赵祷旋(分析师)	021-23183268	zhaoyini@gtht.com	S0880526080002

本报告导读：

模型能力提升与调用成本下降推动 AI 商业化加快，多模态和智能体持续拓展产品功能，内容生产、营销、教育和游戏等应用进入快速发展阶段，看好“AI+视频、游戏、营销”带来的投资机会。

投资要点：

- **模型能力提升、使用成本下降，AI 由效率工具向任务助手发展。**头部模型在推理、代码、长上下文、多模态和工具调用等方面持续进步，模型调用价格明显下降。2025 年企业生成式 AI 支出达到 370 亿美元，其中应用层支出为 190 亿美元。随着端侧设备、多模态和智能体能力完善，AI 的使用入口由网页和软件扩展至手机、PC 及智能硬件，可完成的任务也由问答和内容生成延伸至资料整理、办公协作和任务执行。
- **模型层竞争转向能力、生态、成本与商业化的综合比较，头部厂商形成差异化优势。**OpenAI 等模型厂商依靠订阅、企业席位和 API 调用直接收费，并向智能体、广告和企业平台多元化布局；Google 等生态巨头则将模型接入云、搜索、办公软件和终端，借助原有用户、数据及收费体系扩大 AI 使用。具体来看，OpenAI 向综合性 AI 平台拓展业务，Anthropic 聚焦企业客户与 Coding，MiniMax 依托全模态模型和海外消费应用扩大订阅及 API 收入，月之暗面将资源集中于模型效率和 Agent 产品，阶跃星辰则将手机、汽车等高频终端作为模型入口。
- **多模态推动 AI 进入内容生产流程，营销、教育和游戏应用加快商业化。**图像生成与编辑可提高广告素材供给和游戏美术效率，视频模型推动 AI 剧、漫剧增长，并逐步用于影视特效、剪辑和虚拟拍摄；推荐模型可改善广告匹配和转化，智能辅导可提供个性化教学，游戏开发工具则由素材生成延伸至代码、场景和对象编辑。随着生成质量、编辑能力和版权规则完善，AI 应用有望由单点工具进一步进入企业日常生产流程。
- **投资建议：看好 AI+视频、游戏和营销三条主线。**影视公司拥有内容、IP 和分发渠道，可利用 AI 提高素材制作、后期处理和内容运营效率，推荐中文在线、昆仑万维、芒果超媒、上海电影、华策影视；游戏公司可将 AI 用于美术、代码、测试、UGC 工具和智能角色，推荐世纪华通、恺英网络、巨人网络、完美世界、心动公司；营销公司可批量生成多语言素材，并结合用户和投放数据提高广告匹配效率，推荐易点天下、值得买。
- **风险提示：**大模型技术进展不及预期；AI 应用商业化进展不及预期；行业竞争加剧；政策监管、数据合规及版权风险；算力供给及资本开支不及预期。

相关报告

传播文化业《《Last Asylum: Plague》《Lordrush》入围海外收入增长榜，视频生成模型 Vidu S2 发布》2026.09.19

传播文化业《抖音加速布局 AI 应用，2026 亚运会即将开幕》2026.09.12

传播文化业《头部游戏公司中报业绩表现出色，下半年新品有望陆续上线》2026.09.10

传播文化业《《后西游记》上线，AI 视频渗透至长剧市场》2026.09.06

传播文化业《AI 重塑游戏：研发提效、玩法创新与供给扩容》2026.09.05

下载日志已记录，仅供内部参考

目录

1. AI 技术持续迭代，应用商业化加快 3
1.1. 基础设施、模型和应用构成 AIGC 产业链三大环节 3
1.2. 模型能力提升调用成本下降，海外 AI 应用商业化提速 3
1.3. 生成式 AI 使用范围持续扩大，从效率工具走向任务助手 5
1.4. 端侧 AI 拓展手机、PC、智能硬件新入口 6
1.5. 多模态与 Agent 化推动 AI 应用升级 8
2. 模型层：能力持续提升，厂商围绕成本、生态和商业化展开竞争 8
2.1. 长上下文、推理与多模态能力提升 8
2.2. 海外 AI 竞争格局分化，科技巨头与模型厂商各具优势 10
2.2.1. 云业务支撑算力扩张，科技巨头强化基础设施优势 10
2.2.2. 能力竞争：头部厂商相互追赶，迭代速度与调用成本更受关注 11
2.2.3. 生态建设：成熟应用推动 AI 进入更多使用环节 12
2.3. 模型层商业模式 13
2.3.1. API 收费：按 tokens 计价与生成内容量计价 13
2.3.2. 订阅收费：按时间周期计价面向 C 端和 B 端用户 13
2.3.3. 广告收费：依托流量、用户意图与成熟广告体系变现 15
2.3.4. 定制化收费：覆盖本地与云端部署，满足企业差异化需求 15
3. 头部大模型厂商：模型能力加速变现，商业路径逐步分化 16
3.1. OpenAI：多元业务拓宽收入边界，高投入推动平台化扩张 16
3.2. Anthropic：聚焦模型与 Coding 优势，企业收入快速放量 19
3.3. MiniMax：全模态与全球化双轮驱动，消费应用与 API 协同增长 21
3.4. 月之暗面：模型能力推动商业化，订阅与 API 收入提速 24
3.5. 阶跃星辰：多模态切入智能终端，端侧合作打开增长空间 26
4. 应用层：多模态拓展产品形态，海外 AI 应用加快商业化 27
4.1. AI 应用分类 27
4.2. 对话与搜索类 27
4.2.1. 聊天机器人 27
4.2.2. AI 搜索 28
4.3. 图像类 28
4.4. 音视频类 30
4.4.1. 音乐生成 30
4.4.2. 视频生成编辑 31
4.5. 营销类 32
4.6. 教育学习类 34
4.7. 游戏类 35
4.7.1. 游戏开发生成 35
4.7.2. 智能 NPC 36
5. 投资建议 37
6. 风险提示 39

下载日志已记录，仅供内部参考

1. AI 技术持续迭代，应用商业化加快

1.1. 基础设施、模型和应用构成 AIGC 产业链三大环节

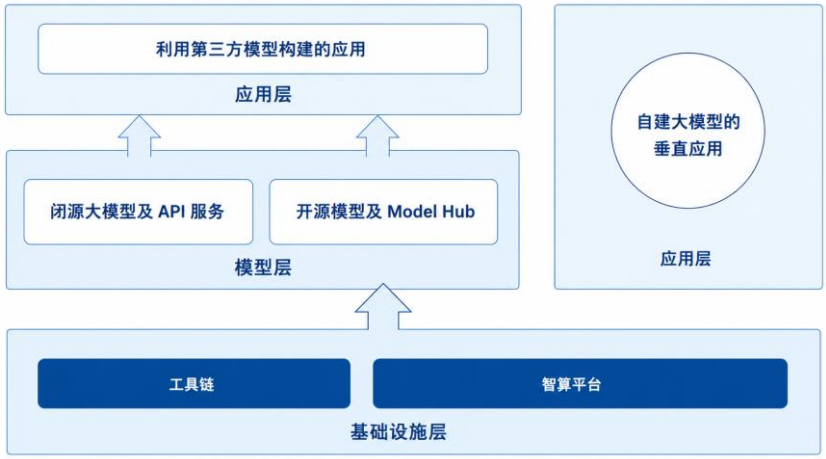
AIGC 产业链主要包括基础设施、模型和应用三大环节，分别提供算力、模型能力和面向用户的产品。基础设施支撑模型训练和推理，模型层向下游输出技术能力，应用层通过具体产品形成用户需求和商业收入。

1）基础设施层为模型训练、推理和部署提供支撑。主要包括 AI 芯片、服务器、数据中心和云平台，以及数据标注、模型微调等数据服务和开发工具。

2）模型层连接上游算力和下游应用，是产业链的核心环节。当前模型分为闭源和开源两类，能力已由文本扩展至图像、音频和视频，并通过 API、MaaS 和私有化部署向企业及开发者提供服务。

3）应用层直接面向用户，是模型能力转化为产品和收入的重要环节。相关产品既包括基于第三方模型开发的通用应用，也包括依托自研模型推出的垂直应用，覆盖对话搜索、办公、编程、图像、音视频、教育、营销和游戏等领域。

图1：AIGC 产业链包含基础设施、模型和应用三个环节



资料来源：启明创投生成式 AI 2023，国泰海通证券研究

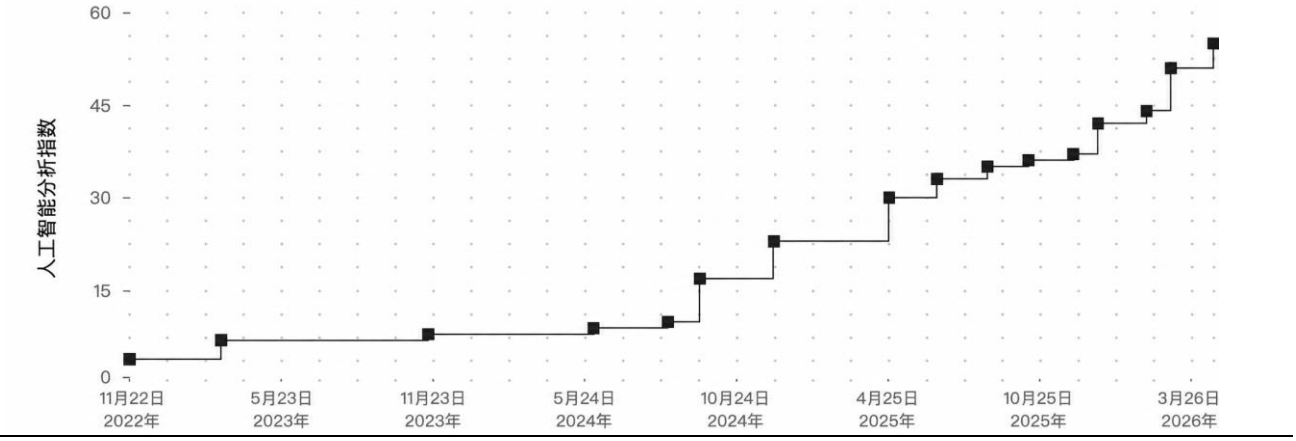
AIGC 产业正从技术竞争转向产品和收入竞争。模型训练和推理需求仍将带动云服务、AI 芯片、服务器和数据中心投入；模型层的竞争重点也由参数规模转向推理能力、多模态能力、调用成本和产品入口。应用能否形成高频使用、持续付费和稳定留存，将直接影响产业商业化进展。基础设施提供算力和开发工具，模型层输出技术能力，应用层通过具体产品形成收入，三者共同推动 AI 产业增长。

1.2. 模型能力提升调用成本下降，海外 AI 应用商业化提速

头部厂商持续推出新模型，模型能力提升明显。据 Artificial Analysis，OpenAI 模型智能指数由 2022 年的个位数升至 2026 年 GPT-5.5 的 55 分，且 2024 年下半年后提升速度加快。当前竞争重点已由参数规模转向推理、代码、工具调用和长任务处理等能力，模型迭代仍是 AI 应用发展的基础。

下载日志已记录，仅供内部参考

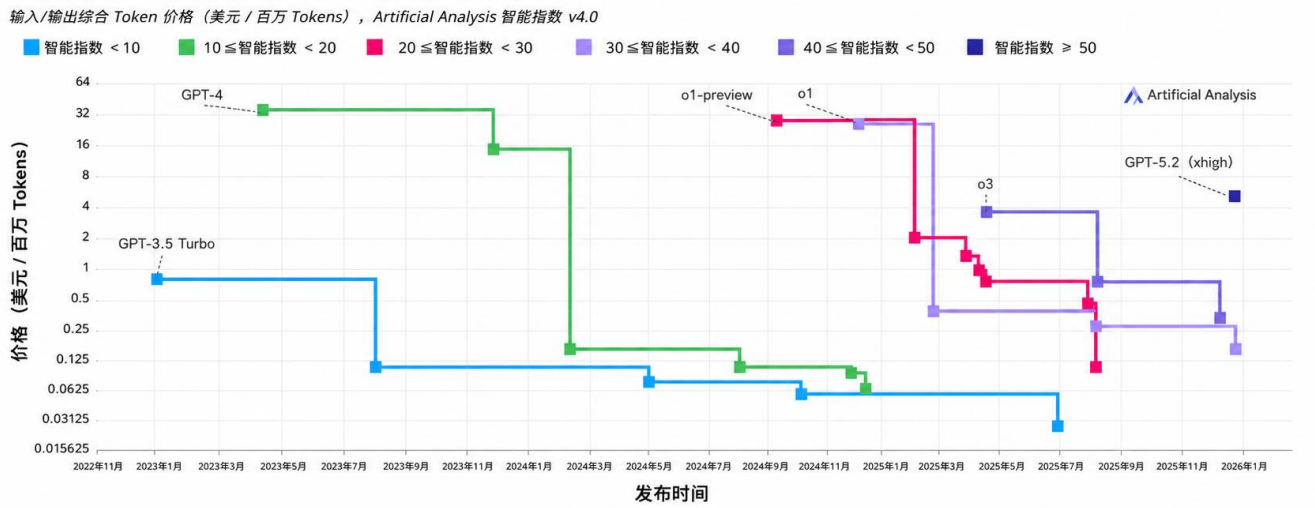
图2：OpenAI 模型智能水平持续提升（2022—2026 年）



资料来源：Artificial Analysis

模型调用价格明显下降，开发者可以用更低成本获得旗舰模型的能力。Artificial Analysis 按模型智能水平比较输入与输出混合价格，达到 o1 智能水平的模型，2026 年的最低调用价格已较 o1 发布初期下降 128 倍。从 GPT-4、o1 到 GPT-5.2，旗舰模型价格也逐步下行。模型能力的提升以及调用成本的下降，正在降低 AI 应用的开发与使用成本。

图3：模型能力持续提升的同时，调用成本显著下降



资料来源：Artificial Analysis State of AI: 2025 Year-End Edition

AI 应用已覆盖通用助手、内容创作、办公和编程等主要品类。据 a16z，ChatGPT、Gemini 和 Claude 等通用助手仍处于用户规模前列；Canva、CapCut 和 Adobe 等内容创作工具保持较高活跃度；Notion、Grammarly 等办公工具继续增长；Cursor、Lovable 等 AI 编程产品进入网页端榜单前列。AI 功能正逐步进入办公、设计和软件开发等高频使用环节。

下载日志已记录，仅供内部参考

图4：前 50 名生成式 AI 网页产品



资料来源：a16z

图5：前 50 名生成式 AI 软件产品



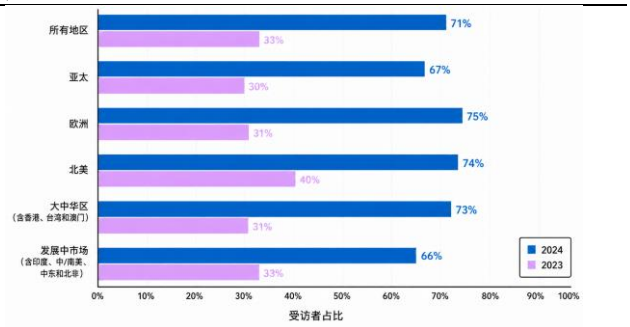
资料来源：a16z

1.3. 生成式 AI 使用范围持续扩大，从效率工具走向任务助手

ToC 端，通用 AI 助手渗透率快速提升，生成式 AI 已进入大众使用阶段。据 OpenAI 2025 年 9 月发布的研究(覆盖 150 万条 ChatGPT 对话)，ChatGPT 周活跃用户已达 7 亿。AI 产品应用场景从通用问答扩展至搜索、写作、图像、视频、角色陪伴，多模态能力支持用户以自然语言处理文本、图像、音视频等多类型内容。

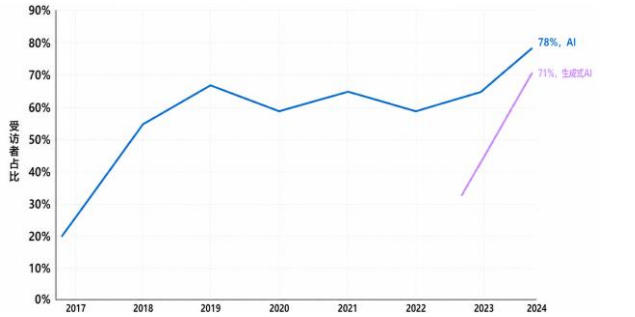
ToB 端，企业采用率及支出规模同步高增，效果易量化环节率先形成付费。据 Stanford AI Index 2025,2024 年 71%受访企业已在至少一个业务职能中常规使用生成式 AI，较 2023 年的 33%大幅提升，与整体 AI 使用率 78%的差距收窄至 7pct。据 Menlo Ventures，2025 年企业生成式 AI 支出达 370 亿美元，为 2024 年的 3.2 倍，其中应用层支出 190 亿美元，占比超 50%。代码生成、销售支持、客服问答、会议纪要、文档处理、营销素材制作等任务标准化程度高，降本增效更易衡量，商业化落地更快。据 Stanford AI Index 2025，供应链与库存管理、服务运营、人力资源、战略与公司财务四大职能成本下降的企业占比分别达 61%、58%、56%、56%;战略与公司财务、供应链与库存管理、营销与销售收入提升的企业占比分别达 70%、67%、66%。

图6：2024 年全球企业人工智能使用率较 2023 年明显提升



资料来源：Artificial Intelligence Index Report 2025

图7：企业人工智能与生成式人工智能使用率于 2024 年大幅提升



资料来源：Artificial Intelligence Index Report 2025

图8：全球企业生成式 AI 支出规模快速增长

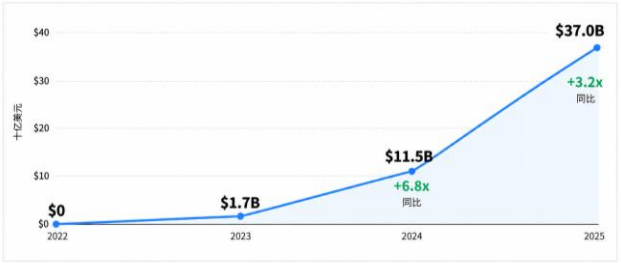
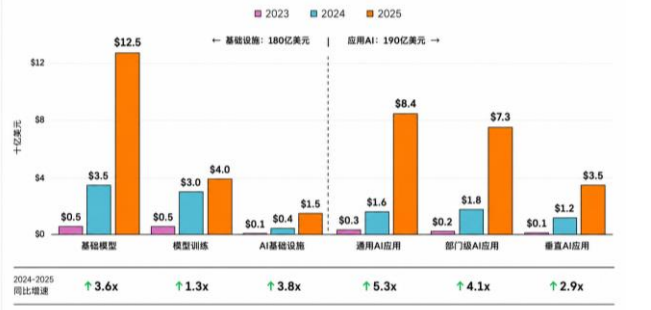


图9：企业生成式 AI 支出增长主要来自应用层

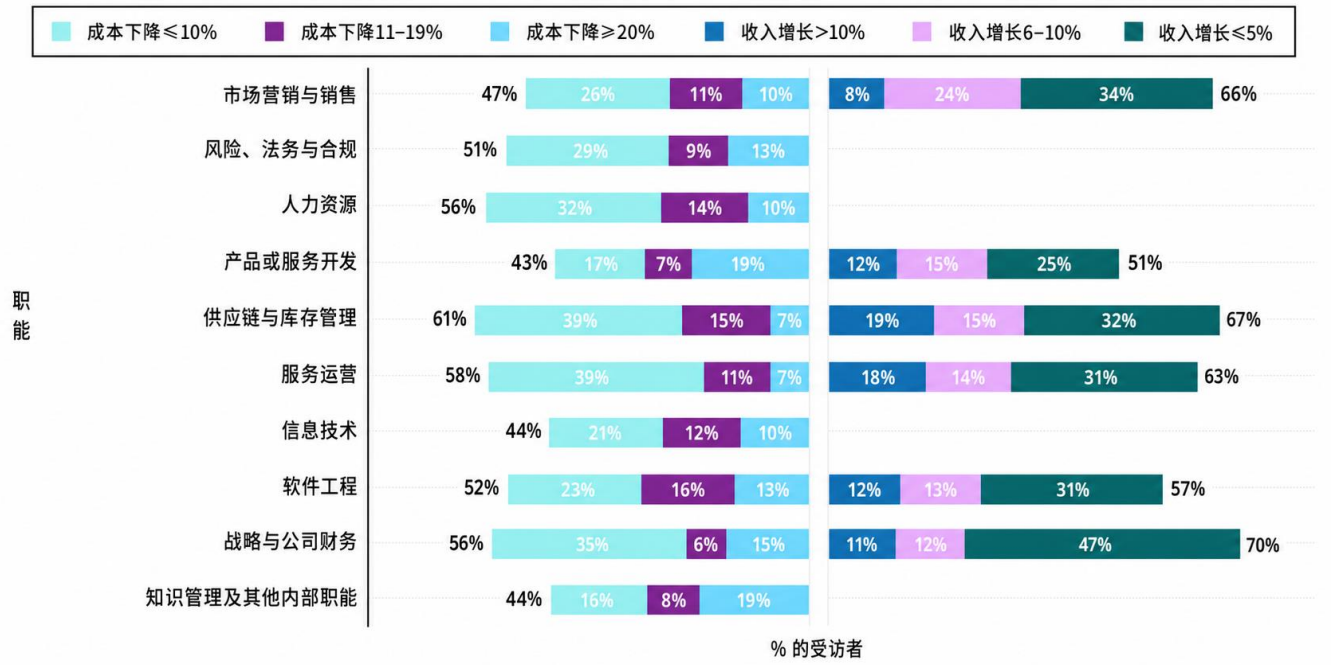


下载日志已记录，仅供内部参考

资料来源：2025:The State of Generative AI in the Enterprise

资料来源：2025:The State of Generative AI in the Enterprise

图10：生成式 AI 在供应链、营销及服务等职能中已实现显著降本增收



资料来源：Artificial Intelligence Index Report 2025

Agent 正推动 AI 从问答、内容生成迈向任务执行。模型推理、长文本处理、多模态理解与工具调用能力提升后，AI 开始承担任务拆解、资料整理、工具调用与结果反馈。面向个人用户，AI 逐步渗透搜索、学习、创作、陪伴、娱乐；面向企业，AI 切入代码开发、客户服务、销售跟进、营销投放、数据分析等具体环节。我们认为，能解决明确问题、嵌入日常流程、结果稳定的产品，更易从试用走向持续付费。

1.4. 端侧 AI 拓展手机、PC、智能硬件新入口

端侧 AI 将部分模型能力部署于本地，更适配低延迟、隐私保护和离线场景。相比云端调用，端侧 AI 响应更快、对网络依赖更低，且减少个人数据上传，已用于语音助手、图像编辑、实时翻译、会议纪要、个人知识管理等功能。**AI PC 和 AI 手机加快放量，渗透率有望从可选配置转向主流配置。**相比传统产品，AI PC 和 AI 手机搭载 NPU、支持端侧模型运行，响应速度、隐私保护、离线处理和跨应用协同上更具优势。据 Gartner，全球 AI PC 渗透率将由 2024 年 15%提升至 2025 年 31%，2026 年进一步达约 55%，2029 年有望成为 PC 市场主流形态。据 Canalys，AI 手机占全球智能手机出货量比重将由 2024 年 16%提升至 2028 年 54%，2023-2028 年出货量 CAGR 约 63%。我们认为，AI 终端渗透率仍处于快速提升初期，随芯片算力、操作系统及系统级 Agent 能力完善，AI PC 和 AI 手机有望承接更多办公、创作、实时翻译和个人助理功能，成为新一轮换机的重要驱动力。

下载日志已记录，仅供内部参考

图11：微软 AI PC



资料来源：Microsoft 官网

图12：谷歌 AI 手机



资料来源：36 氪

AI 智能眼镜已由产品验证进入规模放量阶段。2025 年，EssilorLuxottica 与 Meta 合作的 Ray-Ban Meta、Oakley Meta 等 AI 眼镜销量超 700 万副；2025 年下半年全球智能眼镜出货量同比增长 139%，Meta 市场份额提升至 82%。以 Ray-Ban Meta 为代表的产品已支持免手持拍摄、通话、语音问答、实时翻译、环境识别和内容记录，眼镜与摄像头、麦克风结合，使 AI 能持续获取用户第一视角信息。

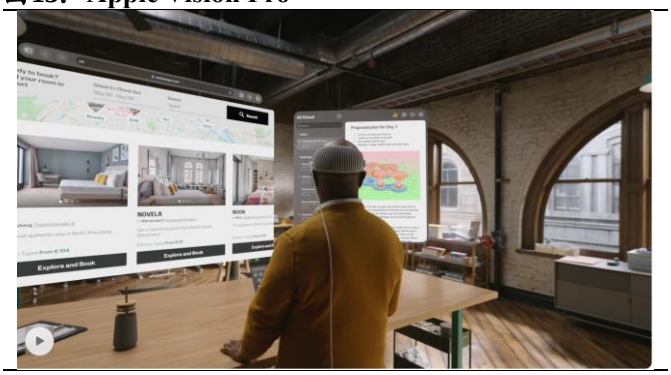
AI 玩具加快渗透，产品供给持续丰富。AI 玩具主要用于对话陪伴、故事生成和互动学习；据 WIRED，截至 2025 年 10 月，中国注册 AI 玩具企业已超过 1500 家，Miko 累计销量超 70 万台，华为智能憨憨上市首周销量达 1 万台。据 Grand View Research，包含 AI 互动机器人、联网教育玩具等产品的全球智能玩具市场规模将由 2025 年的 144 亿美元增至 2033 年的 440 亿美元，2026—2033 年复合增速为 15.8%。

AI 耳机和手表渗透率持续提升。据 Counterpoint，端侧 AI 可穿戴设备渗透率将由 2025 年约 30%提升至 2032 年近 80%，耳机和智能手表是主要增长品类。AI 耳机可提供实时翻译、语音助手、通话降噪和环境自适应等功能，苹果已在 AirPods 中引入由端侧 Apple Intelligence 支持的面对面实时翻译。智能手表则依托持续佩戴和心率、温度、运动等传感器，将 AI 用于运动识别、睡眠分析、个性化健康建议和任务执行。2026Q1，支持端侧 AI 的智能手表出货量同比+70%，占全球智能手表出货量的 25%。

端侧 AI 的应用范围正从手机和穿戴设备扩展至机器人、汽车等物理设备。与 AI 眼镜、手机、耳机、手表等设备主要理解用户指令并反馈信息不同，机器人和汽车还需要持续感知周围环境、预测物体运动并生成控制动作，因此对环境仿真、合成数据和决策训练提出更高要求。2026 年，NVIDIA 发布 Cosmos 3，将文本、图像、视频、环境音和动作统一至同一物理 AI 世界模型，可用于环境仿真、合成数据生成以及机器人和自动驾驶策略训练。我们认为，随芯片算力、续航和端云协同能力提升，AI 有望由手机和 PC 中的功能模块延伸至智能眼镜、玩具、机器人及汽车等，增加用户与 AI 的交互时长和使用入口。

下载日志已记录，仅供内部参考

图13：Apple Vision Pro



资料来源：Apple 官网

图14：物理 AI NVIDIA Cosmos



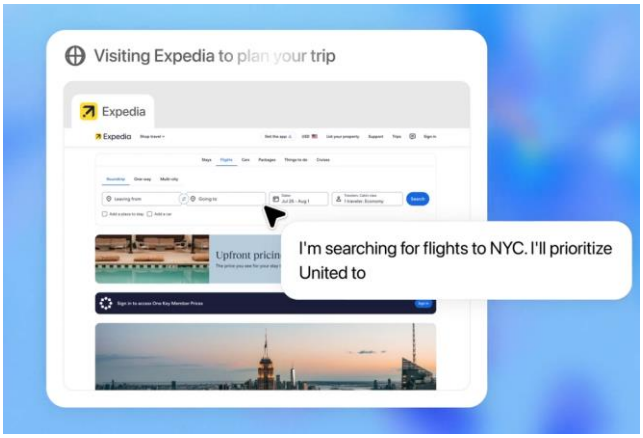
资料来源：Nvidia 官网

1.5. 多模态与 Agent 化推动 AI 应用升级

多模态能力推动 AI 应用从单一功能工具走向综合助手。早期 AI 产品主要用于聊天问答、文本写作和代码补全，当前已整合文本、图像、语音、视频、文件和搜索等能力，ChatGPT、Gemini、Claude、Microsoft Copilot 等通用 AI 助手开始承载更多输入输出形态和使用场景。以 ChatGPT 为例，GPT-5 提升视觉、视频、空间和科学图表理解能力。用户可在同一产品中完成信息检索、内容分析、文本写作和图像视频创作，AI 应用的功能范围扩大。

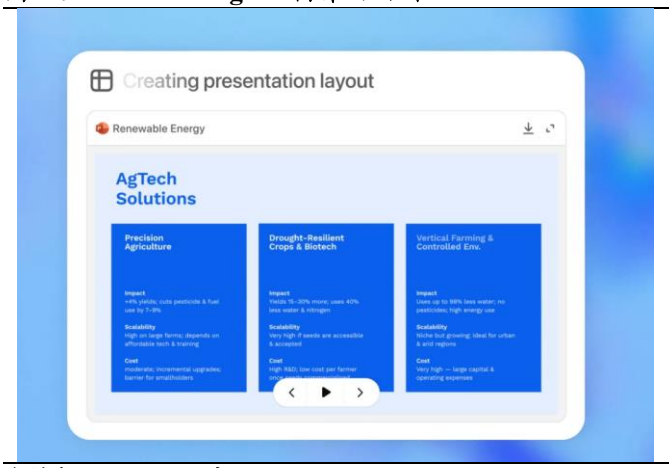
Agent 推动 AI 从回答问题转向完成任务。相比主要提供答案或生成内容的传统 AI 工具，Agent 可围绕用户目标自主拆解任务、检索资料、调用工具并交付结果。以 ChatGPT Agent 为例，其通过整合网页浏览、代码执行和外部应用连接，可完成资料研究、行程预订、表格更新和演示文稿制作。我们认为，Agent 将推动 AI 应用由逐轮问答转向目标委托，能力由内容生成延伸至执行具体任务，并提升使用频次与付费价值，拓展订阅及按量计费空间。

图15：ChatGPT Agent 规划行程



资料来源：OpenAI 官网

图16：ChatGPT Agent 制作幻灯片



资料来源：OpenAI 官网

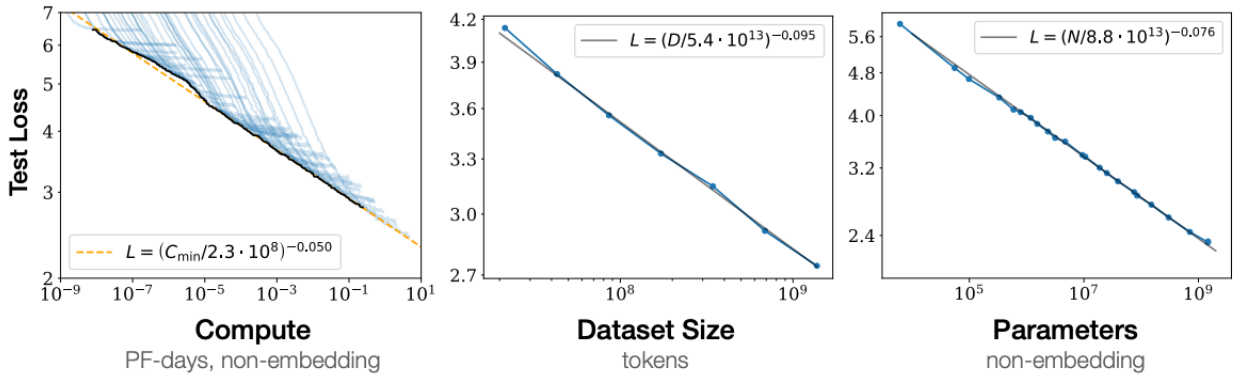
2. 模型层：能力持续提升，厂商围绕成本、生态和商业化展开竞争

2.1. 长上下文、推理与多模态能力提升

Transformer 架构与 Scaling Law 推动大模型能力提升。大语言模型通过大规模文本和代码训练，学习语言规律并根据上下文生成内容。Transformer 的注意力机制提高了并行训练效率，使模型能够扩展至更大的参数和数据规模；Scaling Law 则表明，增加参数、训练数据和算力通常可以继续降低模型损失。海外头部厂商由此持续扩大训练投入，推动模型理解和生成能力提升。

下载日志已记录，仅供内部参考

图17：Scaling Law 的机制

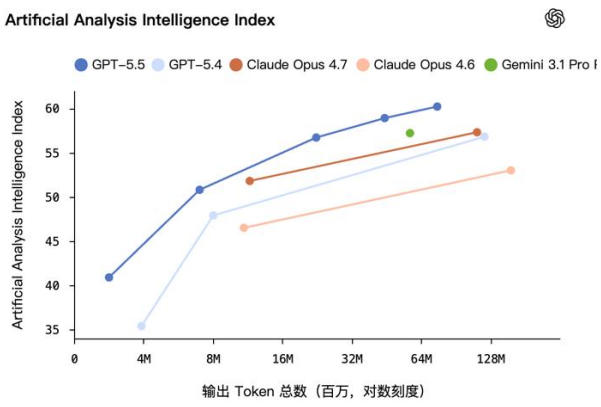


资料来源：Kaplan et al. (2020) Scaling Laws for Neural Language Models

长上下文与推理能力同步提升，模型处理复杂资料和多步骤任务的能力增强。GPT-5.5 支持 110 万 tokens 上下文和 12.8 万 tokens 最大输出，并提供多档推理强度；其在 Terminal-Bench 2.0、Expert-SWE、FinanceAgent 和 FrontierMath Tier 4 上的得分分别为 82.7%、73.1%、60.0%和 35.4%，均高于 GPT-5.4。Claude Opus 4.6 将上下文窗口扩展至 100 万 tokens，在 MRCR v2 百万 tokens 检索评测中得分达到 76%，显著高于 Claude Sonnet 4.5 的 18.5%；Claude Opus 4.8 进一步将 100 万 tokens 上下文和 12.8 万 tokens 最大输出设为默认配置，并通过推理强度设置平衡效果、速度和成本。

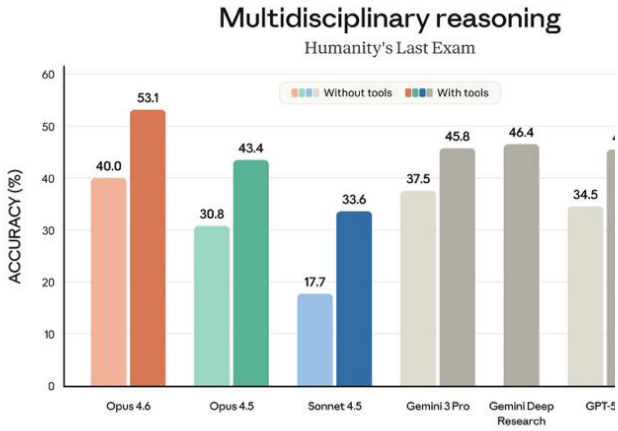
能力提升正在改变复杂任务的完成方式。模型可一次读取完整报告、代码仓库等大规模资料，并在较少人工干预下完成信息筛选、步骤规划和结果校验，减少分段输入、反复补充和中途修正。企业端的价值也由单次问答扩展至代码开发、研究分析、知识检索及 Agent 任务，竞争重点将更多落在完整任务的交付质量，而非单轮回答表现。

图18：GPT-5.5 在推理、代码及长上下文等能力上持续提升



资料来源：OpenAI 官网

图19：Opus 4.6 增强了专家级推理的能力



资料来源：Anthropic 官网

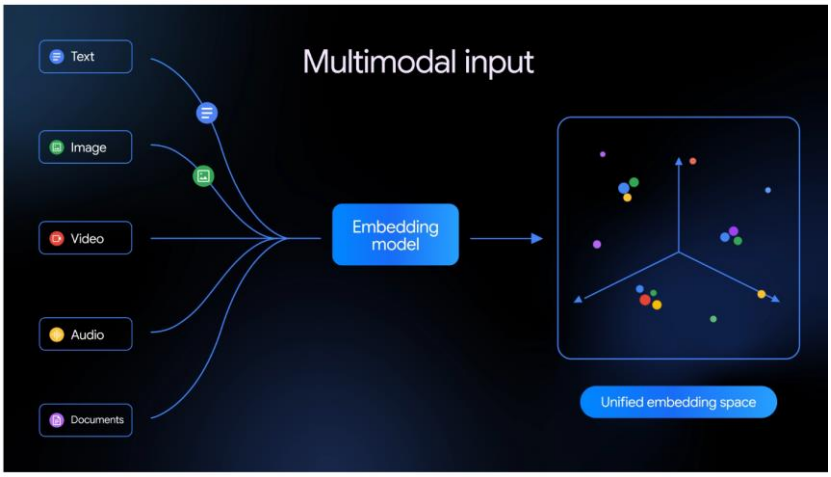
多模态模型由单一内容识别升级为跨模态理解与检索。GPT-4o 在同一模型中处理文本、音频、图像和视频输入，并支持文本、音频和图像输出，音频响应最快为 232 毫秒、平均为 320 毫秒；GPT-5 进一步提升视觉问题求解、视频理解、科学图表和空间推理能力，MMMU 得分达到 84.2%。Google 则将多模态能力用于企业信息检索，Gemini Embedding 2 可把文本、图片、视频、音频和文档映射至统一向量空间，覆盖 100 多种语言，支持跨模态搜索、分类和聚类。

多模态能力提升工作效率，进一步赋能内容产业链。多模态模型可统一处理文本、图像、音频和视频等信息，使各项多模态任务可以在同一产品内连

下载日志已记录，仅供内部参考

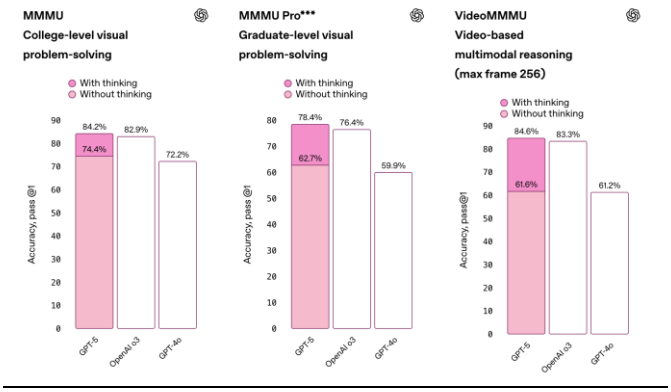
续完成，减少软件切换和人工汇总。随着图像、音频和视频生成质量持续提升，多模态大模型正逐步进入内容设计、素材制作、后期编辑等产业链环节。我们认为，这将为模型厂商和内容公司布局下游应用赛道提供技术基础，推动 AI 由辅助工具进一步融入内容生产过程。

图20：Gemini Embedding 2 将多种输入映射至统一向量空间



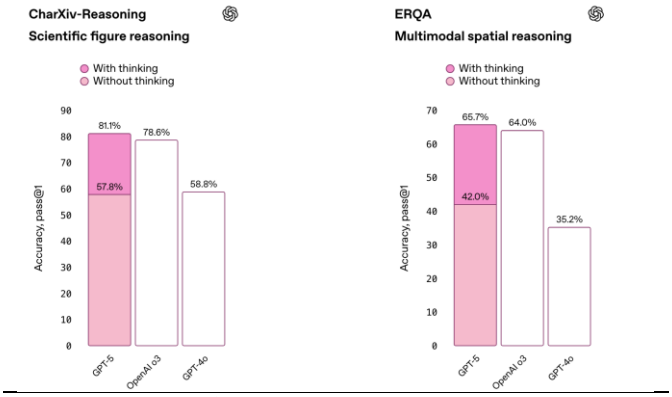
资料来源：wikidocs

图21：GPT-5 多模态理解与视频推理能力对比



资料来源：OpenAI 官网

图22：GPT-5 科学图表与空间推理能力对比



资料来源：OpenAI 官网

2.2. 海外 AI 竞争格局分化，科技巨头与模型厂商各具优势

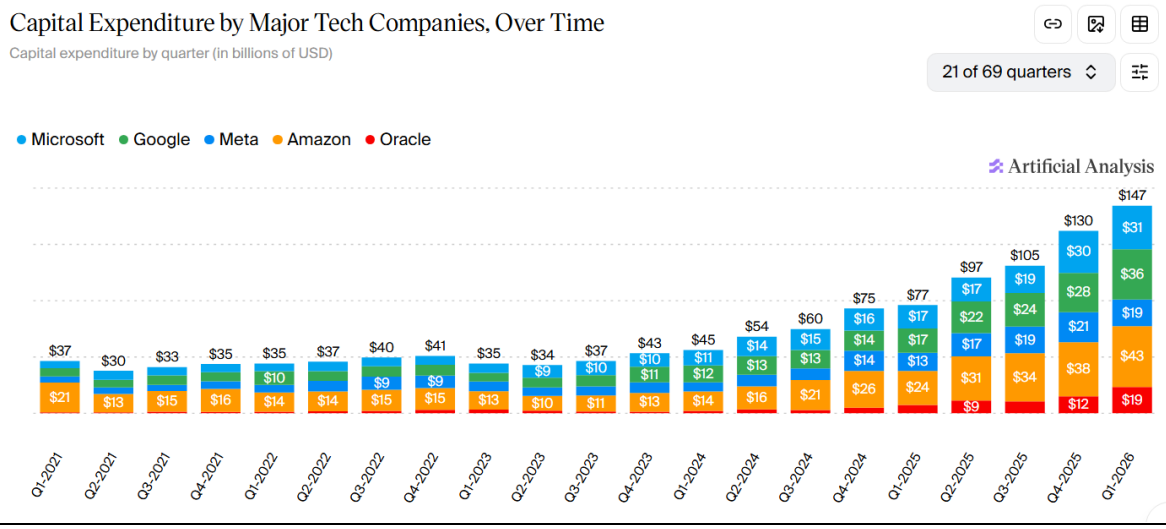
2.2.1. 云业务支撑算力扩张，科技巨头强化基础设施优势

科技巨头以云业务支撑 AI 基础设施扩张。据 Artificial Analysis，微软、Google、Meta、Amazon 和 Oracle 的 AI 相关基础设施支出由 2024Q1 的 450 亿美元增至 2026Q1 的 1470 亿美元，主要投向数据中心、电力、服务器和网络。相关资源既服务于自研模型训练和推理，也可通过云平台向企业销售算力、存储、模型调用及开发工具。2025 年 Azure、AWS 和 Google Cloud 收入分别超过 750 亿美元、达到 1287 亿美元和 587.05 亿美元，同比分别增长 34%、20%和 36%。

自有云平台提高了算力投入的回收效率。云厂商可通过外部客户使用云业务来分摊折旧、电力和运维成本，并依托既有企业客户推广 AI 服务。兼具云平台、数据中心和客户基础的厂商，更容易同时扩大算力供给、降低单位成本并形成稳定收入。

下载日志已记录，仅供内部参考

图23：科技公司在基础设施领域的资本支出



资料来源：Artificial Analysis

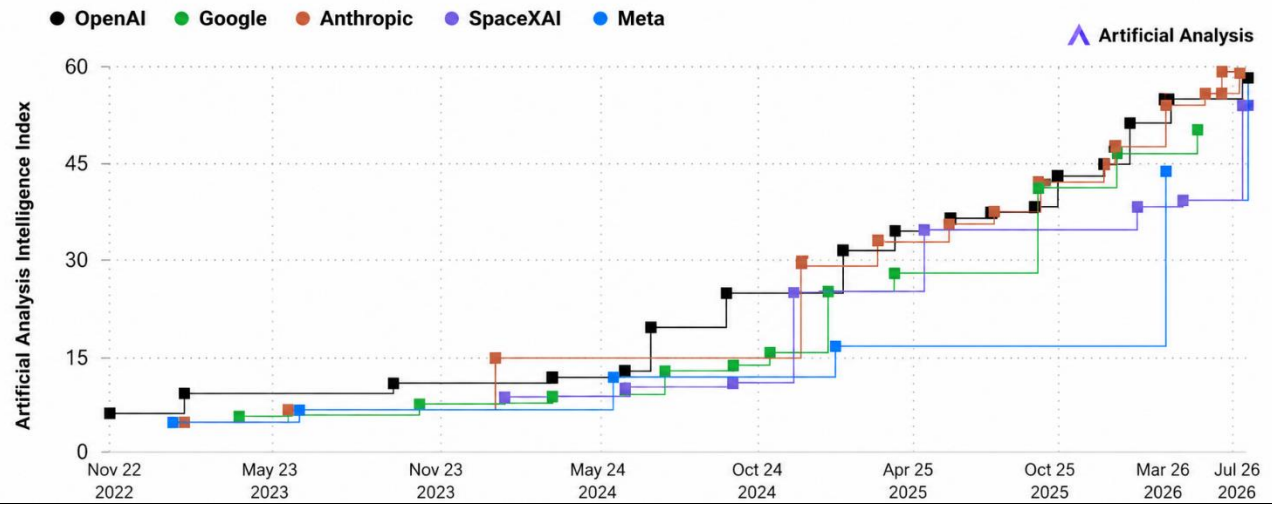
2.2.2. 能力竞争：头部厂商相互追赶，迭代速度与调用成本更受关注

头部厂商模型能力持续提升，OpenAI 与 Anthropic 保持领先，Google、xAI 和 Meta 加快追赶。据 Artificial Analysis，OpenAI 自 2022 年起率先建立能力优势，Anthropic 自 2023 年末开始快速提升，并在 2024 年后与 OpenAI 多次交替领先。2025 年下半年以来，Google、xAI 和 Meta 的新模型能力明显提高，头部厂商由 OpenAI 单独领先转向多家厂商相互追赶。

当前头部模型已集中在相近能力区间，单次版本领先幅度有限。据 Artificial Analysis，Claude Fable 5、Claude Opus 4.8、GPT-5.5 和 Grok 4.5 的模型智能指数分别达到 60 分、56 分、55 分和 54 分，Claude Sonnet 5、GPT-5.4 及 Gemini 3.5 Pro 也处于 50 分左右。Anthropic 和 OpenAI 仍占据榜单前列，但 xAI 和 Google 已缩小与他们的能力差距，模型之间多为个位数差异。

模型厂商依靠快速迭代和模型分层扩大优势。模型能力决定代码开发、研究分析和 Agent 等复杂任务的完成效果。一个模型通常无法长期保持领先地位，会被新模型快速追赶，并且搜索、办公和客服等高频任务并不需要始终调用旗舰模型，厂商会通过轻量化的模型、缓存输入和批量处理来降低成本。我们认为，模型竞争已由保持旗舰模型的能力领先，转向模型迭代速度、任务完成效果和调用成本的综合比较。

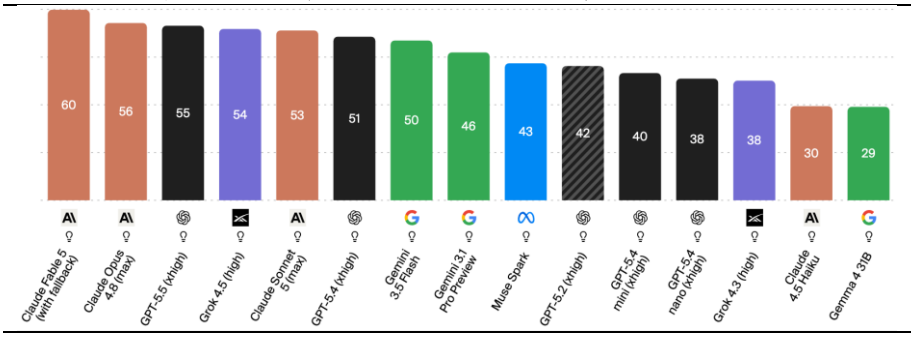
图24：头部厂商模型能力不断发展、相互追赶



资料来源：Artificial Analysis

下载日志已记录，仅供内部参考

图25：2026 年 7 月各头部厂商模型能力指数榜单



资料来源：Artificial Analysis

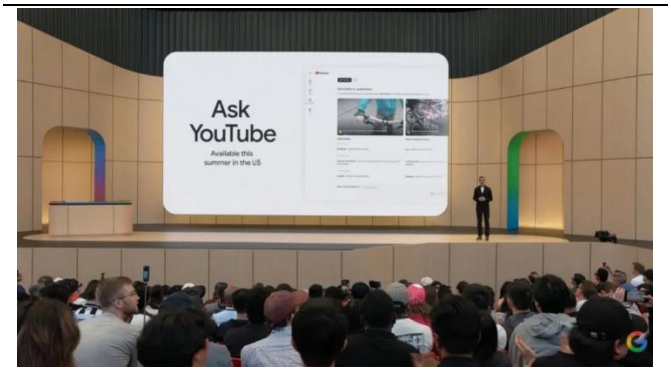
2.2.3. 生态建设：成熟应用推动 AI 进入更多使用环节

科技巨头通过将 AI 嵌入现有软件、平台和终端，直接覆盖既有用户。Meta 依托 Facebook、Instagram 和 WhatsApp 推广 AI。微软则将 Copilot 嵌入 Windows 和 Microsoft 365，截至 2026 年 1 月，Microsoft 365 Copilot 付费席位达到 1500 万，日活跃用户同比增长 10 倍；截至 2026 年 3 月，Meta 旗下应用日活跃用户合计达到 35.6 亿。AI 接入成熟产品后，可直接覆盖原有用户，并沿用数据、账号和付费体系，更快扩大使用。

Google 依托搜索、办公、浏览器和移动系统，形成 AI 产品矩阵。Gemini 已接入 Search、Maps、YouTube、Gmail 和 Chrome 等，覆盖搜索、文档创作、知识整理和日程管理等环节。AI Mode 支持多模态搜索和任务委托，Ask Maps 与 Ask YouTube 可查询地图和视频内容，Docs live、Daily Brief 和 NotebookLM 分别用于语音创作、日程摘要和资料整理。截至 2025 年末，AI Overviews 覆盖超过 20 亿用户，Gemini 应用月活跃用户超过 7.5 亿。Google 还通过 Android XR 眼镜拓展随身交互，使 Gemini 由网页和软件进一步进入日常穿戴使用。

xAI 围绕 X 和 Tesla 拓展 Grok 的使用范围。X Search 支持关键词、语义、用户和帖子线程检索，Grok 可获取并分析平台内的实时文本、图片和视频内容。Grok 进入 Tesla 后，用户可通过语音完成对话、信息查询和导航等操作。X 提供实时内容和社交信息，Tesla 增加车载交互。

图26：YouTube 新增 Ask YouTube 功能



资料来源：硅谷宇宙公众号

图27：Copilot 嵌入微软生态



资料来源：Microsoft 官网

AI 由独立应用工具转为贯穿生态体系的智能助手。AI 接入生态后，可在授权范围内调用生态中的所有数据与应用，由被动回答工具升级为执行多步骤任务的 Agent 助理；用户使用时间越长，AI 对其工作流程、表达习惯和业务规则理解越深，形成“使用增加—理解加深—体验改善—依赖提升”的飞轮，显著增强用户黏性和迁移成本。对模型厂商而言，收入主要来自模型订阅和 API 调用，而对生态巨头则借助 AI 提高原有产品价值，带动 SaaS 订阅、云服务、广告和硬件收入增长。

模型厂商与生态巨头的竞争边界逐步重合，竞争重点将从模型能力延伸至生态与成本的综合较量。我们认为，OpenAI 由订阅和 API 向企业平台、应

下载日志已记录，仅供内部参考

用生态组合、广告及硬件等业务拓展，Google 则持续加强旗舰模型、API 和开发工具，巩固模型层能力。模型厂商需要降低对纯模型收入和外部算力的依赖，生态巨头则需要保持模型竞争力并提高算力利用效率。模型能力决定产品能否吸引用户，应用生态体系决定 AI 能触达的消费者和企业规模，单位推理成本则决定收入能否转化为利润。AI 覆盖的产品和付费环节越多、获客和推理成本越低，收入越容易转化为利润，由此产生的现金流可继续投入模型研发和基础设施建设，巩固长期竞争优势，这也决定了模型层在 API、订阅、广告和定制化部署上的商业化选择。

2.3. 模型层商业模式

2.3.1. API 收费：按 tokens 计价与生成内容量计价

API 按照模型调用量收费，是模型厂商服务开发者和企业客户的主要方式。OpenAI 将文本模型调用拆分为输入、缓存输入和输出三类，其中 GPT-5.5 的价格分别为 5 美元、0.5 美元和 30 美元/百万 tokens，GPT-5.4 mini 则为 0.75 美元、0.075 美元和 4.5 美元/百万 tokens。不同性能和价格的模型可以满足复杂推理与轻量高频任务，缓存输入则有助于降低长文本和多轮调用成本。

表1：OpenAI 旗舰模型调用价格

模型	输入价格（美元/百万 tokens）	缓存输入价格（美元/百万 tokens）	输出价格（美元/百万 tokens）
gpt-5.5-pro	30	—	180
gpt-5.5	5	0.5	30
gpt-5.4	2.5	0.25	15
gpt-5.4-mini	0.75	0.075	4.5
gpt-5.4-nano	0.2	0.02	1.25

数据来源：DevTk.AI，国泰海通证券研究

多模态模型将计价范围从文本 Token 扩展至音频、图像、视频和语音转录。实时语音模型 gpt-realtime-2.1 根据音频、文本和图片的输入及输出 tokens 收费，图像生成模型 gpt-image-2 按照图片与文本 tokens 计价；视频生成模型 Sora 2 按生成时长收费，720p 价格为 0.10 美元/秒，Sora 2 Pro 在 720p、1024p 和 1080p 下的价格分别为 0.30 美元、0.50 美元和 0.70 美元/秒；gpt-4o-transcribe 的语音转录价格约为 0.006 美元/分钟。我们认为，随着模型处理内容由文本扩展至音频、图像和视频，API 收入来源也将由单一文本调用逐步扩展至更多内容生成和处理需求。

表2：OpenAI 多模态模型调用价格

模型	模态	输入价格（美元/百万 tokens）	缓存输入价格（美元/百万 tokens）	输出价格（美元/百万 tokens）	按时长价格	时长单位
gpt-realtime-2.1	音频	32.00	0.40	64.00	—	—
gpt-realtime-2.1	文本	4.00	0.40	24.00	—	—
gpt-realtime-2.1	图片	5.00	0.50	—	—	—
gpt-image-2	图片	8.00	2.00	30.00	—	—
gpt-image-2	文本	5.00	1.25	—	—	—
sora-2	720p	—	—	—	0.10	美元/秒
sora-2-pro	720p	—	—	—	0.30	美元/秒
sora-2-pro	1024p	—	—	—	0.50	美元/秒
sora-2-pro	1080p	—	—	—	0.70	美元/秒
gpt-4o-transcribe	转录	2.50	—	10.00	0.006	美元/分钟

数据来源：GitHub，国泰海通证券研究

2.3.2. 订阅收费：按时间周期计价面向 C 端和 B 端用户

订阅收费通过月度或年度付费将模型服务转化为经常性收入。面向个人用户，模型厂商主要根据模型权限、使用额度、多模态及 Agent 功能设置不同订阅档位；面向企业客户，则通常按照席位收费，并增加团队协作、权限管理、数据安全和企业软件连接等功能。

OpenAI 通过多档个人订阅扩大付费用户覆盖，并以 Business 和 Enterprise

下载日志已记录，仅供内部参考

服务企业客户。个人订阅包括 Free、Go、Plus 和 Pro 月费分别为 8 美元、20 美元和 100 美元，模型权限、使用额度、Deep Research、Agent 及 Codex 能力逐级提升。Business 年付价格为 20 美元/用户/月，提供集中管理、使用分析、SAML SSO 和 MFA 等功能；Enterprise 采用定制价格，进一步增加更长上下文、数据驻留、企业权限控制和服务保障。

图28：OpenAI 个人订阅价格



资料来源：NBAI.club

图29：OpenAI 商业与企业订阅价格

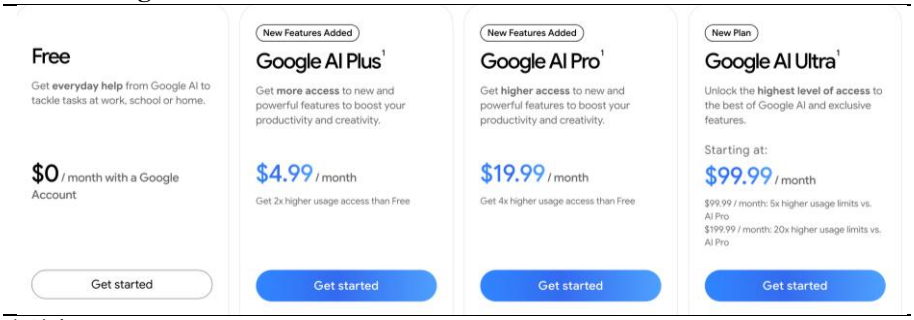


资料来源：NBAI.club

Google 通过 Gemini、Workspace 和云存储打包销售，订阅价值不局限于单一 AI 助手。个人订阅包括 Free、AI Plus、AI Pro 和 AI Ultra，月费分别为 0 美元、4.99 美元、19.99 美元和 99.99 美元，并逐级增加 Gemini、Flow、NotebookLM、办公软件及云存储权益。企业端则以 Google Workspace 为核心，按照用户和月份收费，将 Gemini 接入 Gmail、Drive、Docs、Meet 等产品。

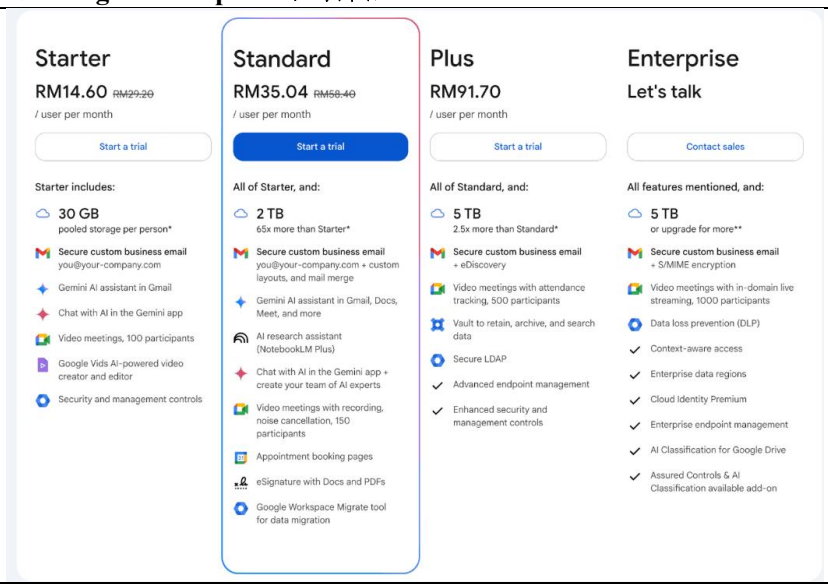
下载日志已记录，仅供内部参考

图30：Google AI 个人订阅价格



资料来源：SaaS CRM Review

图31：Google Workspace 订阅价格



资料来源：飞书

2.3.3. 广告收费：依托流量、用户意图与成熟广告体系变现

广告收费更适合拥有大规模免费流量、明确用户意图或兴趣数据及成熟投放体系的平台。搜索、社交和内容平台可在维持免费使用的同时，根据用户查询、浏览和互动信息匹配广告，并按照展示、点击或转化收费；完整的应用矩阵则有助于增加用户触点、丰富数据并扩大广告库存。Google 和 Meta 分别代表搜索意图驱动与内容推荐驱动两类模式。

Google 将广告由传统搜索结果延伸至 AI 生成的回答。现有搜索、购物和 Performance Max 广告可展示于 AI Overviews 周边或回答内部，平台综合用户查询和 AI 回答内容识别商业意图；AI Mode 也已引入回答内广告，使用户由信息检索进一步转向商品和服务选择。Google 披露，在美国、印度等主要市场，AI Overviews 使相关查询的搜索使用量提升超过 10%，商业查询量同步增加。

Meta 依托社交内容和 AI 推荐系统提升广告匹配效率。广告主可在 Facebook、Instagram 和 Messenger 等平台按展示或用户行为付费，Andromeda 负责从大量广告中筛选候选内容，GEM 进一步优化排序与转化。其中，Andromeda 使部分广告召回率和质量分别提升 6%和 8%；GEM 使 Instagram 和 Facebook 信息流广告转化率分别提升 5%和 3%。2025 年 Meta 广告收入约 1962 亿美元，同比增长 22%，广告展示量和平均价格分别增长 12%和 9%。

成熟流量入口和广告体系决定 AI 广告变现效率。AI 主要提升用户意图识别、广告匹配和转化效率，具备大规模用户、丰富行为数据及完整产品矩阵的平台，更易将 AI 使用量转化为广告收入。

2.3.4. 定制化收费：覆盖本地与云端部署，满足企业差异化需求

本地部署主要面向重视数据安全、自主控制和稳定调用的企业客户。企业

下载日志已记录，仅供内部参考

可在自有数据中心部署开放权重模型，并采购 GPU 服务器、推理软件和定制开发服务，相关费用包括硬件采购、软件授权、系统实施及后续运维。以 NVIDIA 为例，NIM 支持模型在企业数据中心或私有云中运行，并通过 NVIDIA AI Enterprise 提供商业授权和技术支持。

图32：英伟达企业级 AI 全栈平台架构



资料来源：Nvidia 官网

云端部署按模型调用量、专属算力和企业服务收费，前期投入门槛较低。Azure AI、Amazon Bedrock 和 Google Vertex AI 等平台均提供模型调用、微调、企业知识库和 Agent 开发服务，并支持私有网络接入。Azure 预置吞吐量还可按照小时或长期预留收费，企业无需自行采购硬件，可根据实际需求调整算力规模。我们认为，本地部署更适合数据敏感、调用需求稳定的大型客户，云端部署则在上线速度、前期投入和弹性扩展方面更具优势。

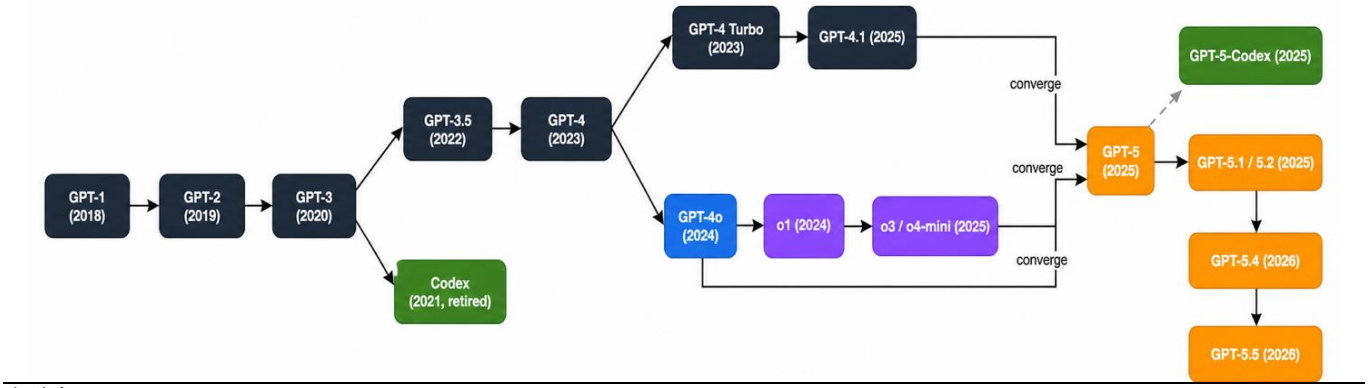
3. 头部大模型厂商：模型能力加速变现，商业路径逐步分化

3.1. OpenAI：多元业务拓宽收入边界，高投入推动平台化扩张

GPT 模型迭代由训练规模扩张逐步转向多模态、深度推理与智能体能力融合。GPT-1 至 GPT-4 主要提升语言理解与复杂任务处理能力，GPT-4o 进一步统一文本、图像和语音交互；此后，o1 及 o3、o4-mini 强化深度推理，GPT-4.1 提升长上下文和代码能力。2025 年起，GPT-5 将通用模型与推理模型统一调度，并延伸至 Codex 等专业分支，模型形态由对话助手逐步向复杂任务执行智能体演进。

下载日志已记录，仅供内部参考

图33：OpenAI 模型发展历程图



资料来源：Hidekazu Konishi

OpenAI 围绕底层模型建立了覆盖个人、企业和开发者的多层商业体系。OpenAI 通过个人订阅、企业席位、API 按量调用和企业定制服务实现收入，并开始尝试广告变现。不同产品共享底层模型和算力基础设施，ChatGPT 主要面向个人用户，API、Business 等则服务开发者和企业客户，形成由消费端向企业端延伸的产品结构。

表3：OpenAI 主要产品矩阵

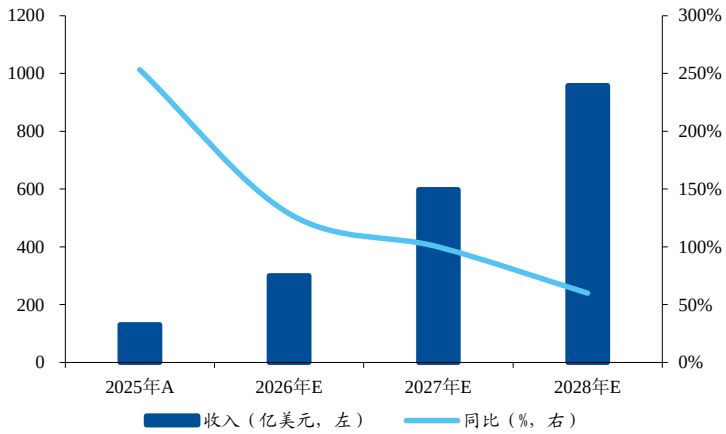
产品及业务	定位	核心功能	收费方式
ChatGPT	面向个人用户的通用 AI 助手	对话问答、联网搜索、文件分析、内容创作、图像与视频生成	订阅收费
ChatGPT Business/Enterprise	面向企业和团队的 AI 工作平台	团队协作、企业知识检索、数据分析等	Business 按席位订阅；Enterprise 定制收费
Codex	面向开发者和研发团队的编程智能体	代码生成、修改、测试及任务执行等	纳入 ChatGPT 订阅
OpenAI API Platform	面向开发者和企业的模型服务平台	提供各种大模型调用	按 Token 收费
Frontier 及企业服务	面向大型企业的智能体建设与管理平台	智能体开发、部署、权限管理及企业软件连接	定制收费
ChatGPT 广告	面向广告主的广告平台	根据用户对话需求匹配商品和服务	按展示等方式收费

数据来源：OpenAI 官网，NBAI.club，国泰海通证券研究

收入快速增长，预计未来三年仍将保持较高增速。据 Epoch AI，公司营业收入由 2024 年的 37 亿美元增至 2025 年的 130 亿美元，同比增长 253.2%。并且预计公司 2026 年收入约 299 亿、2027 年收入约 598 亿美元、2028 年达到约 957 亿美元。收入来源方面，据彭博社，ChatGPT 个人订阅占比 75%，仍是主要组成部分。企业订阅、智能体产品及 API 业务预计成为下一阶段增量，同时广告、硬件等新业务将逐步贡献收入。

下载日志已记录，仅供内部参考

图34：2025-2028 年 OpenAI 收入不断增长

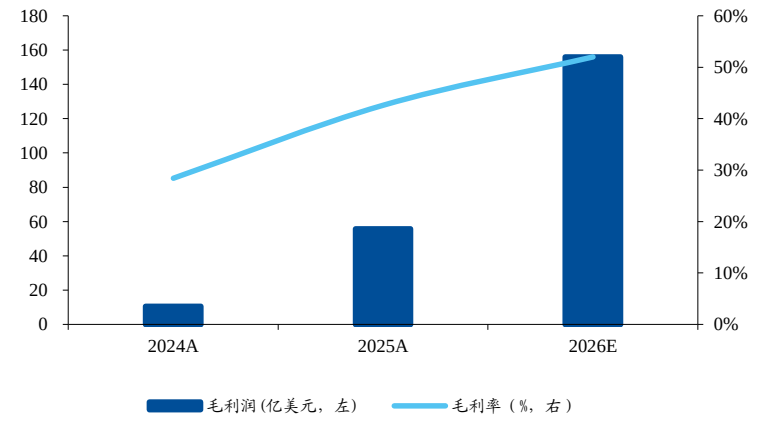


资料来源：Epoch AI，国泰海通证券研究

个人、企业及开发者调用规模同步扩大，商业化基础持续增强。据 OpenAI，截至 2026 年 2 月，ChatGPT 周活跃用户超过 9 亿，消费订阅用户超过 5000 万，付费商业用户超过 900 万。截至 2025 年 11 月，超过 100 万家组织通过企业产品或开发者平台付费，工作场景席位超过 700 万个，Enterprise 席位同比增长约 9 倍。截至 2026 年 4 月，OpenAI API 每分钟处理超过 150 亿个 Token，Codex 周活跃用户达到 300 万；已有超过 9000 家企业客户累计处理 Token 超过 100 亿，其中近 200 家超过 1 万亿。

收入增长快于营业成本，带动毛利率明显提升。据《金融时报》，2025 年营业成本由 26.5 亿美元增至 75.0 亿美元，同比增长 183.0%，低于收入增速；毛利润由 10.5 亿美元增至 55.7 亿美元，毛利率由 28.4%提升至 42.6%，同比增加 14.2 个百分点，据 The Information，2026 年一季度公司营业成本约 35 亿美元，对应毛利润约 22 亿美元，毛利率约 39%，并且预计实现 2026 年全年毛利润约 156 亿美元，毛利率上升至 52%，主要受益于模型运行效率提升和计算成本下降，带动付费产品的单位经济性改善。

图35：OpenAI 毛利率不断提升



资料来源：《金融时报》，The Information，国泰海通证券研究

公司仍将资源集中于模型研发和市场扩张。据《金融时报》，2025 年公司研发费用率由 2024 年的 211.1%降至 146.7%，虽然明显下降，但仍远高于收入规模，反映模型训练、产品研发继续保持高投入。销售及营销费用率由 30.0%升至 43.8%，主要受消费端获客、品牌推广及企业客户拓展加快影响；管理费用率则由 24.5%降至 12.0%，表明收入快速增长已对日常运营和后台管理支出形成一定摊薄。OpenAI 的费用结构仍以研发和市场扩张为主，规模效应目前主要体现在管理费用端。

高估值对收入增长和盈利改善提出更高要求。2026 年 3 月融资完成后，公

下载日志已记录，仅供内部参考

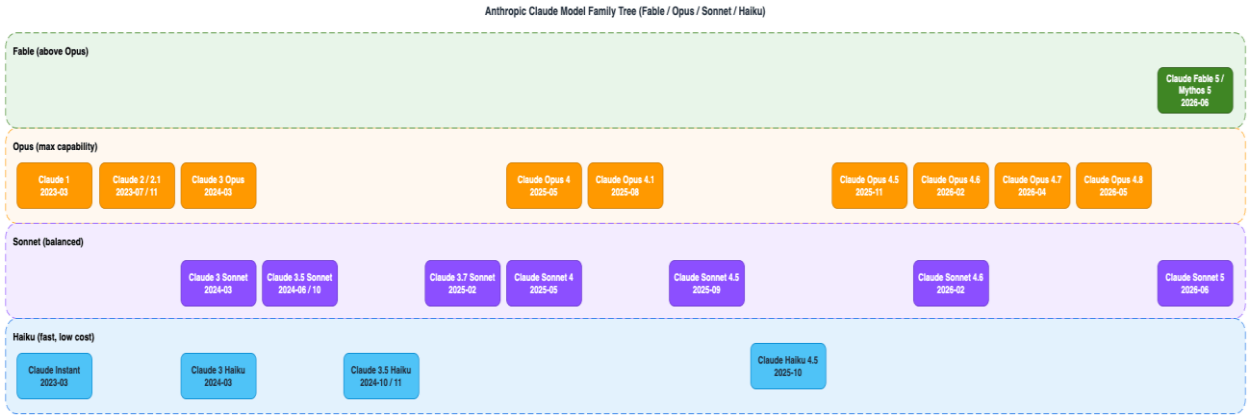
司投后估值达到 8520 亿美元，同期年化收入已超过 250 亿美元，对应 ARR 倍数约 34.1 倍。据 TheFinSense，2026 年 3 月上市 SaaS 公司 ARR 倍数中位数约 3.7 倍，同期 Databricks 估值约为其年化收入的 24.8 倍。OpenAI 估值明显高于公开软件行业平均水平，也高于多数高成长 AI 软件公司。

我们认为，OpenAI 正由纯大模型厂商向覆盖 AI 应用、企业服务、开发者平台和广告业务的综合 AI 公司转型。一方面，公司通过 ChatGPT、Codex 及智能体产品扩大应用入口，并以企业订阅、API 和按任务收费提升收入稳定性；另一方面，广告业务可依托 ChatGPT 的用户规模拓展消费端变现。多元业务布局有助于降低对模型调用和个人订阅的依赖，扩大公司生存空间，并推动其由技术提供商进一步升级为综合性 AI 平台。

3.2. Anthropic：聚焦模型与 Coding 优势，企业收入快速放量

Claude 模型的差异化产品定位兼顾能力、成本与响应速度。早期 Claude 1、Claude 2 系列主要提升通用理解和复杂任务处理能力；Claude 3 开始形成较清晰的 Opus、Sonnet、Haiku 三级体系，其中 Opus 侧重最高能力，Sonnet 平衡性能与成本，Haiku 强调速度和低成本。此后 Claude 3.5、3.7 及 4 系列持续强化代码、推理和复杂任务执行能力，并通过不同版本加快迭代节奏；最新 Claude 5 及 Fable 系列进一步向更高能力层级延伸。我们认为，清晰的模型分层使 Anthropic 既能以高端模型保持模型能力竞争力，也能以中轻量模型覆盖更广泛的个人、开发者和企业日常高频使用需求。

图36：Claude 模型发展历程图



资料来源：Hidekazu Konishi

Anthropic 更侧重企业与开发者，并以无广告和云平台分发形成差异化。据 Anthropic 表示，Claude 将保持无广告，其收入主要来自订阅、API 和企业服务；同时，Claude 已接入 Amazon Bedrock、Google Cloud Vertex AI 和 Microsoft Foundry，借助企业现有的云采购及管理体系扩大客户覆盖。

表4：Anthropic 主要产品矩阵

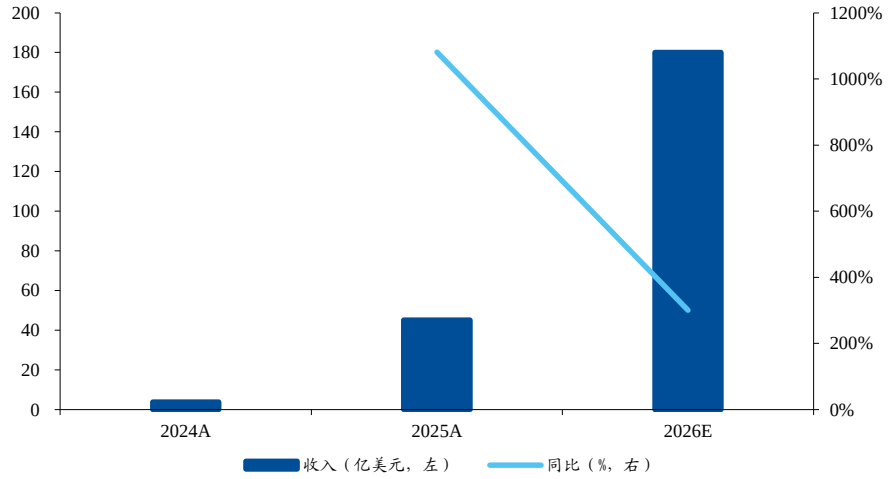
产品及业务	定位	核心功能	收费方式
Claude 个人版	面向个人的通用 AI 助手	对话问答、写作、深度研究等	订阅收费
Claude Team 及 Enterprise	面向组织和大型企业的 AI 工作平台	团队协作、企业知识接入、数据保护及内部工具连接	按席位订阅及定制收费
Claude Code	面向开发者和研发团队的编程智能体	理解代码仓库、生成及修改代码、运行测试、调用终端工具	个人及团队订阅，并按 API 调用量收费
Claude Cowork	面向知识工作者的任务执行工具	读取和修改文件，完成跨工具办公任务	纳入个人及企业订阅，企业版本定制收费
Claude API	面向开发者和企业的模型服务平台	提供各种大模型调用	按 Token 收费
云平台分发	通过主流云平台触达企业客户	通过 Amazon Bedrock、Google Cloud 等平台提供 Claude 模型	由云平台按照模型调用量收费

数据来源：Anthropic 官网，国泰海通证券研究

下载日志已记录，仅供内部参考

企业客户及 Claude Code 推动收入高速增长，Anthropic 商业化进入加速阶段。据 Epoch AI，公司营业收入由 2024 年的 3.81 亿美元增至 2025 年的 45 亿美元，同比增长约 10.8 倍。据 The Information，2026 年一季度公司收入增至约 48 亿美元，单季度收入已超过 2025 年全年水平，并且 2026 年全年收入预计最高约 180 亿美元。收入增长主要来自企业客户扩大模型及 API 使用，以及 Claude Code 在开发团队中的快速渗透，据 Anthropic，截至 2025 年 9 月，公司企业客户超过 30 万家；截至 2026 年 2 月，年化支出超过 100 万美元的客户已超过 500 家。Claude Code 成为重要增量，其年化收入由 2025 年 11 月的 10 亿美元增至 2026 年 2 月的 25 亿美元以上，约占同期公司总年化收入的 18%。

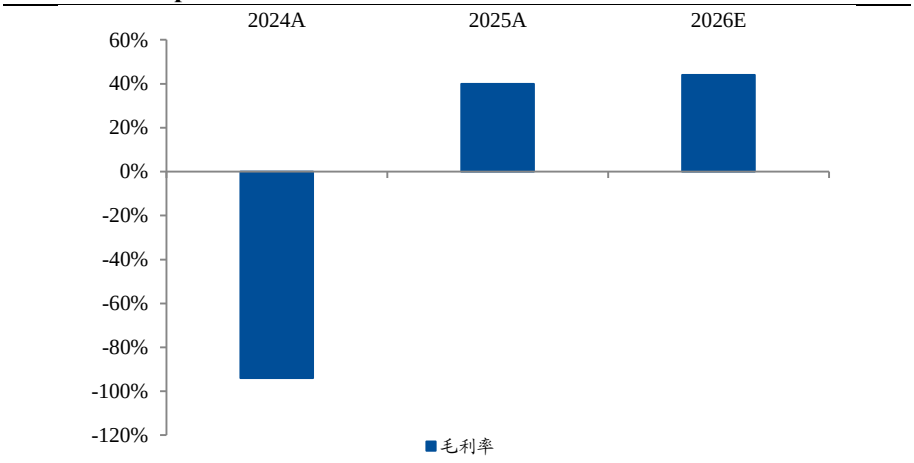
图37：Anthropic 2024-2026 年收入快速增长



资料来源：Epoch AI，The Information，国泰海通证券研究

收入快速增长推动毛利率由负转正，盈利能力持续改善。据 The Information，公司毛利率由 2024 年的-94%提升至 2025 年的约 40%。据 Morningstar，Anthropic 2026 年毛利率有望进一步提升至约 44%。

图38：Anthropic 毛利率提升



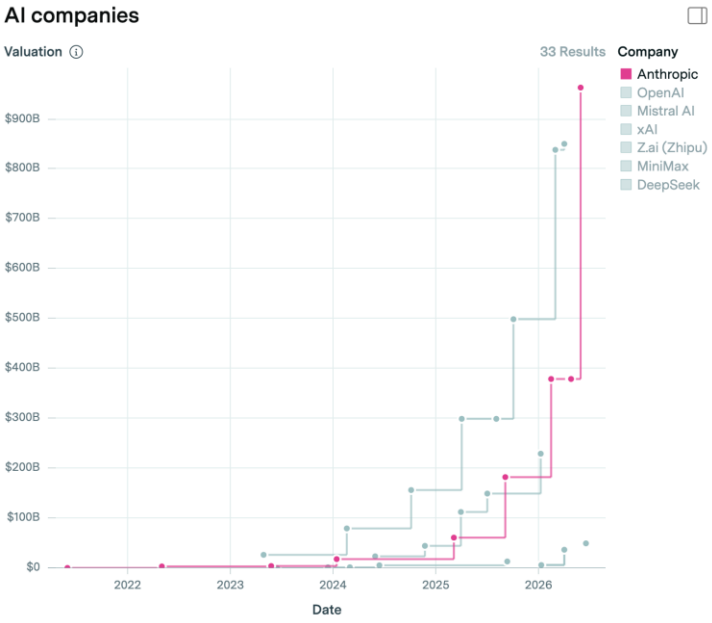
资料来源：The Information，Morningstar，国泰海通证券研究

算力和人员投入仍是 Anthropic 最主要的支出。据 Epoch AI，2025 年公司训练算力和推理算力支出分别约 41 亿美元和 27 亿美元，人员及其他运营成本约 29 亿美元，三项合计约 97 亿美元，相当于同期收入的约 216%。其中，训练成本占收入的比例由 2024 年的约 394%降至 2025 年的 91%，表明收入快速增长已开始降低高额研发投入带来的经营压力。

下载日志已记录，仅供内部参考

企业收入快速增长，推动融资规模持续扩大、估值持续攀升。据 Epoch AI，2025 年 9 月，公司完成 130 亿美元融资，估值升至 1830 亿美元；2026 年 2 月，公司融资 300 亿美元，估值达到 3800 亿美元。2026 年 5 月，公司完成 650 亿美元融资，投后估值跃升至 9650 亿美元，此次估值大幅提升，主要由于公司年化收入于 5 月突破 470 亿美元，企业客户采用持续扩大。

图39：Anthropic 估值持续攀升



资料来源：Epoch AI

我们认为，企业级市场是大模型竞争的关键领域。如果大模型不跳脱出软件行业，企业客户仍是付费能力最强的客户群体。Anthropic 一方面保持基座模型智力水平全球领先，另一方面以安全、透明和可控为差异化定位，降低金融、法律、政务等高合规行业的采购与部署门槛。在产品侧，公司长期聚焦企业客户与编码智能体，Claude Code 已建立企业级 Coding Agent 的先发优势，并可进一步向更多知识工作延伸。在分发侧，公司较早接入微软智能云、AWS Bedrock 和 Google Vertex AI，覆盖主流企业云平台。模型能力、企业信任、产品优势与多云渠道共同构成其企业级市场壁垒。

3.3. MiniMax：全模态与全球化双轮驱动，消费应用与 API 协同增长

早期由文本大模型起步，逐步拓展至语音、视频、音乐和推理模型，形成较完整的多模态布局。公司早期围绕 abab 系列提升通用语言能力，并于 2024 年引入 MoE 架构；随后通过 Speech、Hailuo 和 Music 系列进入语音生成、视频生成及音乐合成等方向。2025 年以来，MiniMax 进一步推出开源文本模型和 M1、M2 推理模型，模型能力由内容生成延伸至复杂推理。整体来看，MiniMax 已由单一文本模型厂商发展为覆盖语言、多模态生成和推理能力的综合大模型平台。

下载日志已记录，仅供内部参考

图40：MiniMax 模型发展历程图



资料来源：MiniMax 招股书，国泰海通证券研究

MiniMax 形成消费端应用与企业端模型服务协同发展的业务结构。个人端覆盖 AI 陪伴、视频生成、语音音乐、智能体等产品，主要通过订阅及应用内付费变现；企业端通过开放平台提供各种大模型 API 及其他企业服务，按照 Token 调用量或定制收费。

表5：MiniMax 主要产品矩阵

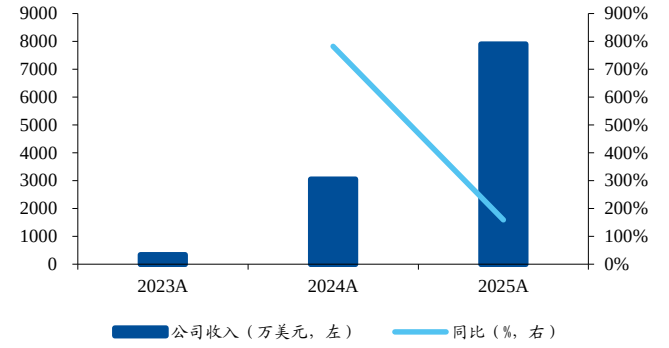
产品及业务	定位	核心功能	收费方式
星野/Talkie	面向个人用户的 AI 角色陪伴产品	AI 角色对话、角色创建、互动剧情及虚拟社交	订阅及应用内付费
海螺 AI	面向创作者的视频生成产品	文生视频、图生视频及视频编辑	订阅及额度收费
MiniMax Agent	面向个人的通用智能体	多步骤任务执行	订阅收费
MiniMax Audio&Music	面向创作者和企业的语音生成产品	文本转语音、声音克隆、多语言配音及音频生成	按订阅及 API 调用收费
MiniMax Code	面向开发者的编程智能体	代码生成、代码库理解、修改测试及软件开发任务执行	订阅、开发套餐及 API 收费
MiniMax 开放平台及其他 AI 企业服务	面向开发者和企业的全模态模型平台	提供各种大模型调用并为企业提供模型授权部署	按 Token 或调用量收费，企业定制收费

数据来源：MiniMax 官网，MiniMax 招股书，国泰海通证券研究

AI 原生产品支撑收入增长，开放平台及海外业务贡献持续提升。据 MiniMax 招股书及 2025 年度业绩公告，公司收入由 2023 年的 346.0 万美元增至 2024 年的 3052.3 万美元，2025 年进一步增至 7903.8 万美元，2024 年和 2025 年分别同比增长 782.2%和 158.9%。其中，AI 原生产品收入由 75.8 万美元增至 2180.5 万美元和 5307.5 万美元，2025 年占比 67.2%；开放平台及其他 AI 企业服务收入由 270.2 万美元增至 871.8 万美元和 2596.3 万美元，2025 年同比增长 197.8%，占比提升至 32.8%。此外，2025 年国际市场收入占比达到 73.0%，从前三季度地域结构看，新加坡和美国分别贡献总收入的 24.3%和 20.4%，其他海外地区贡献 28.4%，已覆盖亚太、美洲和欧洲等多个市场。

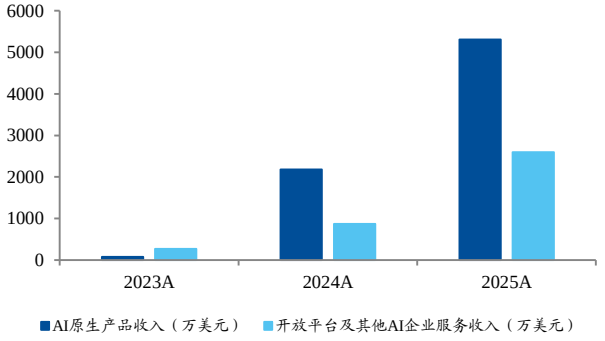
下载日志已记录，仅供内部参考

图41：2023-2025 年公司收入快速提升



资料来源：MiniMax 招股书，MiniMax2025 年度业绩公告，国泰海通证券研究

图42：2025 年 AI 原生产品贡献主要收入，开放平台及企业服务占比提升



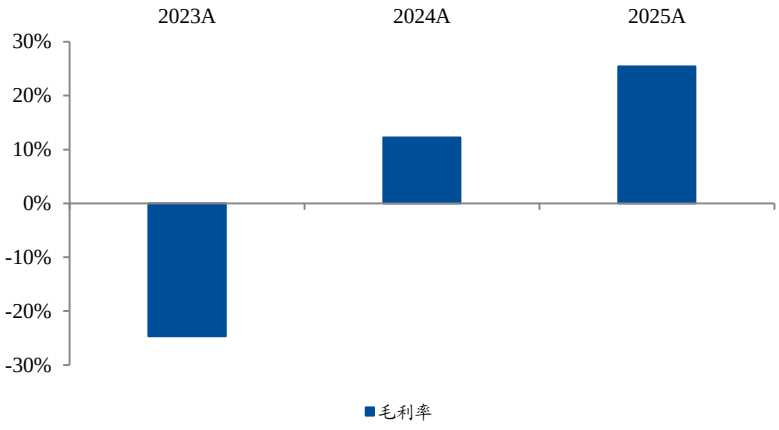
资料来源：MiniMax 招股书，MiniMax2025 年度业绩公告，国泰海通证券研究

个人用户、付费用户和开发者规模同步扩大。据 MiniMax 招股书及 2025 年度业绩公告，截至 2025 年末，公司累计服务超过 2.36 亿名个人用户及 21.4 万家企业客户和开发者。2025 年前三季度，AI 原生产品平均月活跃用户达到 2760 万，付费用户由上年同期的 48.9 万增至 177.2 万，单名付费用户平均支出由 7 美元升至 15 美元；开放平台付费用户由约 400 名增至 2500 名。截至 2026 年 6 月末，公司全球个人用户进一步增至约 3 亿，企业客户和开发者超过 100 万，半年增长约 4 倍。

开放平台用户与模型调用量快速增长，编程需求表现突出。据 MiniMax 2025 年度业绩公告，2026 年 2 月，M2 系列模型平均单日 Token 消耗量增至 2025 年 12 月的 6 倍以上，其中 Coding Plan 相关 Token 消耗量增长超过 10 倍；2025 年四季度，M2 还成为首个在 OpenRouter 实现单日 Token 消耗超过 500 亿的中国模型。我们认为，Token 消耗和编程调用快速增长，表明公司增长正由消费应用进一步延伸至开发者和生产力工具。

模型及系统效率提升推动毛利率连续改善。据 MiniMax 招股书及 2025 年度业绩公告，毛利润由 2023 年的-85.4 万美元增至 2024 年的 373.8 万美元和 2025 年的 2007.9 万美元，毛利率由-24.7%提升至 12.2%和 25.4%。其中，AI 原生产品毛利率由 2023 年的-380.2%提升至 2024 年的-8.1%，反映用户规模扩大和推理效率提升已显著降低单位服务成本，2025 年整体毛利率进一步提升，主要受模型效率、系统效率及基础设施配置优化带动。

图43：2023-2025 毛利率持续改善



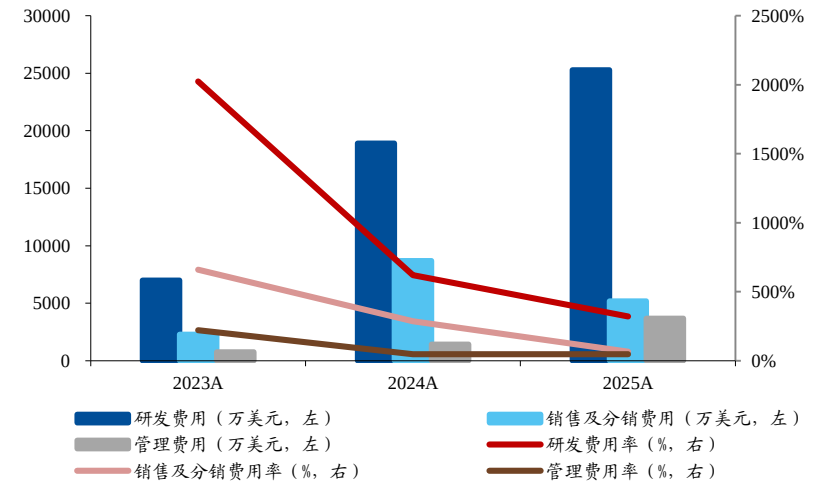
资料来源：MiniMax 招股书，MiniMax 2025 年度业绩公告，国泰海通证券研究

研发仍是最主要的投入方向，但研发及销售费用率已明显下降。据 MiniMax 招股书及 2025 年度业绩公告，研发费用由 2023 年的 7000.2 万美元增至 2024 年的 1.89 亿美元和 2025 年的 2.53 亿美元，研发费用率分别约为 2023.1%、619.1%和 319.8%。同期销售及分销费用分别为 2282.7 万美元、

下载日志已记录，仅供内部参考

8699.5 万美元和 5189.6 万美元，对应费用率由 659.7%降至 285.0%和 65.7%，2025 年销售费用下降主要由于 AI 原生产品更多依靠自然增长和口碑传播。管理费用分别为 761.5 万美元、1438.4 万美元和 3681.3 万美元，对应费用率由 220.1%降至 47.1%和 46.6%，2025 年绝对额增长主要受人员扩张及上市费用影响。

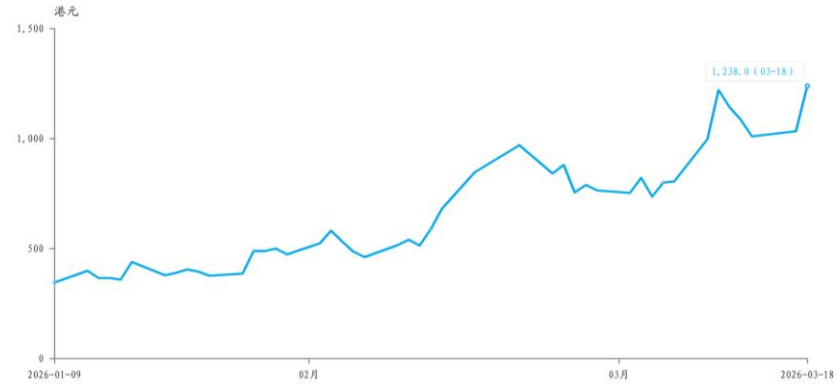
图44：研发、销售、管理费用率均显著下降



资料来源：MiniMax 招股书，MiniMax 2025 年度业绩公告，国泰海通证券研究

模型持续升级与 Agent 业务预期升温，驱动 MiniMax 股价上行。据中国证券网，2026 年 1 月 9 日上市首日股价较发行价翻倍，反映市场对港股大模型稀缺标的及其全球化消费应用的较高预期。2026 年 2 月中旬发布 M2.5 后，模型在编程、工具调用和办公任务中的能力提升，带动市场上调其开发者平台及 Agent 业务预期。2026 年 3 月 18 日 M2.7 进一步强化复杂 Agent 任务和模型自我迭代能力，成为股价再次上行的重要催化。

图45：2026.01.09-2026.03.18 MiniMax 股价上升



资料来源：ifinD，国泰海通证券研究

3.4. 月之暗面：模型能力推动商业化，订阅与 API 收入提速

以长上下文作为切入口，月之暗面逐步将 Kimi 扩展为覆盖多模态、推理、代码和 Agent 能力的模型体系。2023、2024 年公司围绕 Kimi Chat 和 Moonshot 系列强化长文本理解，并通过 API 服务开发者；2025 年以来，模型布局进一步延伸至视觉、语音和推理方向。Kimi K2 系列采用 MoE 架构，重点提升代码生成、工具调用和复杂任务执行能力。Kimi K3，总参数规模达到 2.8 万亿，是首个进入近 3 万亿参数级别的开源模型。2026 年后，公司继续强化原生多模态、Agent 协作和长上下文处理，Kimi 由长文本对话工具逐步发展为综合模型平台。

下载日志已记录，仅供内部参考

图46：月之暗面模型发展历程图



资料来源：月之暗面官网，界面新闻，arXiv，GitHub，国泰海通证券研究

月之暗面围绕 Kimi 构建个人应用、Agent 和 API 服务。月之暗面围绕 Kimi 持续完善模型、产品与开发者生态，并将使用范围由通用问答扩展至编程、知识工作和复杂任务执行。最新模型 Kimi K3 进一步强化各项能力，为订阅和 API 业务增长提供支撑。

表6：月之暗面主要产品矩阵

产品及业务	定位	核心功能	收费方式
Kimi	面向个人的通用 AI 助手	对话、搜索、及任务执行等	订阅收费
Kimi Researcher 等 Agent 产品	面向研究、学习等的深度研究工具	自主完成从任务规划到结果交付的全流程	纳入 Kimi 订阅
Kimi Code	面向开发者的编程智能体	读取代码库、执行终端命令并完成多文件开发任务	纳入 Kimi 订阅；通过 API 接入时按 Tokens 收费
Kimi API 开放平台	面向开发者和企业的模型服务平台	提供各种大模型调用	按输入和输出 Token 收费
企业服务	面向企业客户的模型接入及解决方案	企业办公插件集成和定制化智能体等	企业定制收费

数据来源：月之暗面官网，国泰海通证券研究

新模型订阅和 API 需求增长推动收入快速提升。据财联社，2025 年 9-11 月 Kimi 海内外付费用户数平均环比增长超过 170%，受 K2 Thinking 等模型带动，海外 API 收入较此前增长约 4 倍。2026 年公司商业化继续扩张，据每经网，K2.5 模型更新后，Kimi ARR 在 2026 年 3 月初突破 1 亿美元，在 4 月继续增长至超 2 亿美元，付费订阅和 API 调用都在加速。据新浪新闻，截至 2026 年 6 月，Kimi 产品已进入 200 多个国家和地区，互联网、金融、制造、教育、医疗等行业已成为重要企业客户来源。

经营策略由高额投流获客转向研发和高价值用户。据每经网，2024 年 Kimi 营销投放超过 7 亿元，其中第四季度约 5.3 亿元；2025 年第一季度投放降至约 1.5 亿元，第四季度已不足 10 万元。受投放收缩影响，Kimi App 月活跃用户由 2025 年第一季度的 2165.3 万降至第四季度的 902.7 万，但公司将资源重新集中于模型训练、代码和 Agent 能力，并在产品及商业化上聚焦 Agent。

由消费应用向开发者和企业客户倾斜。公司与阿里、腾讯等产业方形成资本和渠道协同，据财联社，两者参与公司多轮融资，Kimi 模型同时进入阿里云模型服务和腾讯云相关产品体系，借助头部云平台触达更多企业与开发者。

融资与现金储备为持续研发提供支持。据财联社，公司于 2025 年末完成 5 亿美元 C 轮融资，投后估值约 43 亿美元，现金储备超过 100 亿元。据财联社，公司 2026 年 5 月已完成新一轮约 20 亿美元融资，投后估值超 200 亿美元。据证券时报，月之暗面半年内累计融资超 39 亿美元，总融资额超 376 亿元人民币，成为国内大模型创业公司中累计融资最多的企业。

我们认为，Kimi 具备较大的发展潜力。一方面，公司是国内较早探索万亿参数开源模型的厂商之一，团队持续追求模型智能上限，更大规模模型的推

下载日志已记录，仅供内部参考

出有望进一步提升综合能力。另一方面，公司较强的基座模型与算法和工程能力有助于降低推理及任务执行成本，在 Coding 和 Agentic 市场形成竞争优势。此外，公司早期愿意投入资源争夺 C 端用户，并与阿里、腾讯等产业方合作，体现出较强的商业敏感度和执行能力，有望在后续大模型商业化竞争中进一步释放价值。

3.5. 阶跃星辰：多模态切入智能终端，端侧合作打开增长空间

以语言与多模态模型为起点，并进一步布局端侧模型。2024 年，阶跃星辰围绕 Step 系列完善基础语言和视觉能力，并开始拓展语音与视频生成；2025 年，多模态布局进一步加快，陆续推出语音、视频及图像模型，同时通过 Step 3 等模型强化复杂推理能力。2026 年以来，公司一方面通过 Flash 系列提升推理与 Agent 能力，另一方面推出轻量化视觉模型、图像编辑模型和端侧模型。整体来看，阶跃星辰已形成覆盖基础、多模态、智能体和端侧的较完整模型体系。

图47：阶跃星辰发展历程图



资料来源：阶跃星辰官网，财联社、IT 之家，GitHub，arXiv，国泰海通证券研究

阶跃星辰更侧重将模型能力嵌入智能终端和企业系统。公司以自研 Step 系列模型为基础，通过开放平台连接开发者，并围绕手机、汽车及具身智能拓展终端合作，目前已与 OPPO、荣耀、中兴、吉利及智元机器人等企业开展合作。2026 年 7 月，公司进一步发布大模型原生终端品牌 STEPX、智能体操作系统 Step AOS 及首款智能体手机 STEPX Neo，由外部赋能终端延伸至自有软硬件产品。随着终端覆盖范围扩大，预装授权、模型调用和生态服务有望共同构成收入来源。

表7：阶跃星辰主要产品矩阵

产品及业务	定位	核心功能	收费方式
阶跃 AI	面向个人用户的通用 AI 产品	对话问答、图像理解及内容生成等	订阅收费
Step 系列模型 API	面向开发者和企业的模型服务平台	提供各种大模型调用	按 Token、调用量或生成内容收费
企业服务	面向金融、制造、通信等行业客户	企业知识检索、流程自动化及智能体开发等	企业定制收费

数据来源：阶跃星辰官网，国泰海通证券研究

企业和终端需求推动收入快速增长。据《财经》，阶跃星辰 2025 年收入接近 5 亿元，2026 年预计增至约 12 亿元，收入规模有望扩大至上年的约 2.4 倍。公司当前形成 AI 服务、AI 设备和 AI 基础设施三类业务。

商业化重点由消费端应用转向智能终端和开发者服务。据 36 氪，公司通过将模型接入手机、汽车等设备，并随设备出货及实际使用获得收入，该业务已成为核心收入来源。目前阶跃模型覆盖国内约 60%的头部手机品牌，累计装机量超过 4200 万台。同期，开放平台 API 调用量同比增长近 20 倍，活跃开发者数量增长约 5 倍。相较单纯按 Token 收费或承接企业定制项目，这一路径既拓宽了模型分发入口，也有助于降低消费端获客投入。

研发仍处于高投入阶段。据财中社，2025 年公司算力及算法研发投入约 12

下载日志已记录，仅供内部参考

亿元，相当于同期收入的约 240%。据新华财经，2026 年 1 月，阶跃星辰完成超 50 亿元 B+轮融资，募集资金主要用于基础模型研发及“AI+终端”业务拓展。并且据虎嗅，截至 2025 年 5 月，阶跃星辰员工超过 400 人，其中约 80%为技术研发人员。

持续融资为模型研发和终端扩张提供资金支持。据财联社，阶跃星辰于 2024 年末完成数亿美元 B 轮融资。据财新网，2026 年 1 月又完成超过 50 亿元 B+轮融资。持续融资和产业资本参与为公司保持模型研发投入，并扩大手机、汽车等终端合作提供资金支持。

我们认为，阶跃星辰的差异化优势在于较早布局多模态，并进一步将能力重心延伸至 Agent 及智能终端。相较 MiniMax 和 DeepSeek，阶跃星辰并未选择单纯依靠 API 服务企业客户，也不准备复制 Anthropic 早期以 Coding 为核心切入企业市场的路径，而是将手机、汽车等高频终端作为模型进入用户日常使用环节的核心入口。模型进入具体物理载体后，可以形成数据、产品和服务的完整循环。公司希望通过深入终端，自主参与应用、用户与数据的形成，逐步由模型供应商转向覆盖模型、系统和终端服务的智能终端平台。

4. 应用层：多模态拓展产品形态，海外 AI 应用加快商业化

4.1. AI 应用分类

海外 AI 应用已形成覆盖信息获取、内容生产与垂直用途的多元产品体系。结合 AICPB 产品榜及 AIMultiple 对生成式 AI 应用的分类，我们将海外 AI 应用归纳为对话与搜索、图像、音视频、营销、教育学习及游戏六大类。其中，对话与搜索类包括聊天机器人和 AI 搜索；图像类覆盖图片生成与编辑；音视频类包括音乐生成及视频生成、编辑；营销和教育学习类分别对应 AI 营销与 AI 教育；游戏类则进一步包括游戏开发生成与智能 NPC

表8：AI 应用分类

应用大类	细分类目
对话与搜索类	聊天机器人
	AI 搜索
图像类	图片生成
	图片编辑
音视频类	音乐生成
	视频生成编辑
营销类	AI 营销
教育学习类	AI 教育
游戏类	游戏开发生成
	智能 NPC

数据来源：AICPB，AIMultiple《Top 125 Generative AI Applications》，国泰海通证券研究

4.2. 对话与搜索类

4.2.1. 聊天机器人

聊天机器人正由问答工具向综合任务助手升级。该类产品以多轮对话为核心，覆盖知识问答、内容生成、代码开发、文件分析及多模态理解，并通过个人订阅、企业席位和 API 调用收费。随着多模态能力以及联网检索、长期记忆和工具调用能力增强，产品可完成更多复杂任务；长期使用还会加深对用户偏好和常用需求的了解，提高使用黏性和迁移成本。

表9：聊天机器人应用

名称	2026 年 6 月访问量	简介	商业模式
ChatGPT	5.56B	OpenAI 于 2022 年 11 月推出。最新模型为 GPT-5.5 系列，擅长智能体编程、计算机操作、知识工作、科学研究及复杂任务执行。	个人及企业订阅，API 按量计费，并探索广告变现。
Gemini	2.95B	Google 于 2023 年 3 月以 Bard 之名推出，2024 年 2 月更名	Workspace 企业席位及

下载日志已记录，仅供内部参考

		Gemini。最新模型家族为 Gemini 3.5，首发 3.5 Flash，擅长复杂智能体工作流、长程任务、推理与编程。	Cloud/API 调用收费。
Claude	965.59M	Anthropic 于 2023 年 3 月推出。最新默认模型为 Claude Sonnet 5，旗舰模型为 Fable 5，强化规划、工具调用、编程、推理、知识工作及多步任务执行。	个人及企业订阅订阅、API 按量收费。
Grok	220.78M	xAI 于 2023 年 11 月推出。最新模型为 Grok 4.20，具备高级推理、多智能体协作、实时信息检索、图像生成和视频理解能力。	SuperGrok 与 X 会员订阅
Copilot	121.12M	Microsoft 于 2023 年 2 月先以 Bing Chat 推出，。Microsoft 365 Copilot 最新接入 GPT-5.5 Instant，支持日常办公、图像分析及 STEM 任务	Microsoft 365 生态订阅

数据来源：AI 产品榜，OpenAI 官网，TechCrunch，The Guardian，SaaS CRM Review，微软官网，PricingBrief，Anthropic 官网，国泰海通证券研究

4.2.2. AI 搜索

从链式搜索转为直接给出答案，并逐步提高检索效率。该类产品通过联网检索、多来源整合直接生成答案，并逐步支持多模态输入。根据 AI 产品榜，2026 年 6 月 New Bing 访问量约 36.8 亿次，依托传统搜索入口显著领先；Perplexity、Consensus 和 Kagi Search 则分别聚焦深度研究、学术检索和隐私保护。AI 搜索将在长期使用中进一步了解用户关注领域、常用信息源和表达偏好，提高检索效率与用户黏性。

表10：AI 搜索应用

名称	2026 年 6 月访问量	简介	商业模式
New Bing	3.68B	Microsoft 于 2023 年 2 月 7 日推出 AI 驱动的 New Bing 预览，随后整合至 Microsoft Copilot。未为 Bing 公开单一固定底层模型；Copilot 产品线最新公开接入 GPT-5.5 Instant，可联网问答、信息归纳、内容生成和图像分析。	企业端通过 Copilot 订阅收费
Perplexity AI	120.79M	Perplexity 于 2022 年 12 月推出 AI 答案引擎，采用自研与第三方模型结合的多模型架构。默认 Best 模式可自动选择模型，Pro Search 可使用 Sonar、GPT 等。可实时搜索网络并生成带来源引用的综合答案。	个人及企业订阅，API 收费。
Consensus	5.16M	Consensus 于 2022 年 9 月公开推出，定位于科学研究搜索。未公布单一固定底模，而是结合商业大模型与微调开源模型；最新功能为 2026 年 5 月推出的 Research Agent，可规划检索步骤并调用多种研究工具。	免费基础版、个人及团队订阅
Kagi Search	5.06M	Kagi 于 2022 年 6 月 1 日推出搜索引擎公开测试版。Kagi Search 不绑定单一生成模型，可选择 GPT-5.5、Claude Fable 5 等模型，并可结合 Kagi Search 执行联网研究。	以无广告付费订阅为主

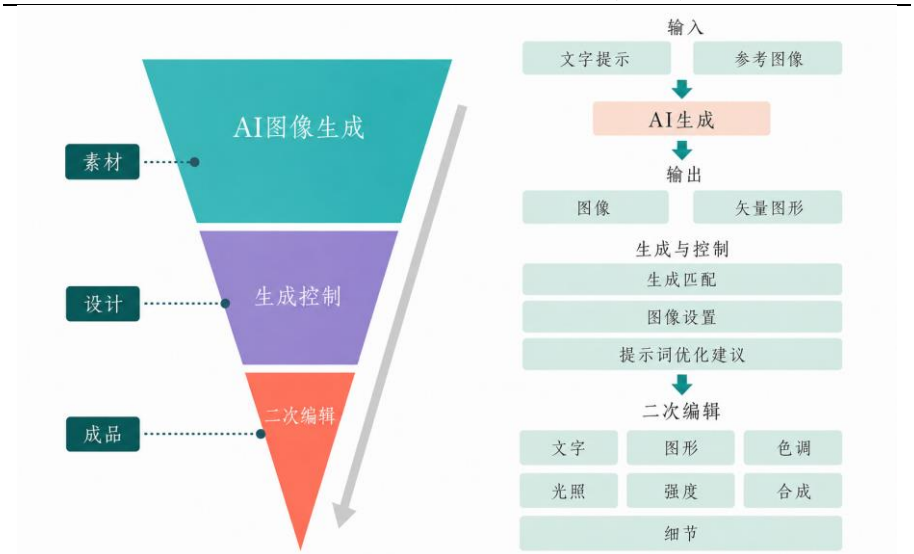
数据来源：AI 产品榜，微软官网，IT History，TechCrunch，Consensus 官网，kagi 官网，国泰海通证券研究

4.3. 图像类

图像类 AI 设计层级包括 AI 图像生成、生成控制与二次编辑。AI 图像生成可根据文字提示或参考图像产出图像与矢量图形，生成控制通过生成匹配、图像设置和提示词优化建议，提高生成结果与设计意图的契合度；二次编辑则围绕文字、图形、色调、光照、和细节等进行完善。生成环节提高产量，编辑环节保证质量。

下载日志已记录，仅供内部参考

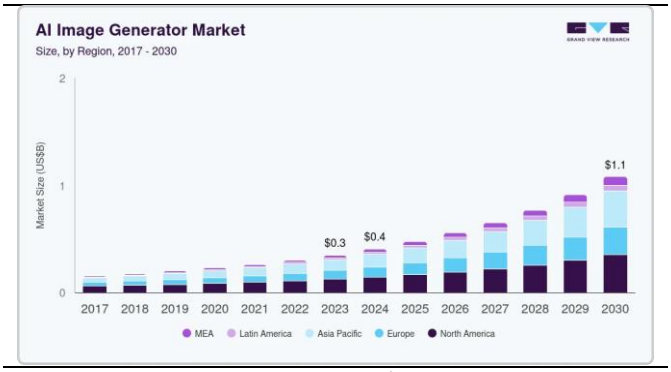
图48：AI 设计层级包括原型生成、辅助设计与精准编辑



资料来源：新智元，国泰海通证券研究

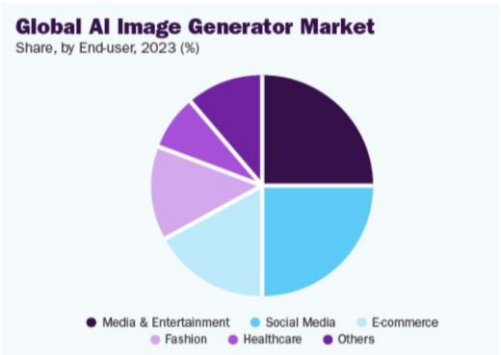
五大创意领域提供增长驱动力，AIGC 图像市场迅速增长。全球 AI 图像生成市场规模由 2023 年的 3.50 亿美元增长至 2030 年的约 10.8 亿美元，2024-2030 年复合增速达 17.7%。从区域结构看，北美仍为当前最大市场，2023 年收入占比达到 37.5%；亚太地区增长速度领先，预计未来将成为 AI 图像生成市场的重要增量来源。全球范围内，媒体和娱乐领域市场占比最高。2023 年，全球 AIGC 图像应用市场规模约 3.496 亿美元，有 5 个主要的应用场景，其中媒体和娱乐领域占比最高，其次为社交媒体、电商、时尚和大健康等多个领域。

图49：2017-2030 世界 AIGC 生图工具市场规模（百万美元）



资料来源：Grand View Research，国泰海通证券研究

图50：2023 年全球 AIGC 生图工具细分应用场景占比

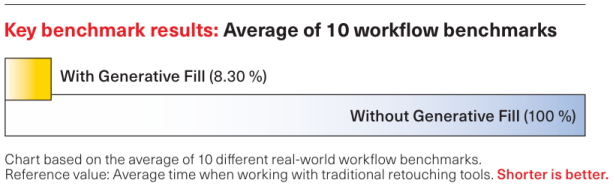


资料来源：Grand View Research，国泰海通证券研究

提高广告素材供给、修改和测试效率。广告主需要围绕不同人群、地区和渠道制作大量版本，传统拍摄和设计成本限制了素材更新速度。据 Amazon Ads，Hisense Mexico 使用 AI 生图后，单项产品可测试的创意版本由 1—2 个增至最多 7 个，制作时间由 4 周缩短至 1 小时以内；同期广告点击率提升 50%、广告投入回报率提升 48%，广告销售成本下降 32%。据 Pfeiffer Report，使用生成式填充完成图片编辑的平均耗时仅为传统工具的 8.3%，对应约 12 倍的效率提升；据 Adobe Express 企业在人员规模基本不变的情况下，内容版本产量可提升至 100 倍以上，营销、销售和人力资源团队制作社交内容的速度可提升至 5 倍。随着生成能力逐步成熟，模型迭代重点将更多转向按要求精确修改，局部控制、风格统一和多人协作等，这些能力将决定生图工具在广告制作中的使用深度。

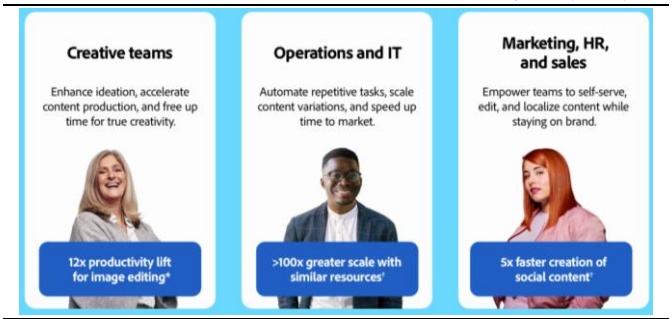
下载日志已记录，仅供内部参考

图51：AI 将图像编辑耗时降至传统流程的 8.3%



资料来源：Pfeiffer Report

图52：生成式 AI 在创意、运营、营销显著提升效率



资料来源：Adobe Express

缩短游戏美术制作周期，并提高内容更新频率。据 Scenari，德国游戏公司 InnoGames 将 AI 用于背景、加载页面、角色和图标制作后，美术团队生产效率提升至原来的 2 倍，新内容制作时间最多缩短 50%。据 Layer，美国游戏公司 Live Play Mobile 则将 AI 用于《Live Play Bingo》的角色和主题背景制作，一套收藏品素材的制作周期由接近 4 天缩短至 1 天，每周节省约 24 小时，美术生产效率提升 300%。我们认为，随着 AI 图片工具进入角色、场景和运营素材的修改环节，游戏公司可以减少重复修图和返工，缩短美术排期并扩大内容供给。

图片编辑将比图片单次生成应用更快进入企业的日常工作内容中。企业需要的并非单单生成一张高质量效果图，而是低成本完成大量修改与适配。未来产品差异将更多体现在局部修改精度等。能够嵌入专业设计软件和企业素材库，并同时服务设计师与非设计人员的厂商，更容易扩大企业订阅量和使用频次。

4.4. 音视频类

4.4.1. 音乐生成

完整歌曲生成率先形成规模，音乐辅助工具围绕专业制作需求继续拓展。该类产品可根据文本、歌词和曲风要求生成人声、旋律及编曲，降低歌曲创作和音频制作门槛。行业用户需求主要集中于快速生成完整歌曲，变声、音轨分离和伴奏制作等工具则面向社交娱乐及专业创作需求形成补充。据 AI 产品榜，2026 年 6 月 Suno 访问量约 7655 万次，断档领先 Voicemod 和 Moises 的 417 万次和 405 万次。

Suno 凭借完整歌曲生成建立明显领先，并通过分层订阅实现规模化收入。用户输入主题、歌词和曲风即可生成包含人声、旋律与编曲的完整歌曲，平台还支持旋律续写、自定义声音和 Suno Studio 多轨制作等，产品由娱乐型生成工具逐步延伸至完整音乐制作平台。Suno 采用免费额度导流，Pro 和 Premier 月费分别为 10 美元和 30 美元，付费套餐提供更多生成额度、进阶模型、优先队列。据 Music Business Worldwide，截至 2026 年 2 月，Suno 累计用户超过 1 亿、付费用户达到 200 万，ARR 达到 3 亿美元。据 Goodwin，2026 年 6 月，公司又以 54 亿美元估值完成超过 4 亿美元融资，用户付费和收入规模均显著领先其他音乐生成产品。

表11：音乐生成 AI 应用

名称	2026 年 6 月访问量	简介	商业模式
Suno	76.55M	通过文本提示或歌词生成包含人声、旋律和编曲的完整歌曲，并提供歌曲编辑、音轨分离及音频上传等功能，覆盖个人娱乐与专业音乐创作。	免费额度导流，Pro 及 Premier 分档订阅，付费套餐提供更多生成额度、优先生成和商业使用权。
Voicemod	4.17M	Text to Song 可根据用户输入的歌词，结合预设曲风和 AI 歌手快速生成歌曲片段，主要面向社交分享和娱乐创作；平台同时提供实时变声和音效板。	基础功能免费导流，主要通过 Voicemod PRO 订阅解锁完整声音库、自定义声音及高级功能。
Moises	4.05M	定位于音乐创作辅助与音频处理，通过 AI 实现人声及乐器分离、伴奏制作、变	免费版提供有限处理次数，Premium 及 Pro 订阅提升上传额度和专业功能

下载日志已记录，仅供内部参考

		调变速、母带处理和歌声转换，更偏向音乐编辑而非纯文本生成歌曲。	

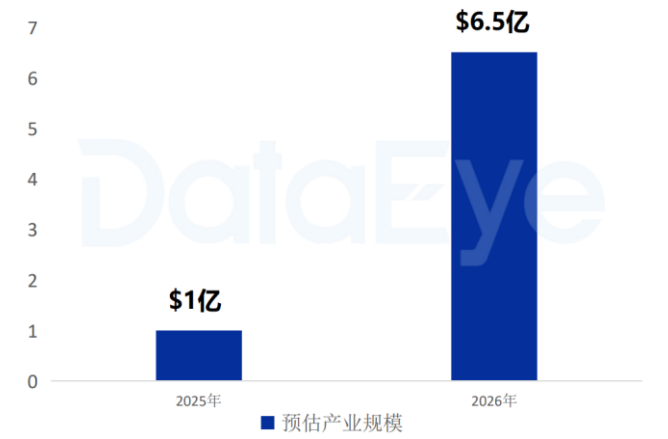
数据来源：AI 产品榜，FirmCritics，CSDN，Voicemod 官网，Moises 官网，国泰海通证券研究

AI 音乐已快速进入流媒体供给端，但数量增长仍快于真实消费增长。Deezer 检测到的全 AI 生成歌曲由 2025 年 1 月的每日约 1 万首、占新增上传量 10%，增至 2025 年 9 月的 3 万首和 28%，2026 年 4 月进一步达到 7.5 万首和 44%，相当于每月新增超过 200 万首。AI 歌曲也开始进入主流榜单，2025 年 11 月，全 AI 生成的乡村歌曲《Walk My Walk》曾登顶 Billboard 乡村数字歌曲销售榜。但截至 2026 年 4 月，AI 歌曲仅贡献 Deezer 总播放量的 1%—3%，其中 85%的播放被识别为欺诈流量，说明当前供给扩张快于真实用户需求增长。随着生成质量提升、平台标识和版权归属规则逐步完善，AI 音乐有望由低成本批量上传转向更高质量的创作与发行，并在流媒体、影视游戏配乐和广告音乐等领域持续扩大应用。

4.4.2. 视频生成编辑

海外 AI 剧与漫剧供给加快增长，成熟短剧平台为内容分发和商业化提供基础。DataEye 预计，海外 AI 剧与漫剧市场规模将由 2025 年的约 1 亿美元增至 2026 年的 6.5 亿美元；2026 年 1—3 月，海外短剧素材榜 TOP100 中 AI 剧与漫剧占比由 2%提升至 16%，3 月上榜作品增至 16 部，反映内容供给和买量投放同步增加。平台端头部格局较为稳定，2025Q1 ReelShort 和 DramaBox 收入分别达到 1.20 亿美元和 1.15 亿美元，显著领先其他海外短剧应用；其中，ReelShort 2025 年 12 月下载量约 2072 万次，创年内新高。我们认为，头部平台积累的用户规模、本地化运营、投放经验和付费体系，有助于缩短 AI 内容进入海外市场的周期，并推动 AI 剧与漫剧由单部内容试水转向持续化供给。

图53：海外 AI 剧/漫剧产业规模预估



资料来源：DataEye

图54：海外短剧榜 TOP100 中 AI 剧/漫剧占比



资料来源：DataEye

图55：海外 AI 剧/漫剧产业规模预估

2025Q1 海外短剧应用收入Top10

短剧APP	2025Q1收入（百万美元）	2025Q1排名	相较2024Q1排名变化	团队背景	所属公司	
ReelShort	120.20	1	-	中国	枫叶互动	头部格局稳定
DramaBox	115.18	2	-	中国	点众科技	
GoodShort	51.13	3	+1	中国	新阅时代	
ShortMax	44.83	4	-1	中国	九州文化	
NetShort	21.50	5	2024.7上线	中国	麦芽文化	中腰部流动性明显
DramaWave	21.12	6	2024.10上线	中国	昆仑万维	
FlickReels	16.59	7	2024.7上线	中国	掌玩网络	
MoboReels	13.37	8	-2	中国	畅读科技	
Kalos TV	12.40	9	-2	中国	青榕文化	
My Drama	11.75	10	2024.4上线	乌克兰	Holy Water	

资料来源：霞光社

图56：海外短剧榜 TOP100 中 AI 剧/漫剧占比



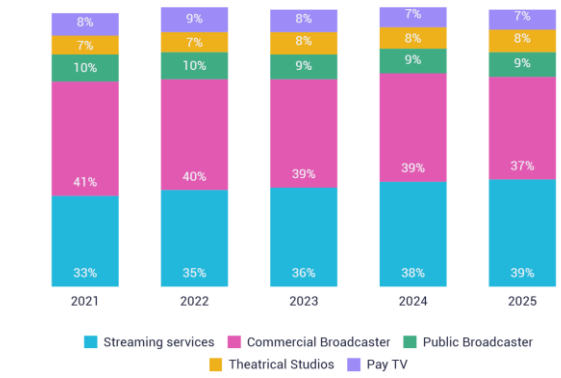
资料来源：点点数据

海外影视内容投入规模庞大，视频 AI 或创造电影行业新增长点。Ampere

下载日志已记录，仅供内部参考

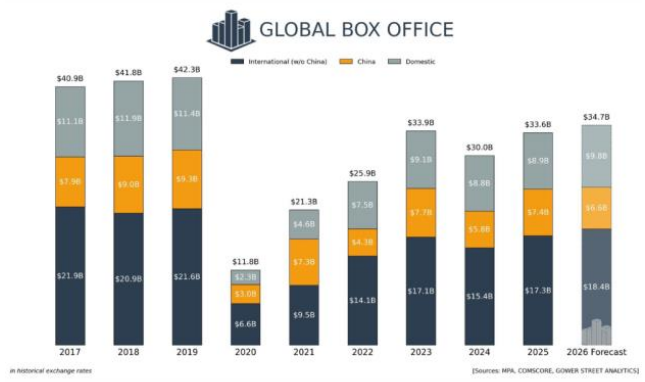
Analysis 预计，2025 年全球影视内容投入达到 2480 亿美元，其中流媒体平台投入同比+6%至 950 亿美元，首次超过商业电视台。电影市场亦保持较大规模，2025 年全球票房约 335.5 亿美元，Gower Street 预计 2026 年将增至 347 亿美元。制作端，据 Morgan Stanley Research，生成式 AI 有望使影视制作成本最多下降 30%，可用于剧本、拍摄、剪辑、声音混制和视觉特效等环节。据 Autodesk，Boxel Studio 在《超人和露易丝》第三季中引入 AI 动作捕捉后，原本一名高级动画师制作 5 秒高质量镜头最长需要 3 周，现可缩短至 2-3 天，并保持相近质量。Netflix 已在阿根廷剧集《永航员》中使用生成式 AI 制作建筑倒塌镜头，相关特效序列的完成速度达到传统方式的 10 倍，使原本受预算限制的镜头得以保留。借由 AI，制片方能够降低设计、虚拟场景、视觉特效和后期制作的时间与成本，使中小预算项目也能使用过去服务于大型制作的视觉效果。

图57：流媒体平台内容支出占比持续提升



资料来源：Ampere Analysis

图58：全球影视票房持续恢复



资料来源：Gower Street

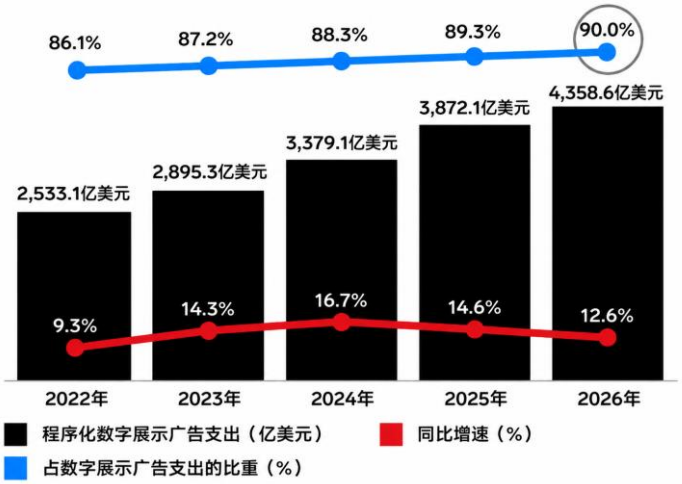
模型技术发展将持续提升优质视频产能。更多优质内容上线将帮助 ReelShort、DramaBox 等平台增加用户时长和付费，提升 ARPU；市场规模扩大也会吸引更多制作团队进入，带动视频生成和编辑需求。对于电影和长剧，短期内 AI 仍主要用于特效、修复、剪辑等后期环节，降低中小预算项目的制作成本；随着人物一致性、镜头衔接和长视频生成能力提升，AI 有望进一步进入分镜设计、虚拟拍摄和部分成片制作，缩短影视制作周期并拓展可制作的内容类型。

4.5. 营销类

程序化广告已成为全球广告投放的主要方式，AI 进一步提高投放自动化和匹配精度。传统人工投放依赖经验和反复试错，程序化系统则可根据用户特征、转化目标和实时反馈自动完成定向、竞价及预算调整。据 EMARKETER，全球程序化数字展示广告支出预计由 2022 年的 2,533 亿美元增至 2026 年的 4,359 亿美元，2022-2026 年复合增速约为 14.5%，占全球数字展示广告支出的比重由 86.1%升至 90.0%。程序化广告规模扩大，为多模态模型进入素材理解、用户意图判断和广告推荐提供了基础。

下载日志已记录，仅供内部参考

图59：2022-2026 年，程序化广告收入及在广告行业占比不断提升



资料来源：EMARKETER

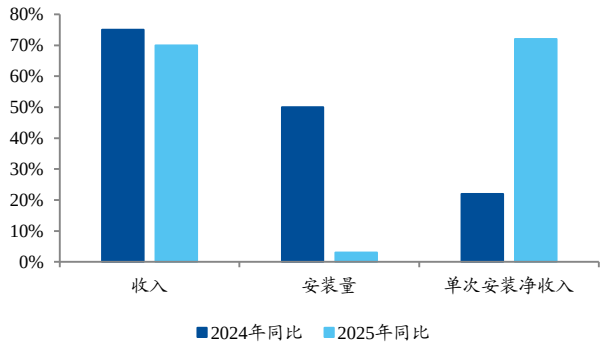
Axon AI 推动 AppLovin 广告增长由安装规模扩张转向单次安装价值提升。AppLovin Ads 依托 Axon 推荐引擎自动完成用户匹配、实时竞价和投放优化，持续提高广告主投放回报。2024 年公司收入同比增长 75%，广告带来的应用安装量和单次安装净收入分别同比增长 50%和 22%；2025 年收入仍同比增长 70%，但安装量仅同比增长 3%，单次安装净收入则同比增长 72%，显示收入增长主要来自广告匹配和变现效率提升。季度来看，公司收入由 2024Q3 的 8.35 亿美元持续增至 2025Q3 的 14.05 亿美元，连续五个季度增长。我们认为，Axon 持续迭代使 AppLovin 在安装规模增速放缓的情况下仍能保持较快收入增长，这也表明 AI 推荐能力能够直接转化为广告收入。

图60：AppLovin 五个季度收入持续增长



资料来源：PPC Land

图61：2025 年 AppLovin 收入增长由单次安装净收入提升驱动

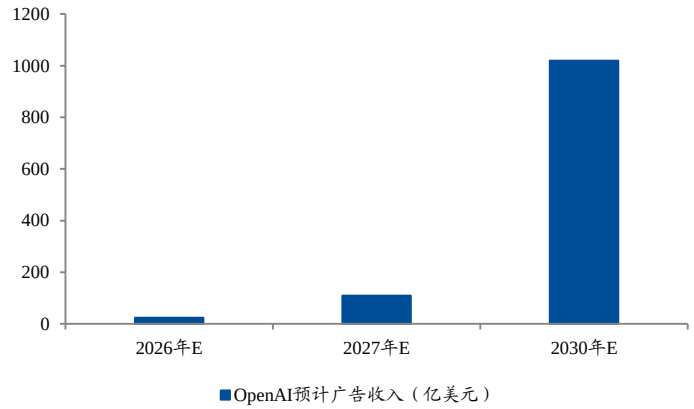


资料来源：AppLovin 2025 年年报，国泰海通证券研究

OpenAI 开始将 ChatGPT 的用户规模和决策需求转化为广告收入。2026 年 2 月，OpenAI 在美国 Free 和 Go 用户中测试广告，广告主要依据对话内容和用户意图完成匹配；上线六周后，广告业务年化收入已超过 1 亿美元。5 月，OpenAI 进一步推出自助式 Ads Manager 和 CPC 竞价，支持按展示和点击收费，将 ChatGPT 由单一订阅产品扩展为广告投放平台。据 The Information，OpenAI 预计广告收入将由 2026 年的约 24 亿美元增至 2027 年的近 110 亿美元，并于 2030 年达到约 1020 亿美元，占当年总收入约 36%。我们认为，以 ChatGPT 为代表的高单位流量价值 AI 应用加快商业化，有望为全球广告市场贡献新增收入。

下载日志已记录，仅供内部参考

图62：Open AI 预计其广告收入提升至 2030 年的 1000 亿美元



资料来源：The Information，国泰海通证券研究

广告收入增长取决于需求识别和转化效果。当广告触达从泛流量曝光转向更明确的需求表达，平台能否连接搜索、推荐、交易和转化数据，将决定广告主的预算分配。ChatGPT 等对话式产品的广告负载通常低于传统信息流，其收入空间更依赖单次转化价值，而非广告数量。长期看，用户信任、效果衡量和交易闭环将成为 AI 广告商业化的主要门槛。

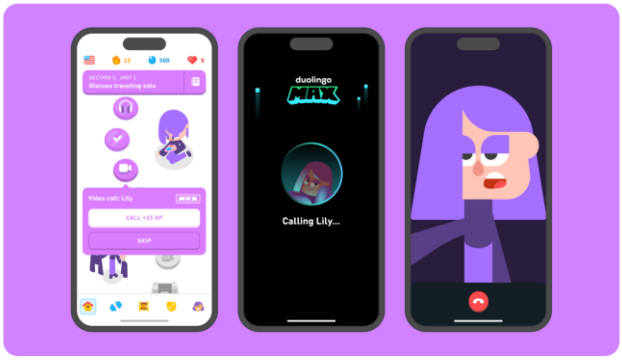
AI 将同时提升营销内容生产和广告投放效率。多模态大模型可降低营销素材制作成本、提高产出效率，程序化广告能力领先的企业有望受益于广告市场增长。

4.6. 教育学习类

一对一教学由稀缺服务转化为规模化产品。传统课堂由一名教师面对多名学生，难以持续适配每个人的基础和学习速度；AI 辅导可以随时回答问题、调整讲解节奏并提供即时反馈。据 2025 年发表于《Scientific Reports》的一项哈佛大学随机对照实验，使用结构化 AI 辅导的学生，学习增益超过课堂主动学习组的 2 倍，完成学习的中位时间为 49 分钟，低于课堂组的 60 分钟。

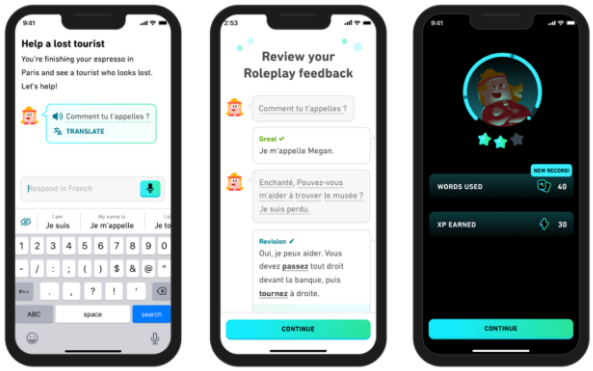
语言学习兼具高频使用和明确付费需求，Duolingo 已加入 AI 教育。Duolingo Max 通过 Video Call、Roleplay 和 Explain My Answer 等功能提供 AI 实时对话与个性化反馈，改变了过去口语对话练习只能依赖真人教师的情况。截至 2025 年末，Duolingo 日活跃用户、月活跃用户和付费订阅用户分别达到 5270 万、1.33 亿和 1220 万，付费渗透率为 9.2%。AI 还显著加快课程发布，公司 2026Q1 发布 2.05 万个学习技能，高于 2025 年季度均值的 7100 个和 2024 年的 1800 个。

图63：Duolingo Max 视频通话功能



资料来源：Duolingo 官方博客

图64：Duolingo Max 角色扮演功能



资料来源：Duolingo 官方博客

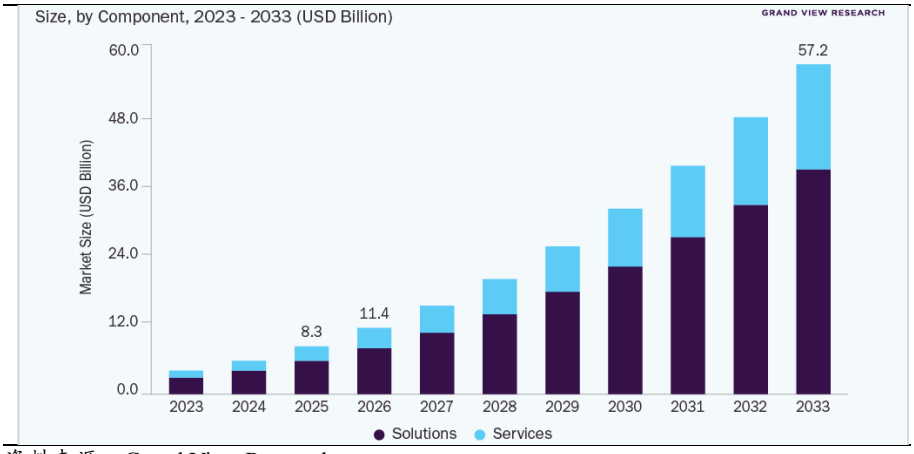
教师辅助工具提升备课和内容制作环节效率。据 Gallup 与 Walton Family Foundation 对 2232 名美国公立学校教师的调查，2024-2025 学年已有 60%

下载日志已记录，仅供内部参考

的教师在工作中使用 AI，至少每周使用一次的教师平均每周节省 5.9 小时，相当于每学年节省约 6 周工作时间。AI 主要用于备课、制作练习材料、调整不同难度的教学内容以及生成测验；Khanmigo 也已提供课程计划、评分标准、测验、学生进度分析等 25 项以上教师工具。AI 节省下来的时间教师可用于单独辅导、课堂互动和反馈答疑。

教育 AI 市场增长将由学校、个人和企业共同推动。Grand View Research 预计，全球教育 AI 市场规模将于 2033 年达到 572 亿美元，2026—2033 年复合增速为 25.9%；其中学习平台与虚拟辅导产品 2025 年市场规模约 38.25 亿美元，预计 2025—2033 年复合增速为 26.4%。K-12、高等教育和企业培训均具备增长空间。

图65：全球 AI 教育市场规模



资料来源：Grand View Research

AI 将推动教育由统一教学向个性化学习转变。我们认为，智能辅导、虚拟教师和自适应学习系统可根据学生基础与进度提供差异化教学，提高教学质量和效率，增强学习兴趣并培养创新能力。AI 将由辅助工具逐步进入教学核心环节，扩大优质教育供给，推动教育服务更加个性化、普惠和高效。

4.7. 游戏类

4.7.1. 游戏开发生成

生成式 AI 已进入游戏研发，现阶段主要用于辅助环节。GDC 对超过 2300 名游戏行业从业者的调查显示，2026 年已有 36%的受访者在工作中使用生成式 AI，其中研究与创意讨论的使用比例达到 81%，代码辅助和日常事务均为 47%，原型制作为 35%。

游戏开发工具从素材生成延伸至代码、场景和对象编辑。Roblox 推出的 Cube 3D 可根据文本生成与游戏引擎兼容的三维模型和环境，并计划由静态网格扩展至具备结构、物理属性和交互功能的游戏对象。Unity AI 则直接嵌入 Unity 编辑器，可读取场景层级、游戏对象、已安装软件包和目标平台，通过自然语言完成代码编写、对象创建、属性调整和素材生成。开发者无需在通用模型、建模软件和游戏引擎之间反复传递内容，AI 开始参与游戏工程开发全流程。

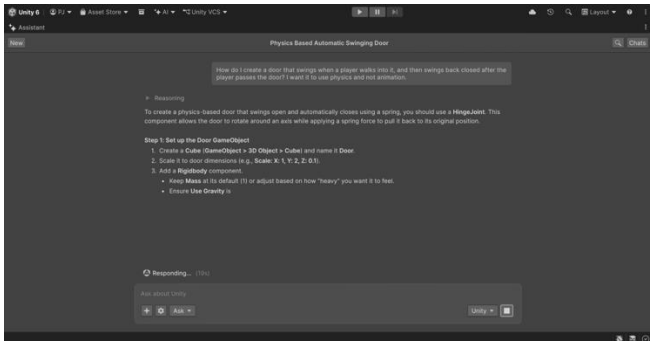
下载日志已记录，仅供内部参考

图66：Roblox Cube 由文本生成 3D 及可交互资产



资料来源：腾讯新闻

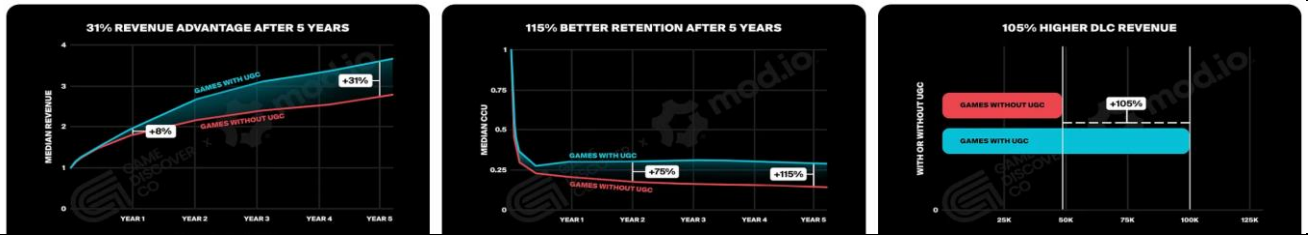
图67：Unity AI 的游戏开发功能



资料来源：Unity 官网

内容制作门槛降低后，UGC 游戏有望率先获得更多内容供给。据 GAMES.GG，支持官方 UGC 或 Mod 的游戏首年收入较其他游戏高 8%，五年累计收入高 31%；五年后的同时在线玩家数高 115%，单项 DLC 收入中位数高 105%。AI 可帮助普通玩家生成模型、关卡、脚本和交互物体，减少三维建模和编程要求，平台再将内容提供给其他玩家，使创作者、游戏公司和平台共同获得收入，进而强化分发平台及 UGC 生态游戏的内容正循环

图68：UGC 提升游戏收入、留存及后续变现能力



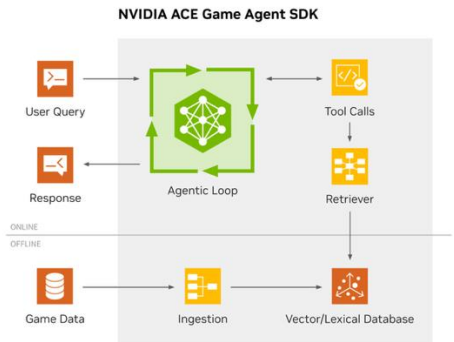
资料来源：GAMES.GG 官网

AI 推动游戏研发提效和 UGC 内容扩张。AI 对游戏开发运营的整体降本增效；Roblox 等海外游戏分发与 UGC 平台凭借门槛更低、能力更强的 AI 创作工具，有望吸引更多开发者和玩家参与内容生产，打造增长飞轮。

4.7.2. 智能 NPC

智能 NPC 推动游戏角色由固定脚本向实时对话和自主行动升级。传统 NPC 主要按照预设对话树和行为规则响应玩家，生成式 AI 则可结合语言模型、语音识别与合成、长期记忆、环境感知和动作规划，使角色理解自然语言并根据游戏状态作出反应。NVIDIA ACE 已提供覆盖语音、智能决策和角色动画的云端及端侧模型，帮助开发者制作可对话、可行动的游戏角色；育碧《Teammates》则允许玩家通过实时语音向 AI 队友下达指令，NPC 可结合玩家需求生成回应并执行相应行动。

图69：NVIDIA ACE Game Agent SDK 工作流程



资料来源：NVIDIA 官网

图70：《Teammates》语言控制 NPC



资料来源：《Teammates》游戏画面

下载日志已记录，仅供内部参考

智能 AI 角色对用户增长和游玩时长有带动作用。《Status》允许玩家创建身份并进入不同虚拟世界，世界中的 AI 角色可根据玩家的发帖、私信和选择作出回应，角色关系和故事也会随互动变化。产品上线 19 天用户数即突破 100 万；据 Inworld，其日活跃用户超过 50 万，用户日均游玩时长达到 1 小时 36 分钟。2026 年 5 月，《Status》累计创建超过 1300 万个世界和 500 万个角色档案。与固定脚本 NPC 相比，可记忆、可回应的 AI 角色能够持续生成新的关系和故事，增加玩家重复进入游戏和长期互动的动力。

智能 NPC 有望提升开放世界、角色扮演和社交游戏的互动深度。具备长期记忆和自主行动能力的角色可根据不同玩家形成差异化关系、任务和剧情，增加重复游玩次数与用户时长，并带动道具、订阅及内容付费。成熟 IP 拥有完整的世界观和人物设定，更容易保持 AI 角色行为与剧情一致，相关价值有望进一步提升。

5. 投资建议

海外 AI 应用的商业化进展，为国内相关公司提供参照。影视方面，海外 AI 视频、短剧和漫剧已验证降低制作门槛、扩大内容供给的可行性，国内短剧市场体量较大，具备 IP 储备、内容制作和分发渠道的公司更容易将技术转化为稳定的内容产出。游戏方面，海外 AI 开发工具、智能 NPC 和 UGC 平台已验证商业化路径，国内具备相关能力的公司，可在现有游戏中提高制作开发效率，并通过智能 NPC、动态剧情和玩家共创增加用户时长与付费。营销方面，海外程序化广告平台已验证智能投放的商业价值，国内具备广告技术、客户资源和流量渠道的平台有望承接同类需求。我们认为，海外映射的核心是寻找业务环节相同、产品能力相近，并拥有 IP、数据、用户和分发渠道的公司，推荐 AI 视频、游戏及营销三条主线。

推荐标的：

1）AI 视频：中文在线、昆仑万维、芒果超媒、上海电影、华策影视

中文在线：AI 全链路赋能构筑核心壁垒，开启平台化增长。公司已构建覆盖逍遥 AI 文本创作、次元神笔视频生成到逍遥 AI Agent 自动化投流的全栈式 AI 基础设施，并在行业内率先实现 AI 剧从技术验证迈向规模化生产，截至 2026 年 6 月月度 AI 剧产能已达 150 部。横琴国际影视城正式启动，旗舰短剧平台 FlareFlow 快速发展。我们认为，随着 AI 剧产能持续爬坡与海外 FlareFlow 平台用户生态加速裂变，公司商业闭环初步形成，价值释放有望加速。

昆仑万维：自研多模态模型与海外内容平台协同，短剧业务已形成收入规模。公司形成 SkyReels 视频生成模型、Mureka 音乐生成平台、Matrix 世界模型及天工智能体等产品矩阵，并将模型能力用于 AI 社交、音乐创作和海外短剧。截至 2025 年 3 月，DramaWave 月活跃用户超过 1000 万；截至 2025 年 6 月末，平台年化流水收入突破 2.4 亿美元，单月流水收入突破 2000 万美元。同期 Mureka 年化收入约 1200 万美元，模型能力已开始转化为应用收入。我们认为，公司同时具备底层模型、AI 原生产品及海外运营能力，有望通过短剧、音乐和社交等产品持续验证多模态模型的商业价值，并推动 AI 业务收入规模扩大。

芒果超媒：内容制作与平台分发一体化，政策调整有望进一步释放内容供给。公司以芒果 TV 为核心，业务覆盖综艺及剧集制作、会员、广告和运营商服务，并依托湖南卫视形成内容制作与平台分发协同。2025 年 10 月，《声鸣远扬 2025》上线；2026 年 2 月，《声起地平线》开播。公司持续推进人工智能在内容创作、生产和分发环节的使用。我们认为，公司成熟的内容制作体系、平台用户基础及台网协同优势，有助于将 AI 带来的制作效率提升转化为更丰富的内容供给，并进一步支撑会员和广告业务增长。

上海电影：影院业务提供经营基础，经典 IP 开发打开第二增长曲线。公司

下载日志已记录，仅供内部参考

形成电影发行放映与 IP 开发运营两条业务主线，上影元围绕经典影视和动画 IP 开展内容焕新、授权合作、衍生商品、整合营销及沉浸式娱乐。我们认为，多模态 AI 有助于降低动画、短视频和互动内容的制作门槛，提高经典 IP 的内容更新频率；公司在 IP 资源、内容制作及衍生经营方面的积累，有望进一步释放版权价值并拓展 IP 授权和商品销售收入。

华策影视：影视内容储备充足，算力业务成为新增业务方向。公司持续推进剧集、电影和微短剧业务，2025 年《太平年》《迷径之上》《你是迟来的欢喜》等项目取得发行许可证，《家业》《势在必行》等项目持续推进。公司同时拓展算力贸易及算力服务，2025 年算力业务收入 1.26 亿元，同比+602.65%；截至 2025 年末，智算规模达到 8000P。我们认为，公司兼具影视内容制作经验和算力资源，有望将 AI 进一步用于剧本辅助、素材生成、后期制作及云端渲染，在提高内容生产效率的同时拓展算力服务收入。

2）游戏：世纪华通、恺英网络、巨人网络、完美世界、心动公司

世纪华通：海外 SLG 产品持续验证，产品与投放数据积累构成 AI 应用基础。公司旗下点点互动形成以 SLG 和休闲游戏为主的海外产品矩阵，《Whiteout Survival》《Kingshot》《Tasty Travels: Merge Game》分别位列 2025 年 Sensor Tower 中国手游境外收入榜第 1、第 3 和第 27 位。我们认为，头部产品积累的大量用户行为、付费及广告投放数据，可用于副玩法迭代、用户分层、素材生成和投放优化，有助于提高海外获客效率并延长产品生命周期。

恺英网络：游戏产品、平台和 AI 投资并行，已覆盖研发及用户交互环节。公司主营传奇、奇迹等优势品类，并通过“996 传奇盒子”拓展平台业务。在 AI 方面，公司投资企业自然选择推出 3D AI 陪伴应用 EVE。公司推出 AI 游戏工具平台 SOON，通过自研垂类模型和 AP 分级品控体系，将 AI 游戏从概念 Demo 推向具备商业变现能力的成品游戏，并凭借全链路自动化生成能力，降低了传统游戏的开发成本与周期，释放了游戏创作的生产力。我们认为，公司现有产品、IP、平台和开发者资源，为 AI 工具进入游戏研发、发行运营及智能角色等环节提供了较明确的使用入口，有望进一步提升研发效率并丰富玩家交互体验。

巨人网络：爆款产品与 AI 原生玩法同时验证，智能角色已进入大规模用户交互。《超自然行动组》日活跃用户超过千万，公司在产品中推出 AI 玩法；截至 2026 年一季度末，相关玩法参与对局数超过 4 亿场，累计生成 AI 假人超过 29 亿个。我们认为，该产品验证了 AI 不仅能够辅助美术、代码和测试，还可直接参与玩家组队、社交及对局过程；公司积累的大规模交互数据有助于持续优化智能角色行为，并推动更多 AI 原生玩法进入游戏产品。

心动公司： AI 驱动 TapTap 从渠道向创作平台跃迁，成长逻辑全面升级。心动公司凭借 TapTap 平台与自研游戏形成的“内容+渠道”双轮驱动。TapTap Maker 作为 AI 游戏创作智能体，上线仅半年已驱动用户发布超 6000 款游戏，日均产出 10-20 款独占内容，通过自然语言开发极大降低了创作门槛，将 TapTap 从分发渠道升级为内容共创平台；同时 AI 已深度嵌入公司研发、运营与算法优化的全链路环节，AI 降本预期叠加 TapTap 平台广告转化率持续提升。我们认为，公司“游戏+AI+平台”的新成长飞轮已初步成型，平台价值重估空间正在打开。

3）营销：易点天下、值得买

易点天下：出海营销渠道和程序化广告能力突出，AI 创意工具完善业务链条。公司形成商业智能营销平台 Cyberklick、数字广告平台 Yeahmobi 和程序化广告平台 zMaticoo 三大核心业务品牌，媒体端与 Google、Meta、TikTok 等全球平台合作，客户主要覆盖跨境电商、游戏和移动应用等行业，并持续

下载日志已记录，仅供内部参考

拓展短剧、泛娱乐等新兴客户。公司推出 AIGC 数字营销创作平台 KreadoAI，可提供数字人、配音、翻译和营销素材生成等功能，并用于电商购物、应用下载、教育培训及企业服务。我们认为，AI 创意生产能够降低多语言、多版本素材的制作成本，并与公司现有投放数据和媒体渠道形成协同，有助于提高广告投放效率并拓展出海营销客户。

值得买：消费内容与用户口碑积累构成差异化优势，AI 购物助手有望向消费决策入口延伸。公司以“什么值得买”为核心，长期沉淀商品信息、用户评价、消费经验及价格数据，主要通过信息推广、互联网效果营销和品牌营销等方式实现变现。2025 年 9 月，公司推出独立 App 并开启 AI 购物助手“张大妈”公测，产品可通过意图识别和任务执行，为用户提供商品推荐、口碑总结、价格比较及购买决策服务。我们认为，公司积累的消费内容、商品数据和用户口碑为 AI 购物助手提供了较强的数据基础，有望推动平台由内容导购进一步向需求识别和消费决策入口升级。

表12：推荐标的盈利预测表

代码	简称	股价（元）	市值（亿元）	EPS（元/股）			PE		
				2025A	2026E	2027E	2025A	2026E	2027E
300364.SZ	中文在线	24.46	178.19	-0.92	0.03	0.18	-27.27	800.32	138.36
300413.SZ	芒果超媒	19.33	361.61	0.66	0.57	0.76	37.23	33.84	25.51
300418.SZ	昆仑万维	42.88	551.02	-1.28	-0.28	0.02	-32.87	-155.76	2081.37
601595.SH	上海电影	23.53	105.46	0.25	0.22	0.35	113.66	105.11	67.56
300133.SZ	华策影视	7.47	140.05	0.10	0.18	0.22	82.19	40.74	33.96
002602.SZ	世纪华通	14.34	1056.96	0.78	1.22	1.46	22.43	11.79	9.81
002517.SZ	恺英网络	16.38	349.95	0.89	1.27	1.43	24.54	12.92	11.43
002558.SZ	巨人网络	23.60	448.53	0.94	2.29	2.60	47.74	10.32	9.09
603444.SH	吉比特	344.58	248.24	24.90	27.18	29.80	17.02	12.68	11.56
002624.SZ	完美世界	9.89	191.86	0.39	0.77	1.07	43.51	12.82	9.25
2400.HK	心动公司	35.86	173.53	3.07	3.34	3.99	18.88	10.74	8.99
301171.SZ	易点天下	34.71	214.69	0.33	0.34	0.49	120.88	102.56	71.41
300785.SZ	值得买	36.23	72.05	0.44	0.54	0.78	100.72	67.28	46.64

数据来源：Choice，国泰海通证券研究，股价时间为 2026/9/23，港元汇率为 1 港元=0.86 人民币

6. 风险提示

（1）大模型技术进展不及预期。大模型能力提升取决于模型架构、训练数据、算力资源及研发投入，推理、多模态、长上下文和工具调用等能力的迭代存在不确定性。若模型性能提升或单位调用成本下降慢于预期，可能限制 AI 产品功能升级，并影响用户使用频率及下游需求增长。

（2）AI 应用商业化进展不及预期。AI 应用收入增长受产品效果、使用成本、用户留存、付费意愿及企业投入回报等因素影响。若用户规模、付费渗透率或企业采购需求增长不及预期，相关公司可能面临收入增速放缓、算力成本较高及盈利改善不及预期的风险。

（3）行业竞争加剧。科技巨头、模型厂商及 AI 原生应用公司持续增加研发、算力和市场投入，围绕模型能力、产品功能、价格及用户入口的竞争不断加剧。若相关公司无法维持技术、产品、内容或渠道优势，可能面临用户流失、获客成本上升、产品降价及市场份额下降的风险。

（4）政策监管、数据合规及版权风险。生成式 AI 涉及个人信息保护、训练数据来源、内容版权、算法透明度及生成内容安全，各国家和地区的监管制度持续完善。若监管要求趋严，或相关公司在数据采集、模型训练、内容

下载日志已记录，仅供内部参考

生成及用户数据使用方面出现合规问题，可能增加内容审核、版权采购和产品整改成本，并影响产品推广及商业化进程。

（5）算力供给及资本开支不及预期。大模型训练和推理依赖高性能芯片、服务器、数据中心、电力及云资源，算力建设受到硬件供应、建设周期、融资能力和应用需求等因素影响。若高端芯片及电力资源供给受限，或科技公司资本开支低于预期，可能延缓模型迭代，并影响 AI 基础设施及相关产业链需求增长。

下载日志已记录，仅供内部参考

本公司具有中国证监会核准的证券投资咨询业务资格

分析师声明

作者具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，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，本报告清晰准确地反映了作者的研究观点，力求独立、客观和公正，结论不受任何第三方的授意或影响，特此声明。

免责声明

本报告仅供国泰海通证券股份有限公司（以下简称“本公司”）授权客户使用。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。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告。

本报告的信息来源于已公开的资料，本公司对该等信息的准确性、完整性或可靠性不作任何保证。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可升可跌。过往表现不应作为日后的表现依据。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司不保证本报告所含信息保持在最新状态。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。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。在任何情况下，本公司、本公司员工或者关联机构不承诺投资者一定获利，不与投资者分享投资收益，也不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。

本公司利用信息隔离墙控制内部一个或多个领域、部门或关联机构之间的信息流动。因此，投资者应注意，在法律许可的情况下，本公司及其所属关联机构可能会持有报告中提到的公司所发行的证券或期权并进行证券或期权交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务。在法律许可的情况下，本公司的员工可能担任本报告所提到的公司的董事。

市场有风险，投资需谨慎。投资者不应将本报告作为作出投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向专业人士咨询并谨慎决策。

本报告版权仅为本公司所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制、发表或引用。如征得本公司同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“国泰海通证券研究”，且不得对本报告进行任何有悖原意的引用、删节和修改。

若本公司以外的其他机构（以下简称“该机构”）发送本报告，则由该机构独自为此发送行为负责。通过此途径获得本报告的投资者应自行联系该机构以要求获悉更详细信息或进而交易本报告中提及的证券。本报告不构成本公司向该机构之客户提供的投资建议，本公司、本公司员工或者关联机构亦不为该机构之客户因使用本报告或报告所载内容引起的任何损失承担任何责任。

评级说明

投资建议的比较标准

投资评级分为股票评级和行业评级。

以报告发布后的 12 个月内的市场表现为比较标准，报告发布日后的 12 个月内的公司股价（或行业指数）的涨跌幅相对同期的沪深 300 指数涨跌幅为基准。

	评级	说明
股票投资评级	增持	相对沪深 300 指数涨幅 10%以上
	谨慎增持	相对沪深 300 指数涨幅介于 5%～10%之间
	中性	相对沪深 300 指数涨幅介于-5%～5%
	减持	相对沪深 300 指数下跌 5%以上
	无评级	对于个股未来市场表现与基准指数相比无明确观点
行业投资评级	增持	明显强于沪深 300 指数
	中性	基本与沪深 300 指数持平
	减持	明显弱于沪深 300 指数

国泰海通证券研究所

地址 上海市黄浦区中山南路 888 号 邮编 200011 电话 （021）38676666