OpenAI产品谱系与生态位研究报告

报告日期：2026.5.15　　报告人：wang hua

执行摘要

OpenAI 已从一家以前沿研究为主的实验室，演化为一家“全栈 AI 基础设施与应用公司”：上层是面向大众的 ChatGPT、Sora、Deep Research、Operator、Codex 等终端产品；中层是面向开发者的 API 平台、Responses API、Agents SDK、Realtime API、工具调用与企业部署能力；底层则是持续迭代的 GPT、o 系列、音频、图像、视频、内容审核与安全评测体系。其核心战略不是只卖“模型”，而是同时占据模型、平台、分发、企业工作流、内容获取与安全治理多个位置。[1]

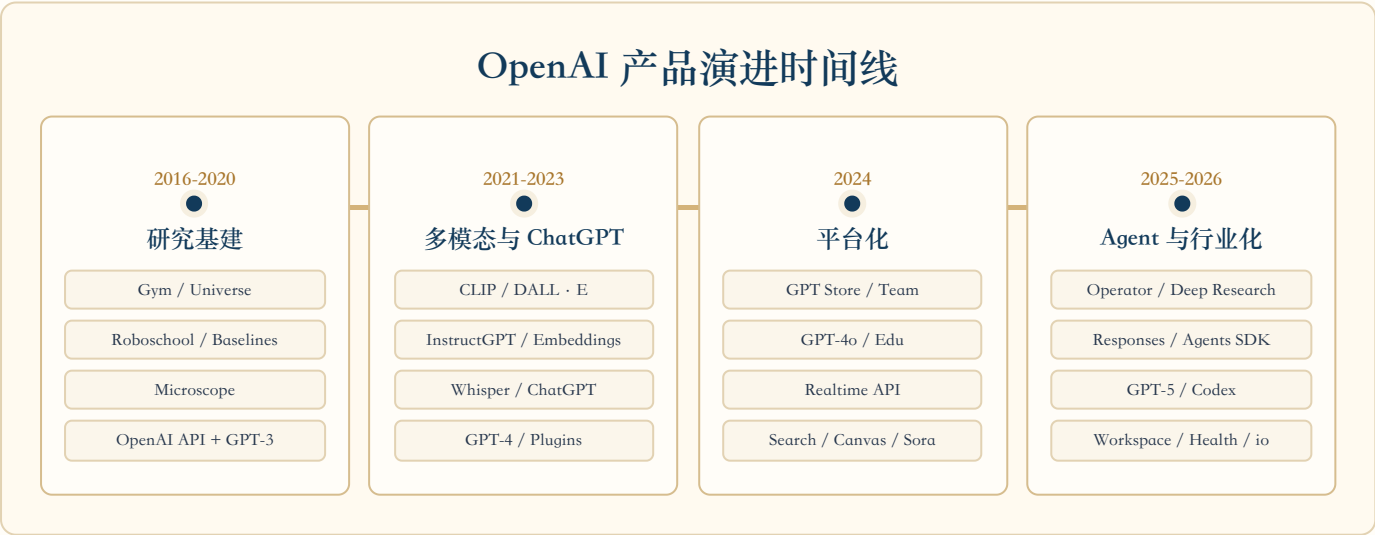
OpenAI 当前生态位的关键特征有三点。第一，它既是研究机构，也是 API 供应商、消费级应用开发者、企业软件商与事实上的标准推动者。第二，它通过 ChatGPT 获得超大规模终端分发，再把这种分发反哺开发者与企业生态；官方到 2025 年已公开提到 ChatGPT 超过 8 亿周活用户、OpenAI 直接服务超过 100 万家企业客户、超过 900 万付费企业用户使用 ChatGPT 工作，API 平台处理能力达到每分钟逾 150 亿 token，企业收入占比亦已超过 40%。第三，它正从“单次问答”转向“可执行任务的 agent 平台”，用 Responses API、Agents SDK、Computer Use、Web Search、File Search、Apps in ChatGPT、Workspace agents 等产品，把模型从回答器变成行动层。[2]

商业上，OpenAI 已形成多元收入结构：消费者订阅（Free/Go/Plus/Pro）、企业席位（Business/Enterprise/Edu）、API 按量计费、定制模型与咨询式部署、内容授权/数据合作、以及自 2026 年起测试的广告收入。治理上，它从 2019 年的 capped-profit 架构，过渡到 2025 年确定、2025 年 10 月正式完成的“非营利基金会控制下的 PBC”结构，以兼顾超大规模融资需求与使命治理。监管和信任层面，OpenAI 一边推进 Preparedness Framework、Model Spec、System Card、Moderation API、Red Teaming Network 等安全工具，一边面临 FTC 调查、版权诉讼、公司结构审查与欧盟平台监管等外部压力。[3]

如果用一句话概括：**OpenAI 构建的不是单一产品矩阵，而是一个以 ChatGPT 为分发入口、以 API/agent 平台为开发中枢、以安全治理为护城河、并向硬件与行业垂直场景外延的 AI 操作层生态。** 这使它既像“新时代的云 API 公司”，又像“AI 版应用商店 + 工作平台 + 新型操作系统”。[4]

产品谱系与时间轴

下图只抽取 OpenAI 产品化路径中的关键节点；日期均来自官方发布页或官方文档。[5]



关键产品目录

下表以“重大发布”为口径，覆盖模型、API、开发者工具、消费应用、插件/应用、研究平台与硬件动作。若 OpenAI 未公开定价，则标注“未披露”或“包含于计划/API 通用计费”。技术谱系为高层概括，而非完整训练细节。表中“官方来源”均为 OpenAI 官

方博客、文档或官方属性页面。

发布时间	产品/发布	核心能力	目标用户	变现/定价	技术谱系	官方来源
2016-04	OpenAI Gym	强化学习环境与基准，便于算法复现与比较	RL 研究者、教育用户	免费/开源	早期 RL 研究基础设施	6
2016-12	Universe	把游戏、网站、应用统一成可训练环境，让 agent 通过像素、键鼠操作任务	通用智能/RL 研究者	免费/开源	建立在 Gym 思路之上，强调“任意程序都可成环境”	7
2017-05	Roboschool / Baselines	机器人仿真环境；RL 算法实现基线	RL/机器人研究者	免费/开源	OpenAI Gym 生态延展	8
2018-05 / 2018-11	Gym Retro / Spinning Up	大规模游戏 RL 平台；深度 RL 教程与代码	RL 研究者、学习者	免费/开源	Gym 生态深化与教育化	9
2020-04	OpenAI Microscope	神经元可视化与可解释性研究平台	研究者	免费/研究平台	解释性研究工具	10
2020-06	OpenAI API + GPT-3	通用 text-in/text-out API；少样本泛化	开发者、创业公司、企业	API 按量计费；最初申请制，后取消 waitlist	GPT-3（175B 参数 few-shot 论文）	11
2021-01	CLIP / DALL · E	多模态文本-图像对齐；文本生成图像	研究者、创意工具开发者	研究发布；商业化当时未披露	GPT-2/3 式扩展到视觉；DALL · E、CLIP 互为关键前置	12
2021-07	Triton	Python 风格 GPU 编程语言，提高 kernel 开发效率	研究/基础设施工程师	免费/开源	训练与推理基础设施	13
2021-08	Codex 初代	自然语言转代码	开发者	包含在 API 平台能力中；后续独立演化	明确由 GPT-3 在 GitHub 代码上微调扩展而来	14
2022-01	InstructGPT / Embeddings	通过 RLHF 更好服从指令；向量语义检索与分类	API 开发者、企业应用	API 按量计费	InstructGPT 建立在 GPT-3 上；Embeddings 明确为 GPT-3 后裔	15
2022-03 / 2022-11	DALL · E 2 / DALL · E API	更高质量文本生图、编辑、变体；后开放 API 公测	创意个人、设计团队、开发者	DALL · E Beta；API 按量计费	DALL · E 2 明确相对 DALL · E 1 升级；与 CLIP latent 技术相关	16
2022-08	Moderation Endpoint	免费审核文本内容，后续成为平台安全基建	API 开发者、平台治理团队	免费	GPT-based 分类器	17
2022-09	Whisper	多语言 ASR 与翻译	研究者、开发者	开源；后 API 商业化	大规模 68 万小时多语音频训练	18
2022-11	ChatGPT	对话式 AI 研究预览，后成为核心消费入口	普通消费者、知识工作者	首发免费	GPT-3.5 系列 + RLHF；运行于 Azure 超算	19
2023-02 / 2024-12	ChatGPT Plus / Pro	更高额度、更快响应、新功能优先；Pro 面向重度用户与研究级智能	个人付费用户	Plus $20/月；Pro $200/月	ChatGPT 订阅层，而非独立模型	20
2023-03	GPT-4	多模态输入（文本+图像）、更强推理与专业考试表现	开发者、企业、专业用户	ChatGPT Plus / API	GPT 系列里程碑；后衍生 Turbo、4o、4.1、4.5	21
2023-03 / 2023-06	Plugins / Function Calling	让 ChatGPT 访问第三方服务；让开发者定义函数、让模型调用工具	ChatGPT 用户、开发者	插件生态；API 功能按量计费	从“聊天”向“工具调用”过渡	22
2023-08 / 2023-09	ChatGPT Enterprise / 语音与图像	企业级安全与隐私、长上下文、分析能力；ChatGPT 看图听说	大企业；移动端用户	企业定制价格	Enterprise 基于 GPT-4；语音/视觉为后续 4o 与 Realtime 铺路	23
2023-11	DevDay 组合发布	GPT-4 Turbo、Assistants API、TTS、DALL · E 3 API、多模态平台能力、GPTs	开发者、构建者、普通创作者	API 按量；GPTs 面向 ChatGPT 用户	从单模型 API 走向“可调用工具的助理/应用”	24
2024-01 / 2025-08更名	GPT Store / ChatGPT Team→Business	GPT 分发商店；团队协作工作区	创作者、小团队、中小企业	Team/Business 订阅；商店分发	GPT Builder + Workspace	25
2024-04 / 2024-05	无登录使用 ChatGPT / GPT-4o / ChatGPT Edu	降低使用门槛；统一文本、视觉、音频的 omni 模型；高校版 ChatGPT	大众、开发者、大学	ChatGPT 免费层；Edu 面向机构销售	GPT-4o 支持实时多模态，平均音频响应约 320ms，API 更便宜	26
2024-07 / 2024-08 / 2024-10	GPT-4o mini / Structured Outputs / Realtime API	低成本小模型；严格 JSON schema 输出；低时延语音多模态接口	开发者、企业应用	API 按量计费	4o 家族：面向工具调用、成本优化与语音化应用	27
2024-07 / 2024-10	SearchGPT Prototype / ChatGPT Search / Canvas	AI 搜索原型后并入 ChatGPT；带来源的联网回答；独立编辑式协作界面	普通用户、研究用户、写作/编码用户	包含于 ChatGPT 各层计划	Search 成为 ChatGPT 内生工具；Canvas 强化工作流	28
2024-08 / 2024-09 / 2024-12	GPT-4o Fine-tuning / Multimodal Moderation / Sora	4o 微调；文本+图像审核；文本/图像/视频生视频	开发者、平台治理者、创作者	API 按量；Sora 包含于 ChatGPT 付费层/单独入口（公开产品页）	4o 家族；Sora 明确沿用 DALL · E 3 的 recaptioning 技术	29
2024-05 / 2025-03	Model Spec / Preparedness 体系升级	公开模型行为规范；前沿风险评估与部署门槛	开发者、研究者、监管与公众	非直接收费；属治理基础设施	安全与产品统一治理层	30
2025-01 / 2025-02	Operator / Deep Research / o3-mini / GPT-4.5	浏览器自动执行任务；多步互联网研究；轻量推理；更强非推理式对话模型	Pro/Plus 用户、开发者	包含于 ChatGPT/API 计划	Operator 的 CUA 建立在 GPT-4o 视觉 + RL；Deep Research 建立在 o 系列能力之上	31
2025-03 / 2025-04	Responses API / Agents SDK / 新音频模型 / GPT-4.1 / o3 / o4-mini / 新图像 API	统一 stateful 响应接口与内建工具；代理编排；新 STT/TTS；长上下文 4.1；推理模型全面支持工具；新图像生成 API	开发者、企业产品团队	API 按量；Batch/Flex/Priority 等差异化	平台从 Chat Completions 向 Responses/Agent 原语迁移	32

发布时间	产品/发布	核心能力	目标用户	变现/定价	技术谱系	官方来源
2025-05 / 2025-08 / 2025-10	Codex 云端编码代理 / GPT-5 / gpt-oss / Apps in ChatGPT / ChatGPT agent / gpt-realtime / Pulse / io 硬件计划	云端并行软件工程 agent；统一型旗舰 GPT-5；开放权重模型；ChatGPT 内应用平台；会研究也会行动的 agent；生产级语音；主动推送式资讯；硬件团队并入	开发者、重度工作流用户、生态伙伴	Codex 有独立工作/企业售卖；GPT-5 API；gpt-oss Apache 2.0；Apps 走平台分发；硬件未商业化	Codex、Apps、Agent 把 OpenAI 从模型公司推向“工作操作层”；io 代表迈向硬件接口	33
2026-01 至 2026-04	ChatGPT Go / Health / Education for Countries / Workspace agents / ChatGPT Images 2.0 / GPT-5.4 / GPT-5.5 / GPT-5.3-Codex / AgentKit	低价层订阅；健康专属体验；国家级教育计划；团队共享 agent；更强图像生成；最新旗舰模型与编码模型；可视化/工程化 agent 开发套件	大众市场、医疗健康用户、政府/教育、团队、开发者	Go $8/月（美国价，部分地区本地化），Health/国家教育/Workspace agent 多为机构或待定；GPT-5.5 API 公布 token 价格；其余多含于平台/企业售卖	明确转向“产品分层 + 行业垂直 + 组织级 agent”	34

影响力最高的十个产品

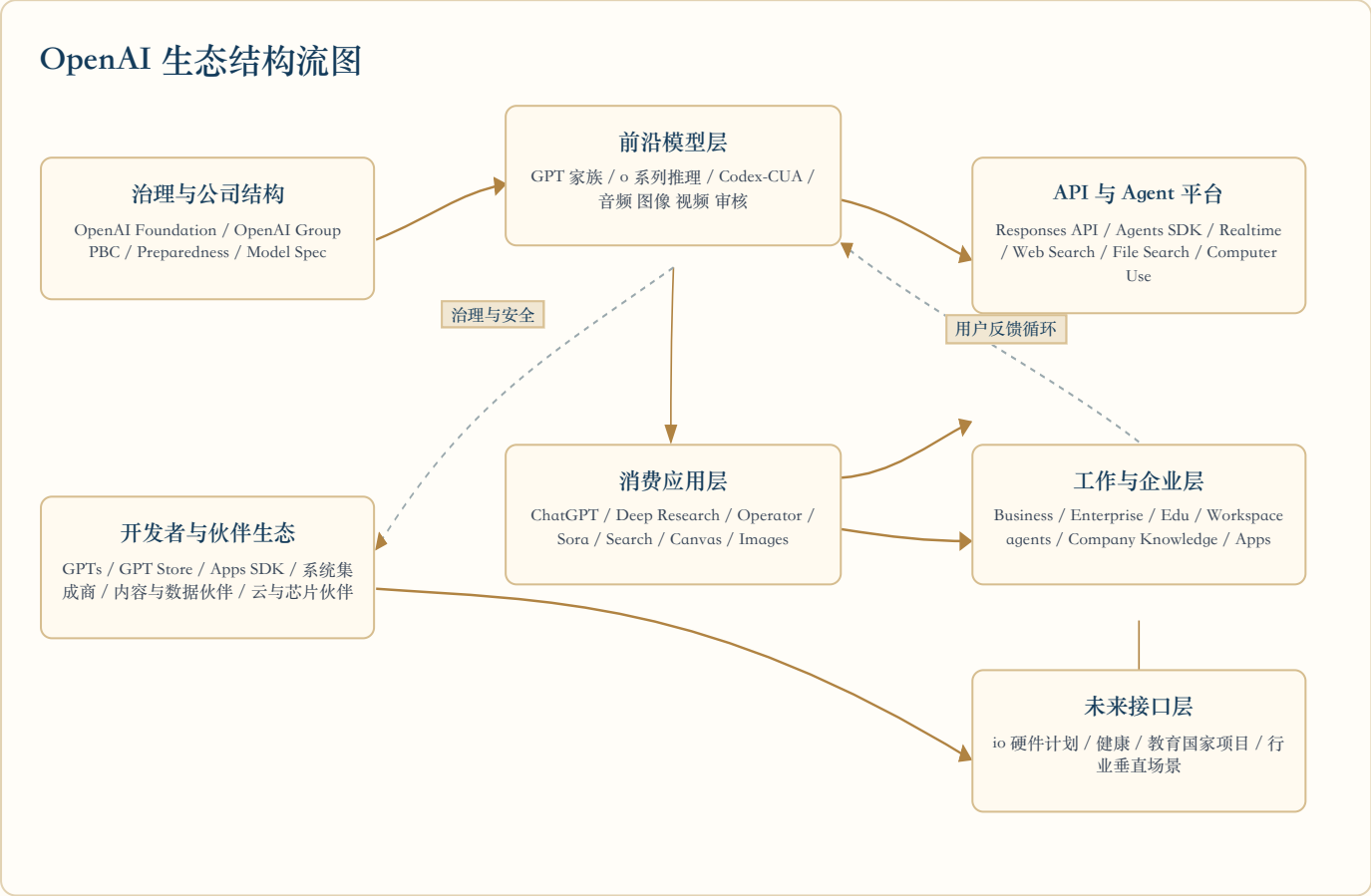
下面这张表把“影响力”理解为对 OpenAI 生态扩张、收入生成或使用规模的综合拉动，而不是单纯模型性能。凡遇到 OpenAI 未公开披露的产品级收入或精确用户数，均标注“未披露”。

产品	影响力判断	公开使用/收入信号	主要变现方式	说明
ChatGPT	极高	OpenAI 官方在 2025 年称 ChatGPT 已有超过 8 亿周活用户；2024 年 4 月时公开口径为超过 1 亿周活用户。35	Free/Go/Plus/Pro、广告、工作版导流	ChatGPT 是 OpenAI 的最大分发入口，也是几乎所有后续产品的流量源。36
API Platform	极高	2025 年 11 月 OpenAI 称直接服务超过 100 万企业客户；2026 年 4 月称 API 处理能力超 150 亿 token/分钟。37	按量计费、企业协议、定制模型	这是 OpenAI 从“模型能力”变现为“平台能力”的核心层。38
ChatGPT Business / Enterprise / Edu	极高	2026 年 2 月 OpenAI 称超过 900 万付费企业用户依赖 ChatGPT 工作；2026 年 4 月称企业收入占比超 40%。39	每席位订阅、企业合同	它把 ChatGPT 从消费品转成组织软件。40
Codex	很高	2026 年 4 月 OpenAI 称 Codex 超过 300 万周活开发者。41	企业版、Business Codex、平台增值	OpenAI 在“AI coding”赛道的关键产品，决定其能否守住开发者心智。42
GPT-4o 系列	很高	具体产品收入未披露；但 GPT-4o 被并入免费层，并成为多模态主力。43	API、ChatGPT 各层计划	4o 是 OpenAI 从“文本 GPT”迈向“原生多模态/实时交互”的转折点。44
GPT-5 系列	很高	产品级用户数未披露；但 GPT-5.5 价格与 1M 上下文等被正式公开。45	API、ChatGPT 高端层	GPT-5 之后，OpenAI 更强调“统一系统”而非碎片化模型 SKU。46
Responses API + Agents SDK	很高	官方称自 2025 年 3 月上线后，已有“数十万开发者”使用 Responses API，处理“数万亿 token”。47	API 按量、平台黏性	这是 OpenAI agent 时代的核心开发接口。48
Apps in ChatGPT / GPTs / GPT Store	很高	Apps SDK 面向开发者的官方卖点是可触达“8 亿+ ChatGPT 用户”。49	平台分发、未来广告/交易/生态抽成潜力	这让 OpenAI 不只是“卖模型”，而是“卖注意力和入口”。50
Deep Research	高	公开用户数未披露；但已扩展到 Plus/Team/Enterprise/Edu/Free，并有明确额度层级。51	作为订阅层高价值功能	它显著提高了 ChatGPT 对高价值知识工作的渗透率。52
Sora	高	用户数与收入未披露。53	订阅拉动、创意生态	Sora 让 OpenAI 从文本/图像进入视频与“世界模拟”叙事。54

生态位与竞争格局

OpenAI生态图

下图概括的是 OpenAI 在价值链中的位置，而不是组织架构图。它同时覆盖研究、平台、应用、企业、内容、伙伴和未来硬件。各层关系来自 OpenAI 的产品页、企业页、合作公告与公司结构披露。[55]



OpenAI在生态中的角色

从角色上看，OpenAI 至少同时扮演五种身份。它首先仍是研究机构：OpenAI 官方“About”页定义自己为 “AI research and deployment company”，且持续发布 CLIP、Whisper、GPT-4、o1、Sora、HealthBench、FrontierScience 等研究或系统卡。第二，它是 API 供应商：从 2020 年 OpenAI API 到今天的 Responses API、Realtime API、File Search、Web Search、Computer Use，平台已不再是单一文本补全，而是 agent 的后端。第三，它是消费应用开发者：ChatGPT 本身就是全球最大 AI 终端之一。第四，它是企业软件与工作流平台：Business、Enterprise、Edu、Workspace agents、Company Knowledge 等，把 OpenAI 拉进了知识工作操作层。第五，它还是规则与标准影响者：通过 Preparedness Framework、Model Spec、公开 System Card 与安全评估，OpenAI 把自己的部署方法论做成了行业参考系。[56]

这种生态位的特别之处在于，OpenAI 并未停留在“把最好模型卖给别人”。它一方面保留基础模型研发，另一方面主动下沉到应用层，用 ChatGPT 把用户入口抓在自己手里；同时又向上伸展到平台层，让第三方开发者和企业在其基础设施之上构建产品。换言之，OpenAI 同时占了 **模型层、工具层、平台层、应用层、企业层、分发层** 六个位置。正因如此，它更像 AWS + iOS App Store + Notion/Slack + Google Search + GitHub Copilot 的某种混合体，而不是传统单点 SaaS。[57]

价值链与网络效应

OpenAI 的第一条网络效应，来自 **消费者分发反哺企业与开发者**。官方一再强调，ChatGPT 的超大用户规模使企业导入的培训成本更低、试点更快、ROI 更容易兑现；Apps SDK 又直接把这种分发暴露给开发者，让开发者可以面向“8 亿+ ChatGPT 用户”构建应用。也就是说，ChatGPT 既是收入引擎，也是生态引擎。[58]

第二条网络效应，来自 **平台工具化**。Responses API 统一了状态管理、web/file/computer 等内建工具；Agents SDK 负责编排、评估与护栏；企业版则提供数据接入与管理控制。开发者越往 Responses/Agents 迁移，OpenAI 越像“agent 云平台”。这会提升切换成本，并把“模型性能差异”转化为“平台粘性差异”。[59]

第三条网络效应，来自 **内容和数据合作**。OpenAI 与 News Corp、Reddit、Stack Overflow、TIME、Hearst、Condé Nast、Axios、Axel Springer 等达成内容或 API 合作；这些合作一方面改善 ChatGPT/Search 的内容新鲜度与可信来源供给，另一方面缓解训练与展示环节的版权与授权摩擦。它因此不只是“抓取网络”，而是试图把优质内容方纳入生态治理。[60]

相对主要对手的竞争位置

与 Google 相比，OpenAI 的强项是产品定义速度、面向大众的单一 AI 入口，以及围绕 ChatGPT 形成的高强度品牌与使用习惯；Google 的强项则是把 Gemini 深度嵌入 Workspace，并在 2026 年推出 Gemini Enterprise Agent Platform，把模型、构建、治理、企业 DevOps 与云基础设施整合得更强。前者强在“通用 AI 前台 + agent 原生体验”，后者强在“企业套件 + 云平台 + TPU/搜索资产”。[61]

与 Anthropic 相比，OpenAI 仍拥有更强的消费端分发，但 Anthropic 在企业和编码工作流上的压力明显增大。Anthropic 官方把 Claude Code 和 MCP 生态放在很核心的位置，而 Ramp 的 2026 年 4 月企业支出样本显示，Anthropic 一度以 34.4% 超过 OpenAI 的 32.3%，说明 OpenAI 在“工作流深度”上并非稳操胜券。OpenAI 的应对就是把 Codex、Workspace agents、Business Codex、Responses API/Agents SDK 做得更像“完整工作系统”。[62]

与 Meta 相比，OpenAI 的路线是封闭前沿模型 + 平台与应用变现；Meta 则以 Llama 4 等开放权重模型、低成本部署与“可微调、可蒸馏、可到处部署”为核心卖点。Meta 的开放策略会持续压缩 OpenAI 的“纯模型溢价”，迫使 OpenAI 把竞争焦点从模型分数，转到工具链、工作流、分发、企业治理与真正可执行的 agent 上。换言之，OpenAI 的护城河更像“操作系统”，Meta 的护城河更像“开放模型供给”。[63]

与 Microsoft 的关系则更复杂：Microsoft 一方面是 OpenAI 的关键伙伴，Azure OpenAI 与 Foundry 仍然是 OpenAI 进入大型企业的巨大放大器；另一方面，Microsoft 365 Copilot、Copilot Agents 与 Azure 自身的 agent 平台又让它在企业前台具有更强控制力。因此，OpenAI 与 Microsoft 既是深度共生，也在企业入口、收入分配和平台主导权上存在天然张力。[64]

总的来看，OpenAI 的相对优势是：**用户规模、品牌心智、产品速度、agent 原生化、从 consumer 到 enterprise 的闭环**。相对劣势是：**算力与资本消耗大、企业深度场景竞争激烈、开放模型对其 API 价格形成压制、以及对伙伴云基础设施和外部监管仍有依赖**。这些优劣势都不是“下一代模型更强一点”就能彻底改变的，而会越来越多地由平台与组织能力决定。[65]

商业、治理与监管

OpenAI 的收入结构已经明显多元化。消费者端有 Free、Go、Plus、Pro；其中 Go 官方价格为 8 美元/月（美国显示价格，部分市场本地化），Plus 为 20 美元/月，Pro 为 200 美元/月。企业端有 Business、Enterprise、Edu，以及面向开发团队的 Business Codex；开发者端则是 API 按调用与 token 计费，并叠加 Batch、Flex、Priority 等效率/时延分层。2026 年 1 月，OpenAI 还宣布将在美国对 ChatGPT Free 与 Go 层测试广告，但强调广告将与答案分离、不会影响回答，也不会出现在健康和政治等敏感话题旁边。[66]

企业与渠道合作是 OpenAI 商业模式的第二支柱。Microsoft 是最重要的结构性伙伴：2019 年 OpenAI 为了融资与扩算力创建 OpenAI LP；2023 年双方又宣布扩大合作，Azure 继续作为 OpenAI 的独家云提供方；到 2025 年，双方签署了新协议，为长期合作、责任 AI 与商业安排续约。Apple 则在 2024 年宣布把 ChatGPT 集成到 iOS、iPadOS 和 macOS 体验中。除此之外，OpenAI 还与 Cloudflare、Scale、Accenture/PwC/Infosys 等部署伙伴、Stack Overflow、Reddit 与大量新闻出版商建立合作，形成“模型—内容—分发—企业落地”的联合网络。[67]

治理方面，OpenAI 的路径也越来越清晰。2019 年，OpenAI 通过 capped-profit 的 OpenAI LP 解决“使命优先但需要巨额资本”的矛盾；2025 年 5 月董事会明确决定仍由非营利组织保留控制权，同时将营利实体改组为 PBC；到 2025 年 10 月，OpenAI 官方页面显示新的结构已完成：非营利组织更名为 OpenAI Foundation，营利实体更名为 OpenAI Group PBC，Foundation 继续通过特殊投票和治理权控制 Group，并持有约 26% 股权；微软约持 27%，其余由员工、前员工和投资者持有。OpenAI 将此定位为“让资本结构更常规，同时维持使命治理”的办法。[68]

但这种治理安排并不意味着争议已经消失。外部层面，OpenAI 一直面临监管和法律摩擦：2023 年 FTC 对 OpenAI 启动与消费者保护及数据风险相关的调查；2024 年 FTC 还对微软、OpenAI 等生成式 AI 投资合作关系展开反垄断问询；版权与训练数据争议

持续存在，《纽约时报》诉讼引发 OpenAI 官方公开回应；到 2026 年，欧盟与 OpenAI 就其模型和网络产品的合规、新网络风险工具开放等仍在持续接触。OpenAI 的实际策略是同时推进三件事：加强自我约束、争取官方合作接口、以及通过授权合作降低部分版权冲突。[69]

安全与伦理层面，OpenAI 已经把“安全体系产品化”。最直接的是 Moderation Endpoint 与后来的多模态 moderation model；更高一层的是 Preparedness Framework、System Card、Model Spec、Red Teaming Network、使用 GPT-4 参与内容政策与内容审核，以及面向公共事务的 2024 全球选举策略。这套体系的特点是：它不是只在研究论文里谈安全，而是把安全作为部署流程、文档、API、模型卡、治理委员会乃至外部合作机制的一部分。2026 年 4 月，OpenAI 还宣布 ChatGPT Enterprise 与 API Platform 达到 FedRAMP Moderate，说明其政府与合规市场也在系统推进。[70]

需要注意的是，收入数字并未完全由 OpenAI 官方按产品线分拆公布。公开可见的大多是公司层级的速度指标与媒体报道：例如 Reuters 报道 2025 年中 OpenAI 年化收入已达 100 亿美元、2025 年末年化收入超过 200 亿美元、到 2026 年 2 月末超过 250 亿美元；但这些数字不是经审计的产品别收入，也通常不细分到 ChatGPT、API、Enterprise、Codex 等单项。因此，对“哪一个产品最赚钱”的判断，仍应以“公开推断”而非“官方财报结论”来表述。[71]

技术架构与集成模式

从技术架构看，OpenAI 近几年的演进可以概括为三层。第一层是 **通用基础模型层**，包括 GPT-3、GPT-4、GPT-4o、GPT-4.1、GPT-4.5、GPT-5、GPT-5.4、GPT-5.5 等；这条线主要靠预训练规模、后训练和多模态融合推进。GPT-4o 的意义尤其大，因为它把文本、图像、音频、视频输入和文本/图像输出统一到“omni”叙事里，并把语音交互延迟压低到接近人类对话水平。第二层是 **推理/专用模型层**，包括 o1、o3-mini、o3、o4-mini、GPT-5.3-Codex、computer-use-preview、omni-moderation 等，它们不是追求覆盖一切，而是针对“更强思考”“更强编码/工具调用”“更强 UI 操作”“更强安全分类”做 specialization。第三层是 **模态侧模型层**，如 Whisper、gpt-4o-transcribe、gpt-4o-mini-tts、图像生成模型与 Sora。[72]

这些模型线的“技术谱系”也越来越清楚。InstructGPT 标志着 RLHF 在产品化中的中心地位；o1 系列明确把“训练时/测试时计算”与大规模强化学习作为新的扩展轴；Operator 背后的 CUA 则把 GPT-4o 的视觉理解与 RL 结合，用于 GUI 操作；Sora 明确写明采用了 DALL · E 3 的 recaptioning 技术；新音频模型则建立在 GPT-4o / GPT-4o-mini 架构之上，并通过蒸馏与 RL-heavy 训练提升识别与可控语音生成。也就是说，OpenAI 的技术并非一条直线，而是一个“通用底座 + 对齐/强化学习 + 模态适配 + 场景专用化”的树状架构。[73]

在接口层面，OpenAI 正把平台从旧的 Completions / Chat Completions 迁移到 **Responses API**。官方文档把 Responses 定义为“最先进的响应接口”，支持状态化交互、文本和图像输入、JSON 输出、函数调用，以及 Web Search、File Search、Computer Use 等内建工具；OpenAI 还明确建议新项目优先使用 Responses，并将其视为 Chat Completions 的进化版本。对开发者来说，这意味着 OpenAI 不只是提供“模型推理接口”，而是在提供一套 agent 运行时。[74]

工具调用是这套运行时的核心。OpenAI 官方工具文档目前已把 web search、file search、computer use、tool search、remote MCP servers 等列为一等能力：模型可以先联网检索最新信息，再回答并给出处；可以对自有文件进行关键词+语义混合检索；可以通过 UI 理解和操作软件；还可以通过 function calling 或 MCP 连接第三方系统。值得注意的是，MCP 不是 OpenAI 发明的标准，它来自 Anthropic 阵营的开放协议叙事；但 OpenAI 选择在工具文档与 Apps 体系中兼容类似连接模式，说明它已经把“互操作性”视为扩大平台生态的必要条件。[75]

从典型集成模式看，OpenAI 基本提供了四种主流路线。其一是 **轻量生成式应用**：直接用 Responses API 或 Chat Completions，辅以 Structured Outputs 和函数调用，适合表单自动化、客服草稿、JSON 工作流。其二是 **RAG/知识系统**：用 File Search、Retrieval、Embeddings 和向量仓库，把企业私有文档接进去。其三是 **agent 型系统**：用 Responses API + web/file/computer use + Agents SDK，实现多步任务、工具选择、评估与观察；这正是 Deep Research、Operator、Codex 一类产品的底层范式。其四是 **实时语音/多模态接口**：使用 Realtime API 或新 STT/TTS 模型与 Agents SDK 集成，做电话坐席、实时语音助手、车载/移动交互。[76]

在时延与扩展性上，OpenAI 的做法也越来越“基础设施化”。GPT-4o 发布时就把“接近人类对话速度”作为卖点；Realtime API 则把语音交互抽象为平台接口；2026 年 Responses API 的 WebSocket 模式和 Codex 相关优化进一步压缩了往返开销、首 token 时间和 per-token 成本。与此同时，OpenAI 对外公开的平台吞吐量已达每分钟超过 150 亿 token，这意味着它的瓶颈正在从“能不能做”转向“如何在规模、时延、成本和安全之间保持平衡”。[77]

最后，OpenAI 的一个新信号是 **从纯闭源前沿模型，向“闭源旗舰 + 开放权重补集”演化**。2025 年发布的 gpt-oss-120b 和 gpt-oss-20b 采用 Apache 2.0 许可，强调低成本、工具使用能力与 agent 兼容性。这并不表示 OpenAI 放弃闭源优势，而更像是在开放模型

压力增强后，用开放权重覆盖边缘部署、本地推理和生态实验场景；旗舰层仍继续由 GPT-5.x、o 系列、Codex 等维持高端溢价。[78]

风险、缺口与战略机会

OpenAI 最主要的风险并不是“有竞争对手”，而是 **它正在同时经营三件昂贵而复杂的事情：前沿研究、全球应用分发、企业平台基础设施**。这会带来极高的资本开支、研发组织难度和治理复杂性。OpenAI 自身在结构说明里反复强调，未来需要“数千亿美元乃至数万亿美元”的资本；媒体报道的年化收入虽然增长很快，但也伴随着巨额算力投入和融资需求。对它来说，持续融资能力和计算供给，本身就是产品竞争力的一部分。[79]

第二个风险是 **企业护城河尚未完全锁定**。OpenAI 在消费端领先，但企业端同时面临 Anthropic 的编码/协议化优势、Google 的 Workspace/Cloud 套装优势，以及 Microsoft 的既有企业分发。尤其是当 agent 产品进入“真实工作流执行”阶段后，客户未必只看模型 benchmark，更看权限治理、审计、合规、系统接入、可控性与总体拥有成本。OpenAI 若不能把 Codex、Workspace agents、Business/Enterprise 工作区真正做成企业生产系统，就可能在利润最高的 B2B 市场上被削弱。[80]

第三个风险是 **内容、版权与信任**。OpenAI 已通过 publisher 授权、新闻合作与 opt-out 机制缓和版权摩擦，但纠纷仍未结束。另一方面，随着模型更像 agent、能主动执行任务，风险从“回答错了”升级为“执行错了、越权了、被注入了、说服过头了”。OpenAI 因此才会持续强化 Model Spec、Preparedness Framework、system cards、red teaming 和 computer-use 安全栈；但这说明风险不会因产品成熟而消失，反而会因产品可执行性增强而放大。[81]

第四个风险是 **开放模型与价格压缩**。Meta Llama、以及更广泛的开放模型浪潮，会持续压缩“只卖推理 token”的毛利空间。OpenAI 已经在用 GPT-4o mini、o4-mini、gpt-oss、Batch/Flex、Business Codex 等方式应对：低价模型做流量，旗舰模型做溢价，平台和工作流做黏性。这条路理论上合理，但前提是 OpenAI 能把“模型供应商”升级成“默认 AI 工作平台”。否则开放模型一旦在某些垂直场景足够好，OpenAI 的中端 API 价值就会被替代。[82]

但机会也同样清晰。首先，OpenAI 已拥有极难复制的 **用户规模 + 品牌 + 开发者平台 + 企业进入路径** 的组合，这意味着它最有机会成为“agent 时代的默认前台”。其次，Health、Education for Countries、FrontierScience、Workspace agents 这类产品表明它正向高价值垂直行业深入；如果这些场景跑通，OpenAI 的定位会从“通用 AI 工具”提升为“行业智力基础设施”。再次，Apps in ChatGPT、GPT Store、remote tools/MCP、Codex 插件目录等动作说明它在摸索“平台抽成”和“生态分发”路径。一旦交易、广告、订阅、企业席位、API 调用、内容授权与硬件叠加到同一生态里，OpenAI 的商业形态会越来越接近一个 AI 平台公司，而不只是模型公司。[83]

从战略机会看，我认为 OpenAI 未来最重要的三件事是：一，把 ChatGPT 从“最好用的 AI 聊天界面”升级成“个人与组织的默认 AI 工作台”；二，把 Responses API / Agents SDK / Codex / Apps 进一步统一成一个清晰的 agent 平台叙事；三，把安全、合规、数据接入和行业化能力做成比模型本身更难替代的系统能力。如果这三件事成功，OpenAI 的生态位会稳定在“AI 操作层”；如果失败，它就更可能退回成一家高端模型供应商，在价格和成本之间长期承压。以上判断属于基于现有产品与商业动作的推论。[84]

开放问题与局限

这份报告尽量优先使用 OpenAI 官方来源，但仍存在几个不可避免的局限。其一，OpenAI 未公开披露按产品线拆分的收入、利润和使用量，因此诸如“ChatGPT 比 API 更赚钱”之类表述只能做方向性推断，不能当作财务结论。其二，部分当前价格页对精确数字抓取不完整，因此我对现价只引用了官方明确披露或在官方搜摘中可见的价格；未能稳定抓取到的项目一律写作“未披露/定制/按量”。其三，本报告的“产品目录”按“重大发布”口径整理，不等于每个小版本、每次模型替换或每项帮助中心功能更新的穷尽清单。其四，2026 年非常新的公告（例如广告、监管、融资与行业部署）有时同时来自官方与媒体，若官方未给出同等细节，我已尽量用“据 Reuters 报道”这类表述降低误导风险。[85]