人工智能行业专题（12）

AI Agent开发平台、模型、应用现状与发展趋势

行业研究 · 行业专题

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

摘要

国信证券 GUOSEN SECURITIES

- AI Agent是突破指令执行的智能实体，核心是具备代理权，可主动感知、自主决策并执行复杂任务，区别于LLM（知识输出）和传统自动化（规则执行）。在AGI分级中，Agent处于L3智能体阶段，性能等效90%熟练成年人。
- **AI背景下AI Infra(AI基础设施)层面临重构，客户将重新选择云/AI平台，云厂加大布局AI/Agent平台瓜分新市场：**微软聚焦B端基础设施，是市场上模型支持最全面的平台，工具链和生态整合全面，安全与稳定性强，云收入增速显著领先；谷歌依托 AI Studio兼顾B/C端多场景，多模态强但生态不成熟、市占率较低；亚马逊依托AWS服务中小企业为主，侧重算力销售与便捷部署但工具链分散。国内字节扣子覆盖全场景，开发者与智能体数量领先；阿里百炼主攻B端全行业、服务30余万企业客户，MCP工具链和开源生态丰富。
- **海外模型呈现差异化发展、国内模型层并未拉开显著差异，2025年初伴随深度推理与长上下文模型发布推动Tokens调用量快速提升，推理消耗仍以各家内部场景为主。**根据Openrouter数据谷歌Gemini（长上下文+多模态）、Anthropic Claude（编程等严谨场景）占API市场半壁江山，国内DeepSeek、阿里Qwen系列份额稳步提升。谷歌2025年7月调用量980万亿（较去年增100倍），其中内部需求占比高达97%，AI推理需求已成为TPU发展的核心动力。国内字节豆包2025年5月日均Tokens 16.4万亿（增137倍），内部占比超80%。
- **应用C端重磅搜索产品主要依赖模型能力与生态导流拉开差距，图像与编程类产品发展迅猛。我们判断应用越偏向垂类，则技术门槛越低、产品理解要求越高、竞争越激烈，商业化闭环越容易。硬件端短期以AI手机/PC为核心、长期向多终端无感交互演进。**海外ChatGPT（MAU预计年底超10亿）、Gemini（MAU 4.5亿）领先，国内夸克、元宝依托生态导流。图像类Midjourney（ARR 5亿美元）、可灵（月收入1.5亿元），编程类Cursor（ARR 5亿美元）、GitHub Copilot（Web MAU1.2亿）表现突出。
- **应用B端Copilot/Agent产品形态丰富、持续渗透，机遇与挑战并存。**目前微软Copilot家族月活用户已超1亿，成为B端Agent/Copilot代表性产品，但企业落地仍面临幻觉、数据安全、成本高（Agent调用成本为LLM 15倍）等问题，B端SaaS面临技术革命下（软件制作成本逐步趋近于零）的产业范式重构。从行业来看酒店/餐饮/旅游行业GenAI投入最高。
- **Agent市场规模与发展预测：**根据IDC数据，全球AI IT支出2023-2028年CAGR 22.3%、其中GenAI达73.5%。CBINSIGHTS预计2032年AI Agent营收有望达1036亿美元（CAGR 44.9%）。根据Garnter与IDC,短期（2023-2025）GenAI嵌入现有应用，中期（2025-2027）Agent成核心组件，长期（2027+）自主代理网络主导业务，2035年后Agent将成为认知共生的人类助手、智能体即应用将成主流。

➢ **投资建议：伴随模型能力提升、AI Infra(AI基础设施)需求快速增长，推荐AI云平台厂商：微软（MSFT.O）、谷歌（GOOG.O）、亚马逊（AMZN.O）、阿里巴巴(9988.HK)、腾讯控股（0700.HK），AI芯片厂商英伟达（NVDA.O）与AI Data Infra数据服务厂商。**

➢ 风险提示：宏观经济波动风险、AI技术进展不及预期风险、行业竞争加剧风险、AI幻觉问题影响等。

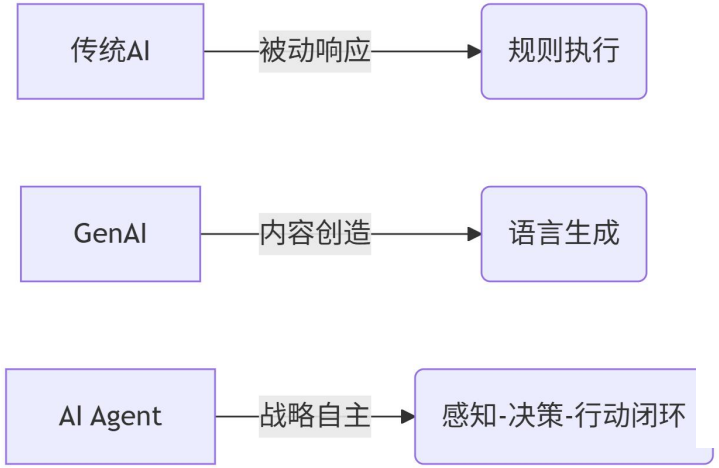
目录

AI Agent：重塑企业智能的核心引擎

国信证券 GUOSEN SECURITIES

- **定义：具备自主性、规划力与执行力的智能实体，超越“指令执行”进入“代理权”时代。**核心突破在于赋予“代理权”（Agency）➡ 主动感知环境、自主规划决策、执行复杂任务。
- **关键特性：**1）自主决策：主动感知环境、制定目标并采取行动；2）动态学习：通过记忆与经验积累实现持续优化；3）跨系统协作：调用工具、API及多Agent协同完成复杂任务。
- **核心模块：**1）感知层：多模态输入（文本/语音/图像）；2）记忆层：短期记忆（对话上下文）+ 长期记忆（知识库）；3）决策层：基于目标规划与强化学习的行动策略；4）执行层：工具调用（API）、跨系统协作（RAG技术）。
- **关键区别：**1）LLM≠Agent：LLM是“知识顾问”，Agent是“战略指挥官”；2）传统自动化：仅规则执行 vs Agent：端到端任务闭环。

图：Agent与传统AI的本质差异



资料来源：Bang Liu，《Advances and Challenges in Foundation Agents》，arXiv:2504.01990，国信证券经济研究所整理

图：基于 LLM 的代理中的推理模式比较

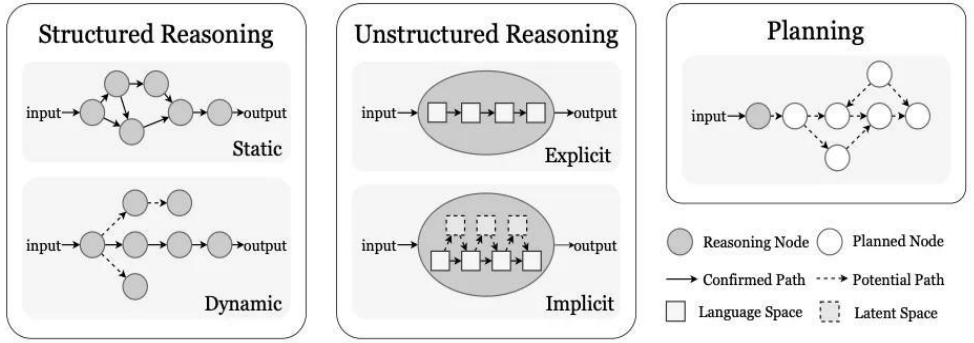


Figure 2.2: Comparison of reasoning paradigms in LLM-based agents.

资料来源：Bang Liu，《Advances and Challenges in Foundation Agents》，arXiv:2504.01990，国信证券经济研究所整理

图：Agent之间的知识共享和专业分工

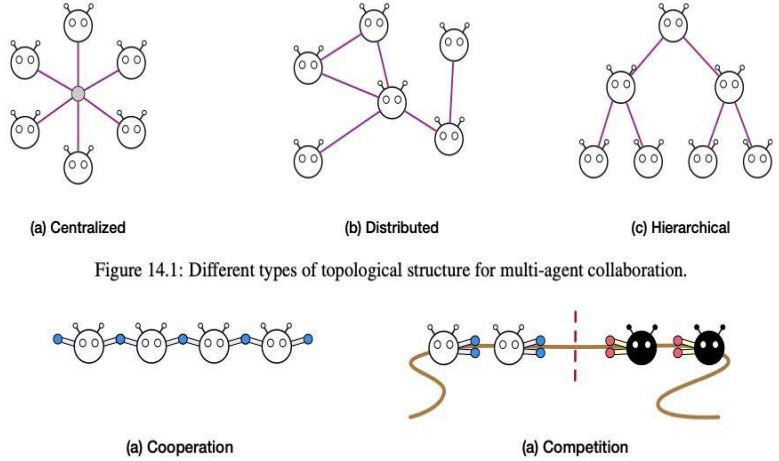


Figure 14.1: Different types of topological structure for multi-agent collaboration.

资料来源：Bang Liu，《Advances and Challenges in Foundation Agents》，arXiv:2504.01990，国信证券经济研究所整理

AGI的分级:Agent处于L3智能体阶段

- AI正站在一个关键新阶段。参考OpenAI对AI的五级分级，AI已不仅仅是能进行对话的聊天机器人（L1），而是逐步进化到智能体（L3）阶段：一个能思考、并能主动采取行动的AI系统。

图：AGI的分级

		性能	技能
L5 组织者 可以完成组织工作的AI	未来	L5：超越100%的人类表现 L4：相当于99%的熟练成年人	L5：能够执行复杂的组织任务。可以完成超越人类极限的工作，如预测未来、与动物交流等，达到科幻中的智能水平。 L4：能够自主进行发明和创新。不仅能完成既定任务，还能提出新的解决方案，帮助人类开拓新的技术和思路。
L4 创新者 能协助发明创造的AI			
L3 智能体 能思考，还可以采取行动的AI系统	目前阶段	L3：相当于90%的熟练成年人	L3：能为用户采取行动，完成一系列任务，不仅能理解和推理，还能实际执行具体操作。
L2 推理者 像人类一样能够解决问题的AI	过去	L2：相当于50%的熟练成年人 L1：相当于或略优于未受培训的人类	L2：可以进行基本的推理和问题解决能力。 L1：能够用人类语言，进行基本交流互动。
L1 聊天机器人 具有对话能力的AI			

资料来源：OpenAI对于AGI的分级、DeepMind对AGI的分级、国信证券经济研究所整理

代理的架构

- **代理核心构成：**

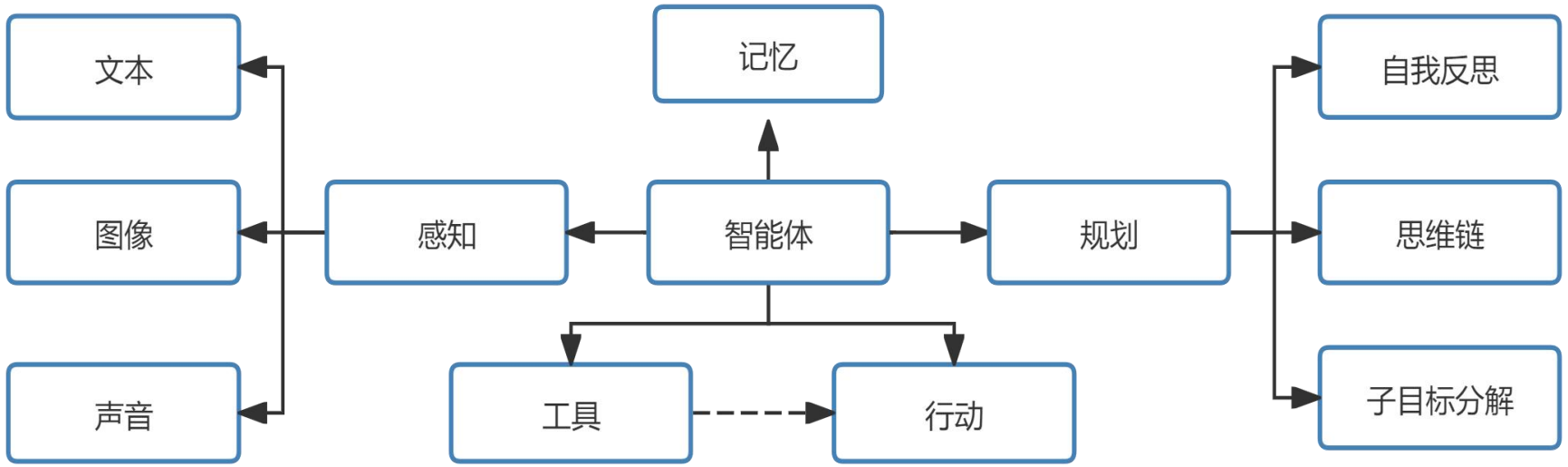
① 记忆（Memory）：存储、组织和检索短期上下文信息和长期知识的系统，支持自我反思优化。

② 感知（Perception）：多模态环境识别（视觉/听觉/传感器）。

③ 规划（Planning）：LLM驱动的目标分解与行动序列生成，复杂推理和决策，将任务分解为步骤，并根据需要进行调整。

④ 工具使用（ToolUsage）：调用API/代码/搜索等扩展能力，与外部应用程序、API、数据库和其他软件交互的集成功能。

图：AI代理架构图



资料来源：Bang Liu，《Advances and Challenges in Foundation Agents》，arXiv:2504.01990，国信证券经济研究所整理

Agent相关技术创新:MCP与A2A

- **MCP（Model Context Protocol模型上下文协议）- AI的“万用接口”，标准化AI模型（LLM）与外部数据源、工具间的交互方式**

➢ Anthropic在2024年11月推出并开源MCP协议，支持并行工具调用（如Web搜索、代码执行）和动态工作流编排。**核心组件**:

① MCP Server：通过MCP协议对外提供服务的轻量级应用，可提供工具执行、资源访问、预定义Prompt等资源。

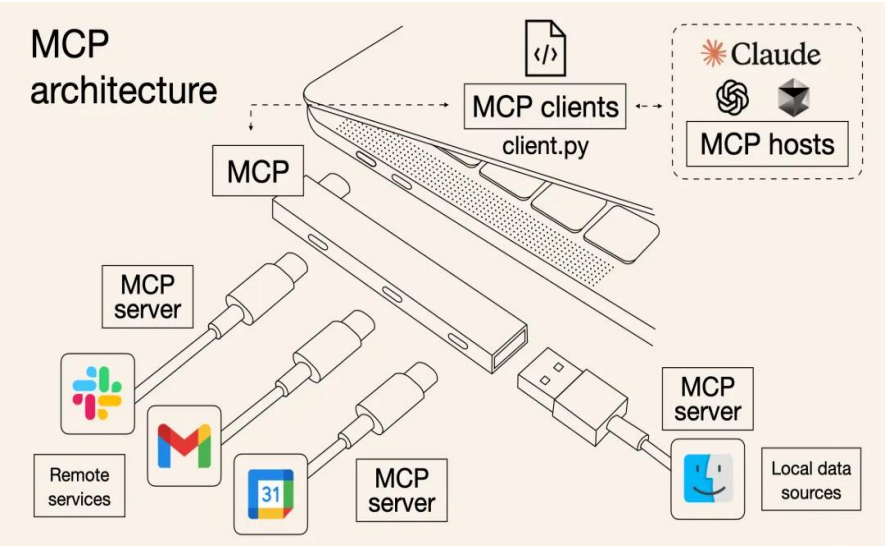
② MCP Client：充当LLM和MCP Server之间的桥梁，通过MCP Client SDK实现Host应用与MCP Server的交互。

③ MCP Host：承载AI模型的应用，如Claude Desktop和Cursor这类智能助手应用和IDE。

- **A2A(Agent-to-Agent)：让不同厂商或框架的AI代理彼此直接通信、协同**

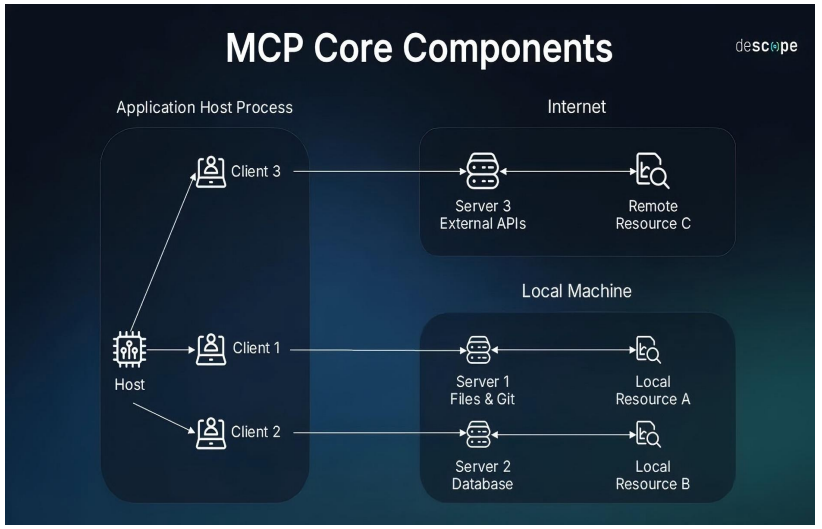
➢ 2025年4月Google推出的开放协议，得到了Salesforce、SAP等50多家科技公司的支持参与。基于其设定的Agent Card、Task、Message等概念和相关认证策略，如“智能体卡”概念用于描述智能体的身份、功能和服务接口。

图：MCP 架构



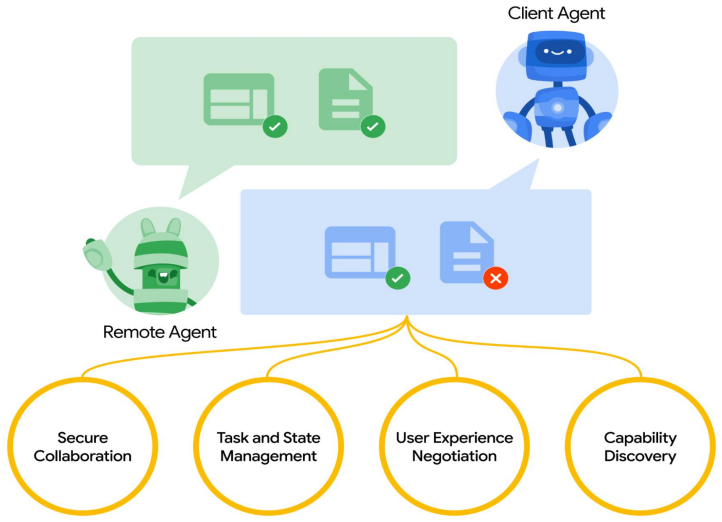
资料来源：Descope，国信证券经济研究所整理

图：MCP 核心组成部分



资料来源：Descope，国信证券经济研究所整理

图：A2A 工作原理



资料来源：谷歌公司官网，国信证券经济研究所整理

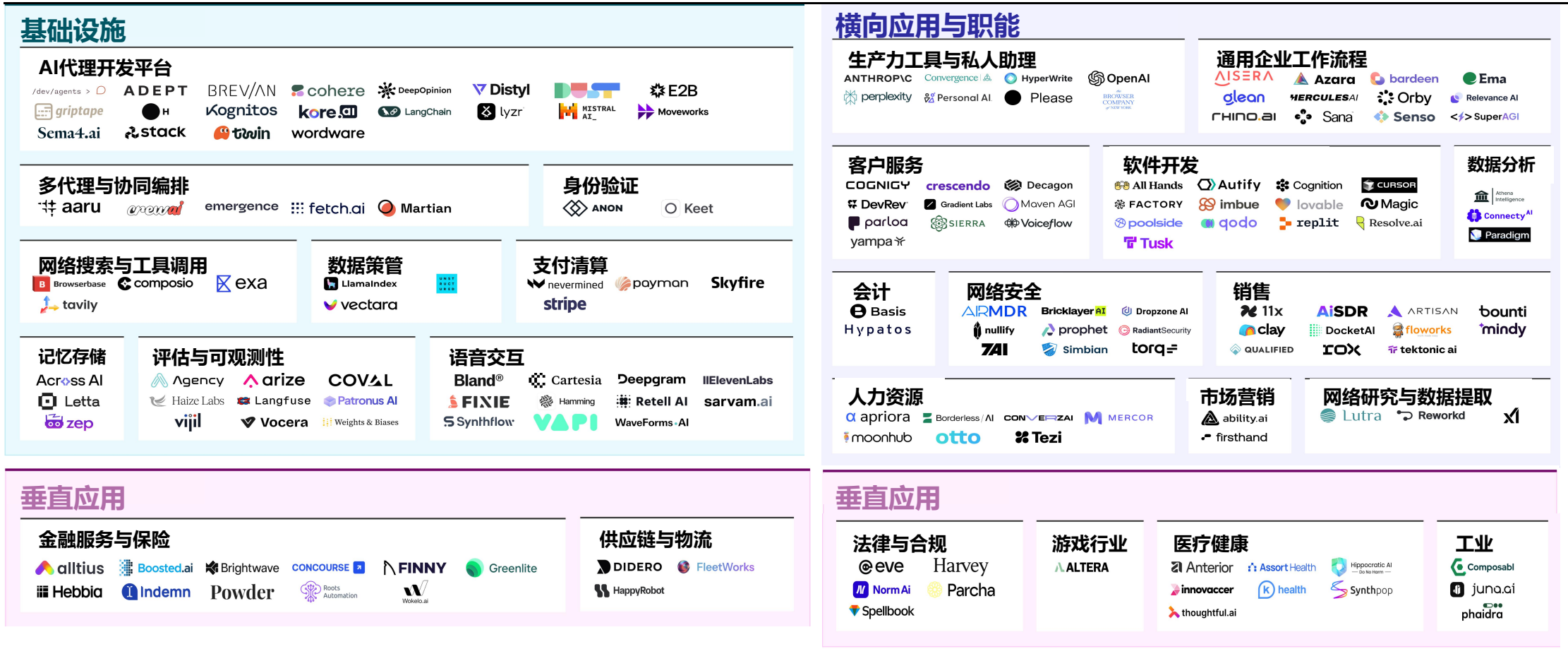
AI Agent市场图谱

① **基础设施类Agent**：聚焦底层支撑，涵盖开发平台、多代理协同、数据策管等多环节，为AI应用构建基础。

② **横向职能类Agent**：服务B/C端客户，从生产力工具到客服、人力等，跨行业适配，优化流程提效、助力日常运营。

③ **垂直应用类Agent** ：深耕金融、医疗等特定行业，贴合行业流程、法规与需求，深度融合专业知识，形成行业专属解决方案。

图：Agent市场图谱



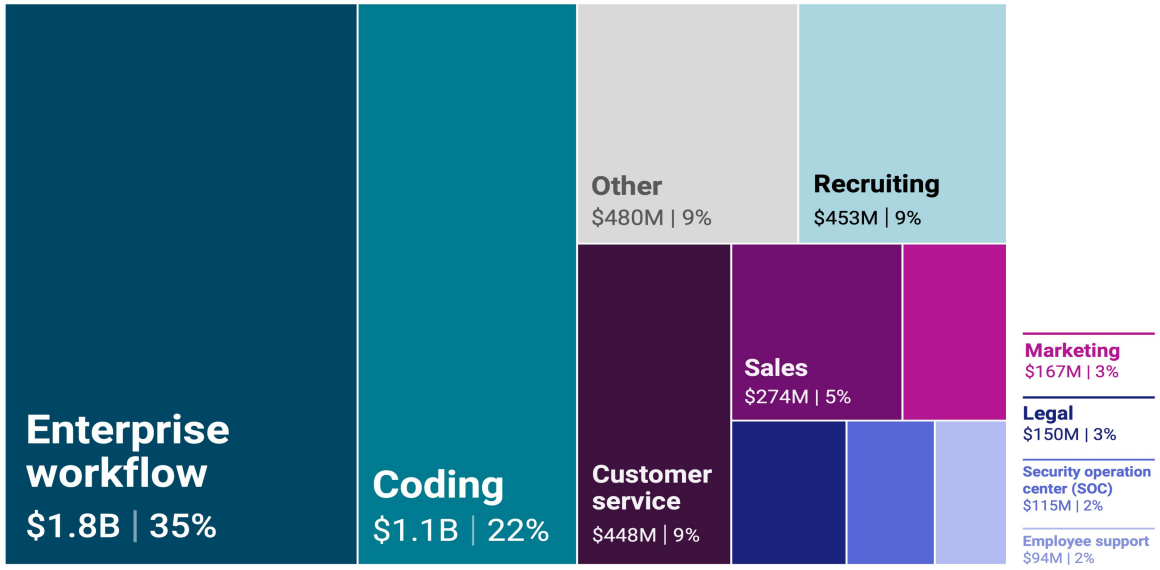
资料来源：cbinsights，国信证券经济研究所整理

Private companies only (as of 2/17/2025). Company inclusion based on CB Insights Mosaic health scores and funding recency. This map is not exhaustive of the space and categories are not mutually exclusive; companies are organized according to their primary focus.

AI Agent发展：增长迅猛，客户服务、软件开发占比较高

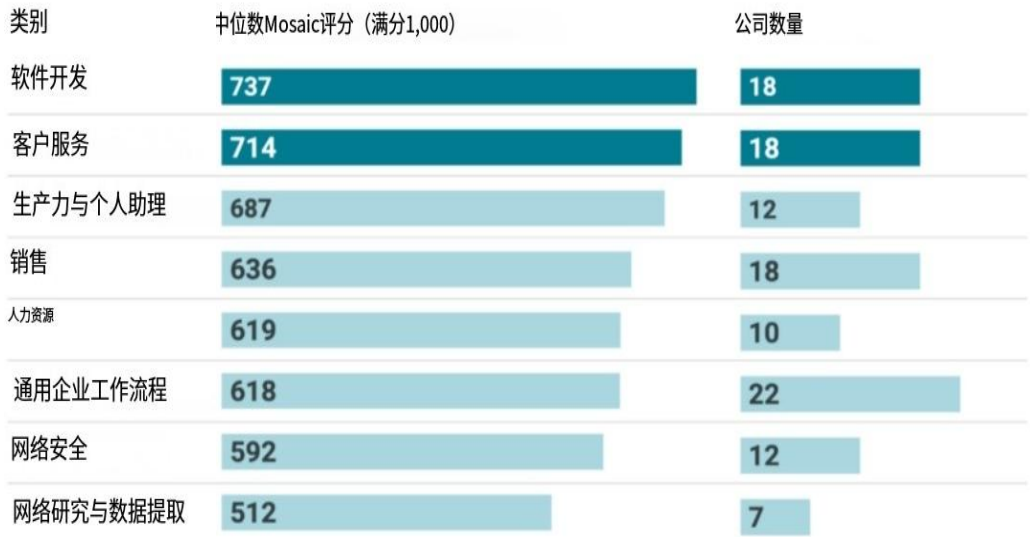
- **头部细分市场：**根据CBINSIGHTS企业工作流、编码两大领域2024年营收均超10亿美元，前者覆盖通用生产力、研究等场景，后者因AI编码工具爆发，半年内催生独角兽，速度达AI行业平均4倍。
- **核心驱动与玩家：**2024年科技巨头主导营收，微软Microsoft Copilot（2024年收入约8亿美元）、GitHub Copilot（2024年收入约6亿美元），总占整体市场超25%份额。初创企业增长迅猛，如Cursor年ARR从100万升至2亿美元。
- **垂类市场**:调查主要针对企业端，客户服务、软件开发为高潜力赛道，24年底调查显示，64家组织中2/3计划12个月内用AI代理支持客服。垂类AI Agent覆盖企业通用场景（人力、营销、安全运营等），较基础设施、垂直赛道商业落地更成熟。

图：2024年主要企业人工智能代理和Copilot市场的年度收入估计细分



资料来源：CBINSIGHTS(截止2024年底)，国信证券经济研究所整理

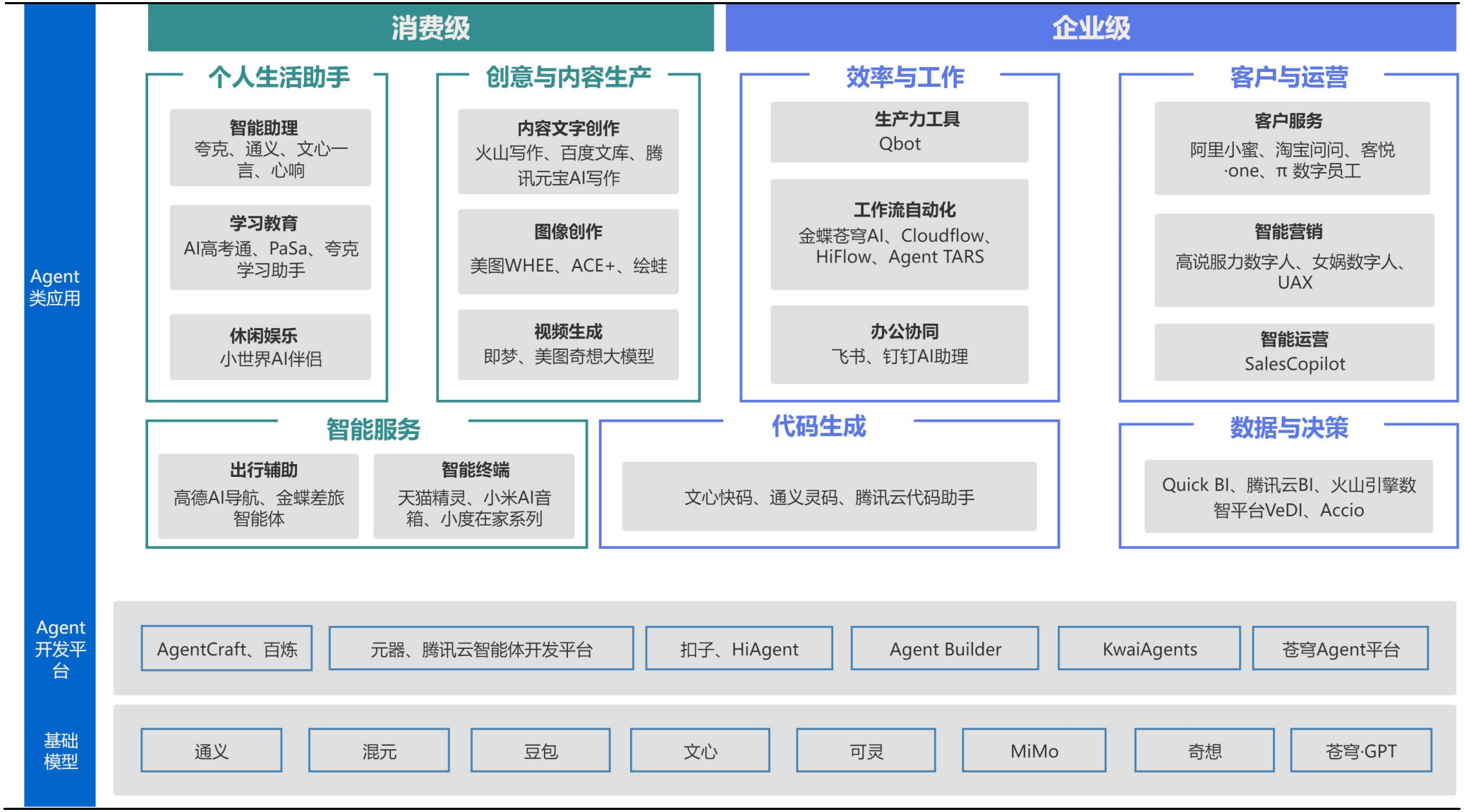
图：有前景的垂类Agent市场



资料来源：CBINSIGHTS(截止2025年2月19号)，国信证券经济研究所整理

国内AI Agent生态：阿里、腾讯、字节、百度、快手、小米、美图、金蝶

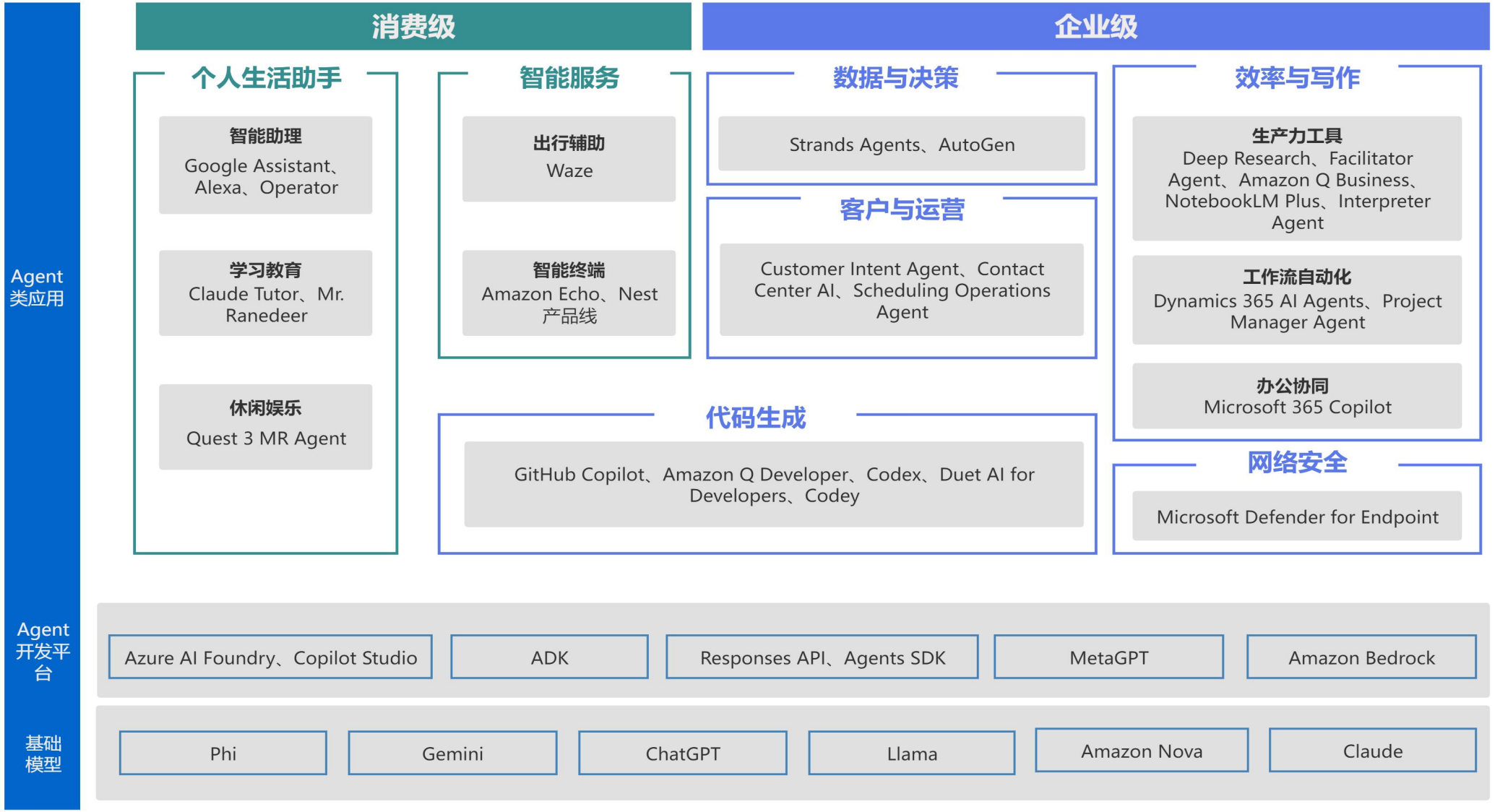
图：国内AI Agent生态



资料来源：各公司官网，国信证券经济研究所整理

海外AI Agent生态：微软、谷歌、OpenAI、META、亚马逊、Anthropic

图：海外AI Agent生态



资料来源：各公司官网，国信证券经济研究所整理

目录

Agent平台典型架构与核心环节

国信证券 GUOSEN SECURITIES

Agent是GenAI从 “概念验证” 迈向 “企业级应用” 的关键桥梁，其核心在于通过流程重构与数据整合释放AI的规模化价值。

图：Agent开发平台架构

底层基础设施

- **算力资源：**提供 CPU、GPU、TPU等计算资源，支撑Agent的模型推理、训练及数据处理（如分布式计算集群）。
- **存储系统**:结构化/非结构化数据存储（数据库）与缓存，存储Agent配置、用户信息、日志等。
- **网络与通信层：**API 网关用于统一管理外部系统与Agent的接口调用，消息队列支持Agent间、Agent与外部系统的通信，以及MCP、A2A等各类协议适配。

Agent核心模块

- **智能体管理：**Agent生命周期管理，创建、初始化、启动、销毁Agent实例，支持批量部署与版本控制。实时跟踪Agent运行状态（如资源占用、任务进度），提供监控面板。多Agent协作管理，协调分工交互。
- **感知决策与交互：**利用CV、NLP技术对各类输入感知处理并理解，提供各类大模型作为Agent决策推理的基础，利用API接口执行操作或输出生成。

支持服务模块

- **数据管理与处理**:数据采集，从内部系统、外部API、用户交互中收集数据。数据清洗、标注、特征提取、知识库管理。数据补充，外部与合成数据填补数据缺口。
- **安全与风控：**管理用户、Agent、外部系统的访问权限。加密存储敏感数据、控制数据传输加密。风险防控检测异常请求（如恶意攻击、高频调用），防止模型被滥用（如提示词注入防护）。

开发运维工具

- **开发者工具：**如低代码开发平台工具，调试工具支持Agent对话流程调试、模型输出可视化。SDK/API文档提供标准化接口，方便开发者集成Agent能力。
- **运维与监控：**日志系统收集全链路日志（请求、处理、错误），支持检索与分析。告警机制设置阈值（如响应超时、错误率）。监控资源瓶颈（CPU、内存、网络），自动扩缩容或优化资源分配。

扩展与生态

- **插件市场与标准适配层：**通过标准化接口、可插拔组件支持开发者根据需求自定义功能、集成外部资源，允许开发者开发“功能插件”（如特定领域插件、自定义工具调用），供用户按需安装。
- **外部协同与开发者生态**:通过工具链、合作机制、共享资源等吸引第三方参与者（开发者、企业、服务商），形成“平台-开发者-用户”的协同生态。整合外部成熟服务（如支付接口、地图服务、OCR识别），形成行业解决方案模板库，建立社区与知识共享平台。

资料来源：各公司官网，Bang Liu，《Advances and Challenges in Foundation Agents》，arXiv:2504.01990，国信证券经济研究所整理

Agent平台重要性：AI Agents由90%的软件工程和10%的AI组成

国信证券 GUOSEN SECURITIES

- 冰山上端如Perplexity、Cursor等知名的AI Agent产品是用户能直接接触到的表象。
- 但冰山水下部分支撑AI Agents的软件工程体系占比达90%，体现了AI Agent背后大量的工程化、系统化工作。而Agent平台正是把这些AI Infra和工具链集成、优化、整合的关键，复杂的工程链条包括：

① 前端开发（如 Streamlit、Flask等框架）

② 记忆模块（如 zep、memo等工具）

③ 认证系统（如 Auth0、Okta 等）

④ 各类工具（如 Google Search、DuckDuckGo 等）

⑤ 基础大模型（如 OpenAI、Gemini 等）

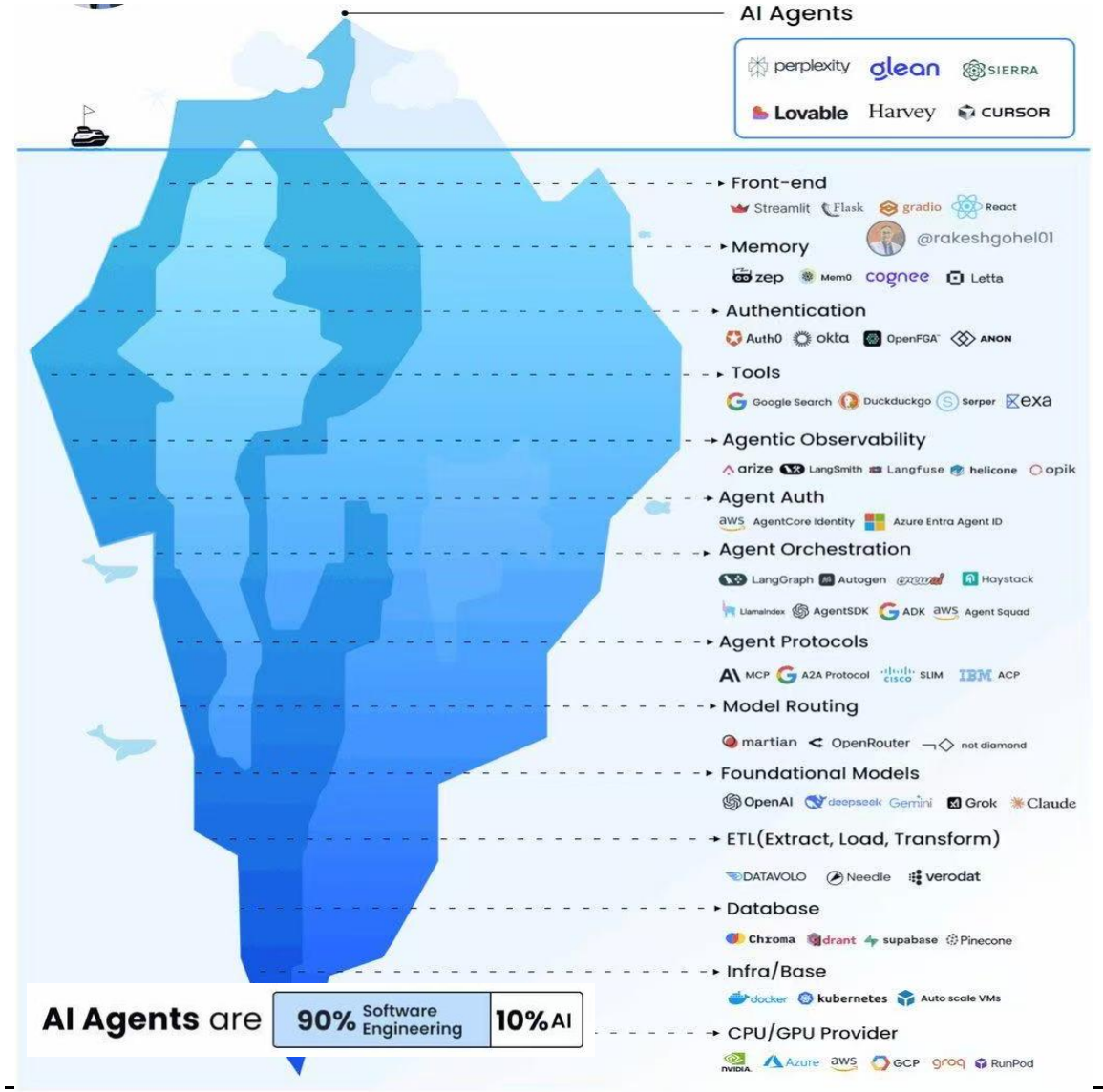
⑥ 数据处理（ETL）

⑦ 数据库（如 Chroma、Pinecone等）

⑧ 基础设施（如 Docker、Kubernetes 等）

⑨ 算力提供（如 NVIDIA、AWS 等）

图：AI Agents由90%的软件工程和10%的AI组成



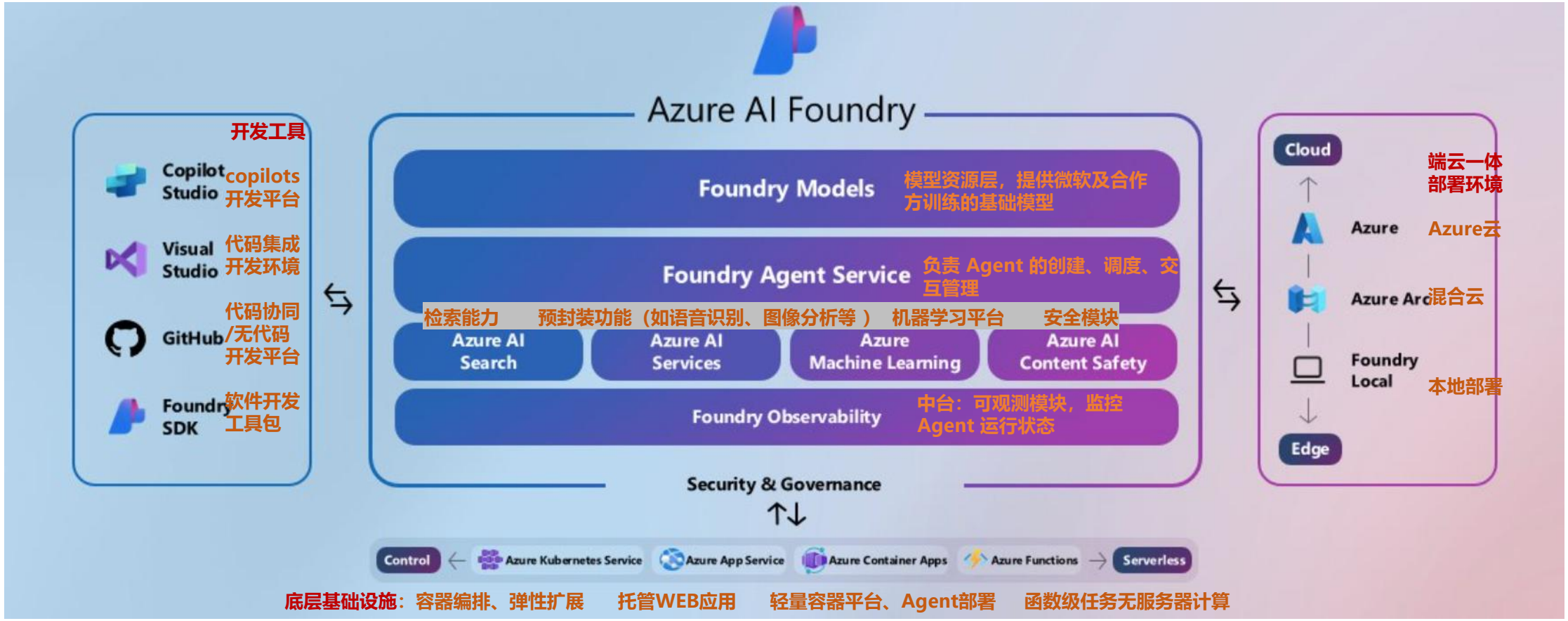
资料来源：Rakesh Gohel，国信证券经济研究所整理

Agent平台案例：微软Azure AI Foundry（Open Agentic web）

2025年微软Build大会题目开放代理式网络（Open agentic web），强调发布多项工具含 GitHub Copilot代理、Azure SRE代理，助力开发及运维。

Azure AI Foundry：构建企业级 AI 模型与智能体全生命周期管理体系。已被80%的财富500强企业使用，25Q2处理的tokens超500万亿，同比增长超7倍，agent service客户数达1.4万。集成Grok-3等1900+模型。

图：Azure AI Foundry 架构



资料来源：公司官网，国信证券经济研究所整理

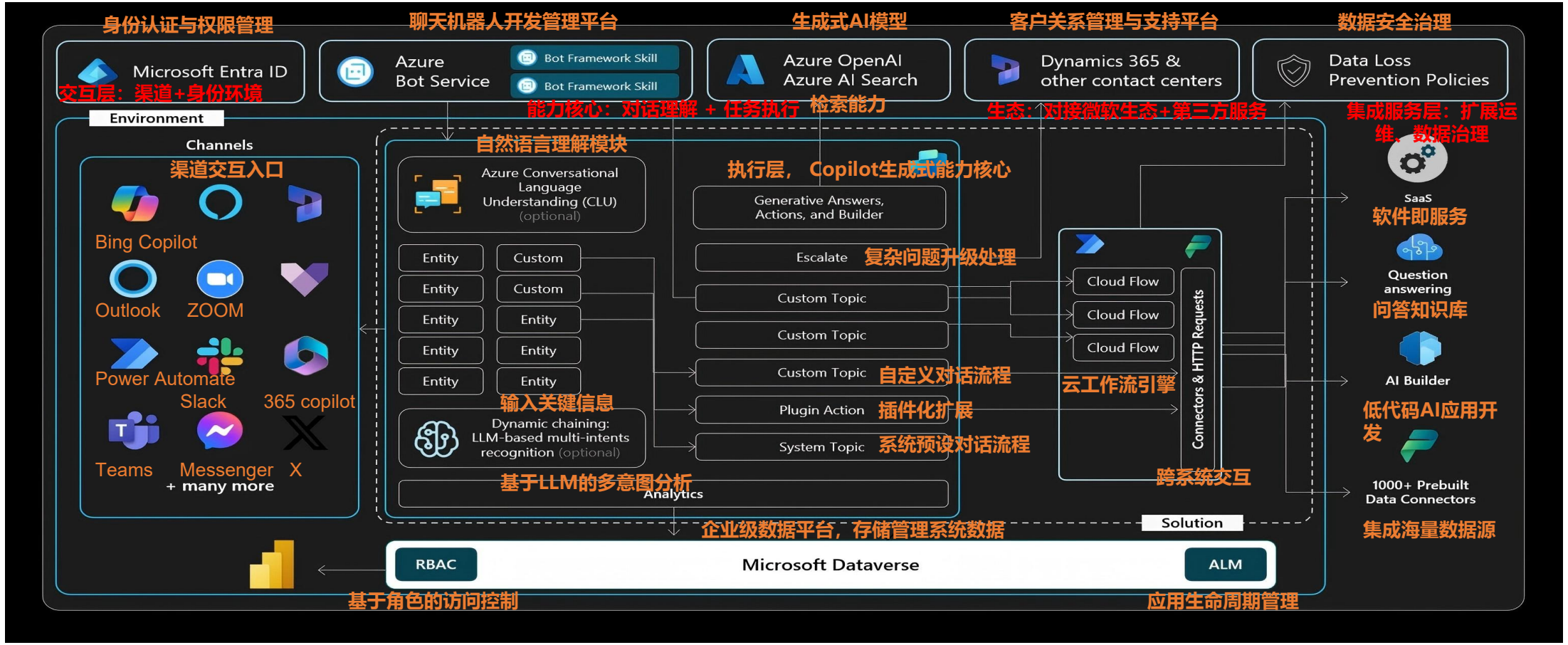
Agent平台案例：微软Copilot Studio

- **Copilot Studio：低代码构建复杂智能体工作流，实现跨角色任务自动化。**

① 低代码多Agent编排：支持通过组件连接Copilot、Azure或自建Agent，协同处理复杂任务（如法律文档生成+合规审查）。

② 协议支持：支持 MCP 协议，可调用符合开放标准的第三方智能体。

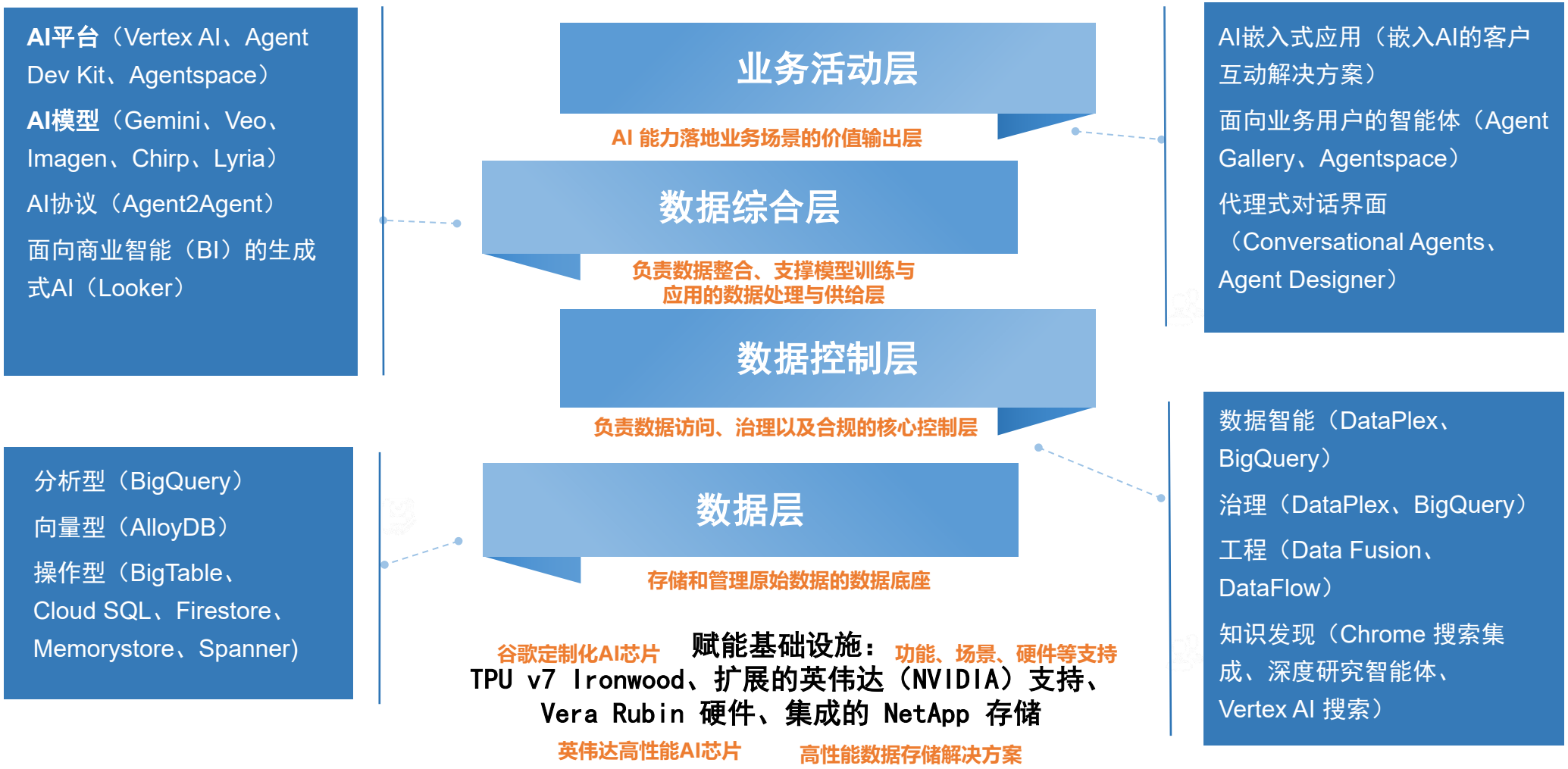
图：Copilot Studio（AI copilot的定制化平台）架构



资料来源：公司官网，国信证券经济研究所整理

Agent平台案例：谷歌数据与AI产品架构

图：谷歌AI平台架构



资料来源：公司官网，国信证券经济研究所整理

海外Agent平台布局：微软、谷歌、亚马逊对比

国信证券 GUOSEN SECURITIES

总结：微软聚焦B端基础设施，是市场上模型支持最全面的平台，工具链和生态整合全面，安全与稳定性强；谷歌依托 AI Studio兼顾B/C端多场景，多模态强但生态不成熟、市场占有率低；亚马逊/Anthropic依托AWS服务中小企业为主，侧重算力销售与便捷部署，Claude模型实用性强，但工具链分散。

图：海外云厂Agent平台对比

类别	微软	谷歌	亚马逊/Anthropic
产品定位	聚焦 “AI平台 + 云服务”，以 Azure AI Foundry 为核心，提供一站式Agent开发与部署生态，偏向 To B 场景	依托 AI Studio 构建 Agent 平台，兼顾 To B 与消费级场景，侧重多模态模型能力	依托 AWS 构建 Agent 部署平台，侧重算力销售与部署灵活性，服务中小企业为主
产品架构	采用Azure AI Foundry（云端）和Windows AI Foundry（本地）双模式，整合 LLM、Web Search、向量数据库，与Azure存储/ VM打通，Copilot Studio支持智能体开发	基于Gemini模型构建代理式架构，通过Agent2Agent协议推动跨平台Agent的互操作性，整合Web Search 等工具	依托MCP协议开发Claude工具链，提供VS Code插件等IDE深度集成方案
模型	市场上模型支持最全面的平台，支持OpenAI 模型、开源模型（如 DeepSeek）及未来可能接入claude，几乎无自研核心模型	以 Gemini 模型为核心，Gemini 2.5 Pro后能力达到业界认可	Bedrock 平台集成多厂商模型 + 自研 Nova（文本、图、视频生成）与 Anthropic合作密切（Claude模型）
用户相关	主要用户为企业客户（尤其是 AI 公司、需要工具链整合的企业）， 用户对 LLM 相关应用依赖性高（如 Outlook、Word 等）	用户包括开发者、企业及消费级用户，AI Studio平台活跃度低，市场信任度和占有率不足	主要用户为中小企业， 侧重满足企业对算力和灵活部署的需求
行业与场景	覆盖广泛 To B场景，包括企业办公、内部信息检索（Business Chat），销售、服务、财务、供应链等业务流程自动化	覆盖企业级模型开发、消费级办公（如 Notebook LM）、数据分析、网络安全及多模态场景（视觉 + 文本处理）等	聚焦中小企业AI部署场景，如模型快速上线、算力密集型任务支持等。Claude模型擅长金融、法律、编程等知识密集型且严谨性的场景
优势	工具链和生态整合全面，与云服务深度协同， 规模和市场份额大，安全与稳定性强	多模态模型能力强\迭代快，优于微软同类产品	算力资源丰富，部署灵活性高，适合中小企业快速落地，模型能力达到实用阈值
劣势	受安全领域投入分流影响AI应用进展，应用多模态能力较弱	平台生态成熟度不足，AI Studio 活跃度低，进入 AI 市场较晚，市场信任度和占有率待提升	工具链分散，生态整合能力弱于微软

资料来源：.各公司官网，彭博社，华尔街新闻，国信证券经济研究所整理

海外云厂25Q2云收入：微软云受AI驱动加速明显、势头领先

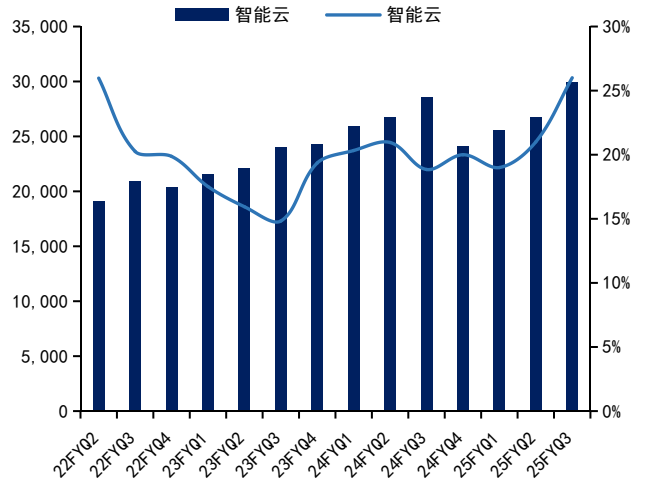
国信证券 GUOSEN SECURITIES

- **微软云**：云与Azure加速增长，本季度智能云收入为299亿美元，同比+26%，环比加速，其中Azure同比+39%，超业绩指引的34-35%。Azure年收入超过750亿美元，本季度未披露AI贡献Azure占比，上季度AI贡献Azure的16个点。尽管本季度新增数据中心容量，但需求仍高于供应，预26财年上半年仍有算力容量限制。

- **AWS（Amazon）**：25Q2 AWS业务收入309亿美元（同比+17.5%），目前AI相关收入继续保持三位数增速，数十亿美元规模。AWS目前仍处于供应能力不足，25Q2积压订单1950亿美元，同比+25%，原因最大的是电力限制，另外芯片和组件数量不足、芯片交付节奏延迟、服务器良率不达预期等，预计未来几个季度仍无法满足需求。

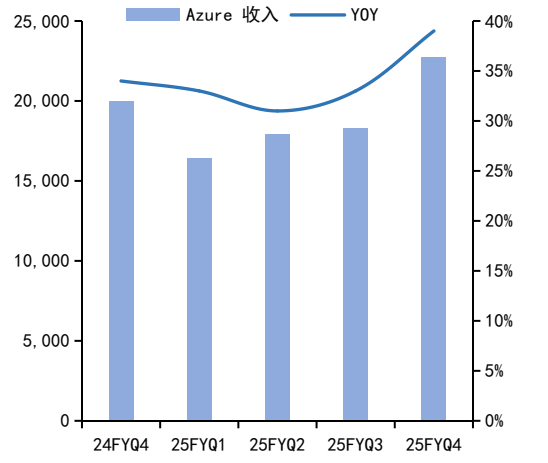
- **Google Cloud**：谷歌云收入136亿美元（同比+32%，环比+11%）。超过2.5亿美元的交易数量同比增长一倍，25H1签署的超过100万美元的交易数量与2024年全年持平。未完成订单在第二季度环比增长18%，同比增长38%，在本季度末达到1060亿美元。尽管公司加快了服务器的部署速度，但预计到26年供需环境仍然紧张。

图：智能云收入与增速变化（百万美元，%）



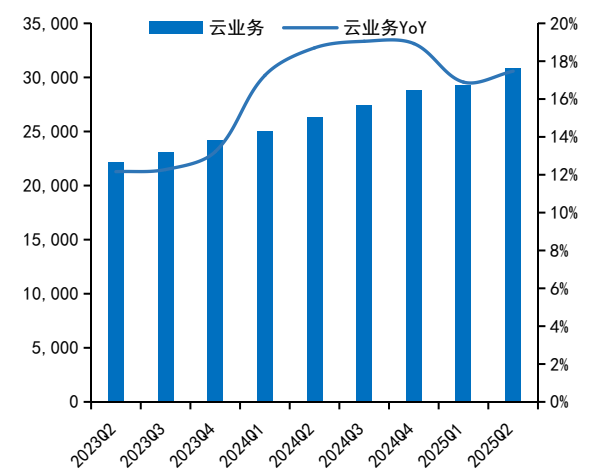
资料来源：公司财报、国信证券经济研究所整理

图：Azure收入变化（百万美元，%）



资料来源：公司财报、国信证券经济研究所整理

图：AWS收入与增速变化（百万美元，%）



资料来源：公司财报、国信证券经济研究所整理

图：谷歌云收入与增速变化（百万美元，%）

资料来源：公司财报、国信证券经济研究所整理

国内Agent平台布局：字节、阿里、腾讯对比

图：海外云厂Agent平台对比

	字节扣子	阿里百炼	腾讯元器
产品定位	做到全场景覆盖，目标拿下国内全部Agent市场，以流程引擎为核心，聚焦数据处理、流程自动化 》C端主推行业专家型Agent（如编程、旅游、律师），聚焦功能深度与用户体验，而非“全能型”服务。 》平台侧重优先提供解决方案，如视觉Agent方案，而非直接面向B/C端做产品，平台型产品已完成构建收费模式	偏向构建通用智能体，侧重B端企业级应用，全行业覆盖（如金融、政务、医疗等），适用复杂推理、跨行业生态集成，支持复杂业务场景（如电商售后、广告投放优化）	基于腾讯混元大模型的一站式智能体制作平台，主打轻量化低代码开发，聚焦社交（QQ / 微信客服）、游戏等垂类场景快速落地
核心功能	① 服务覆盖大型企业（私有化部署，起步价80多万+并提供公有云可选方案） ② 中间层面（HiAgent企业开发级平台、扣子平台分企业专业版/普通版/个人版） ③ C端用户（扣子空间、豆包专业专家Agent）	- 内部业务重构：整合淘宝、天猫、饿了么、飞猪、1688等电商板块，推动底层技术架构共用与数据互通（如用户画像、智能对话信息）。统一技术架构赋能核心业务（电商、阿里云、钉钉等），依托一整套技术栈与模型框架，实现人机交互模式升级。 - 外部应用赋能：通过阿里云+开源社区辐射B/C端产业，完善 “芯片层-模型层- Agent层-应用层” 的全栈生态。	主要依托自身社交、游戏生态进行Agent 部署 ① 双模式创建（提示词/可视化工作流）；②预置插件（位置服务、图像识别等）及MCP（如微信支付、快递100等）； ③支持QQ智能体 /微信客服发布，含分享权限管理
组织架构	由豆包系、火山引擎（HiAgent）、扣子团队及整体 Agent 平台支持团队提供技术支持，形成 “平台中心 + 多应用” 模式，另有内部团队负责豆包专业 Agent 自研	通义实验室负责底座模型及智能体底层框架输出与整合，业务部门基于底层技术构建场景化 Agent（如电商、阿里云、钉钉等）	依托腾讯混元大模型团队+微信生态部门，未明确独立 Agent 部门
用户相关	开发者约170 万，扣子平台有效智能体超200 万，日新增智能体 4000个（文心智能体日增约1000个） B端客户量10 万级别，今年目标20-30万，AI解决方案客户预计下半年达到阿里百链平台客户数的1/2-2/3。外部调用集中在教育、工具、服务类功能	目前服务30余万企业客户，预计今年能到40-50万，繁花计划筛选15-20家头部标杆客户（单客户体量3000万-5000万，部分过亿），优先迁移云上客户至AI平台，通过头部示范拉动中长尾客户重构业务	日新增智能体约700-800个，客户量少于字节扣子和阿里百炼，早期以小B商家和社交、游戏领域客户为主，通过亿级 token免费额度吸引尝鲜用
行业与场景	扣子Agent贡献量前三为教育、生活服务、效率工具，调用量前三位为情感聊天、教育、生活服务（旅游、交通、穿搭等）；	全行业覆盖，核心五大优势行业贡献超80%收入，包括互联网、金融、汽车、医疗健康、教育	聚焦社交（情感陪伴、社群运营）、游戏（剧情互动），逐步拓展电商、工具类场景
优势	生态融合强（与抖音、微信等打通），低代码平台吸引大量开发者，智能体数量领先，场景化模型功能丰富（如生成博客对话、视频理解），与火山引擎形成生态联动增长	全行业覆盖能力强，MCP工具链和开源生态丰富（近4000个MCP 服务，年底预计破万），底座模型（QWEN3）接近海外先进水平，客户数量多且覆盖核心行业	在社交、游戏领域有天然生态积累
劣势	行业覆盖较窄，工具链在复杂跨行业集成上弱于阿里	C端布局相对薄弱，智能体推广处于早期（Q2刚推广）	底层模型能力一般，功能、复杂场景支撑弱、依赖插件集成

资料来源：.各公司官网，各公司公众号，彭博社，华尔街新闻，国信证券经济研究所整理

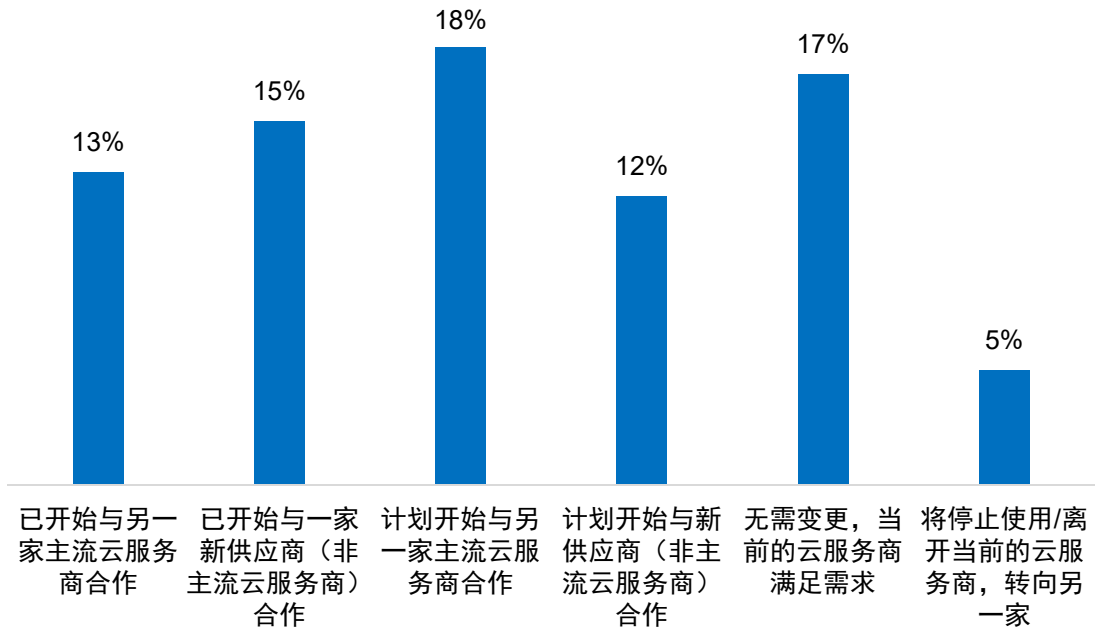
PaaS/Agent平台变迁：客户将根据平台AI/工具部署能力重新选择

国信证券 GUOSEN SECURITIES

PaaS/应用平台供应商面临新的市场份额瓜分机会：根据IDC调查，将有70%的受访企业将更换或新增云/AI平台供应商。仅仅有17%的受访客户认为当前的云提供商可以满足他们的AI/ML/GenAI需求。28%的受访企业已经被迫更换提供商，而另有42%的受访企业计划离开当前的提供商或加入云提供商组合。

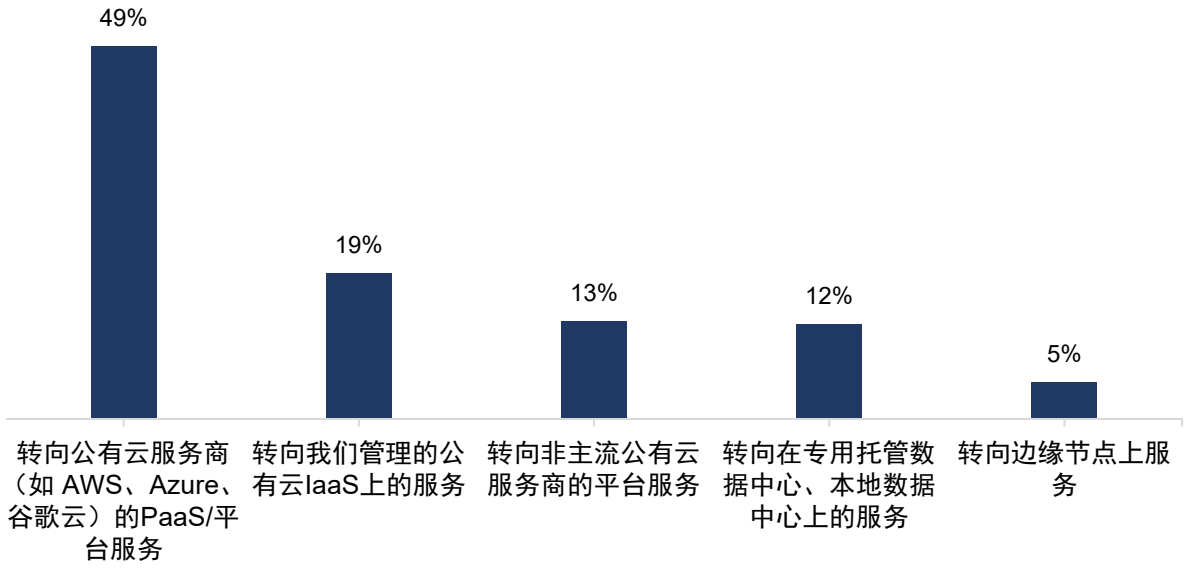
近一半的企业依赖公有云PaaS服务商进行AI方案落地：通过公有云实现对AI模型进行API调用，是企业部署AI应用的主要和关键方式。其他自建、本地托管或边缘节点上使用AI模型占比不到20%。

图：整合AI/ML及GenAI解决方案对组织采用新供应商影响



资料来源：IDC，国信证券经济研究所整理

图：12个月里组织将采用哪些主要方法对AI/ML模型进行API调用



资料来源：IDC，国信证券经济研究所整理

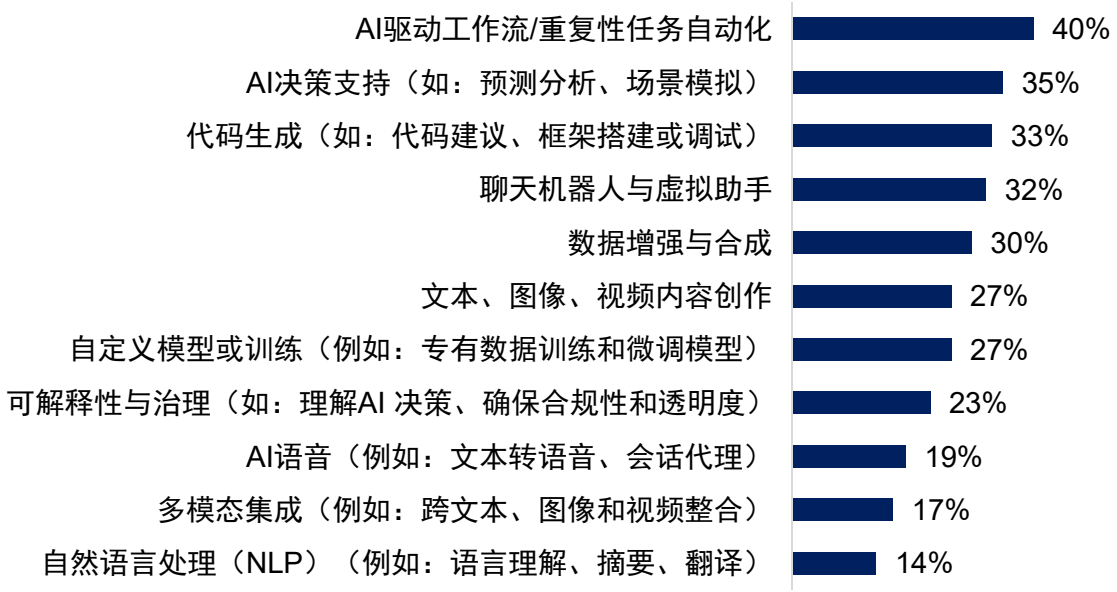
PaaS/Agent平台需求：提高开发交付效率，安全与隐私是最大障碍

国信证券 GUOSEN SECURITIES

PaaS平台的关键需求:围绕GenAI应用开发过程提高交付效率的功能/工具是主要需求，企业客户优先考虑有助于自动执行重复性任务并提供预测分析以协助数据驱动型决策的PaaS/应用程序平台，其次是GenAI相关功能（AIGC等）。AI驱动的工作流自动化（如自动测试代码、模拟重复性任务）是首要需求。决策支持（如预测分析、场景模拟）、代码生成、聊天机器人/虚拟助手和数据增强合成的重要性也较高，超过30%客户会因此考虑平台的选择。

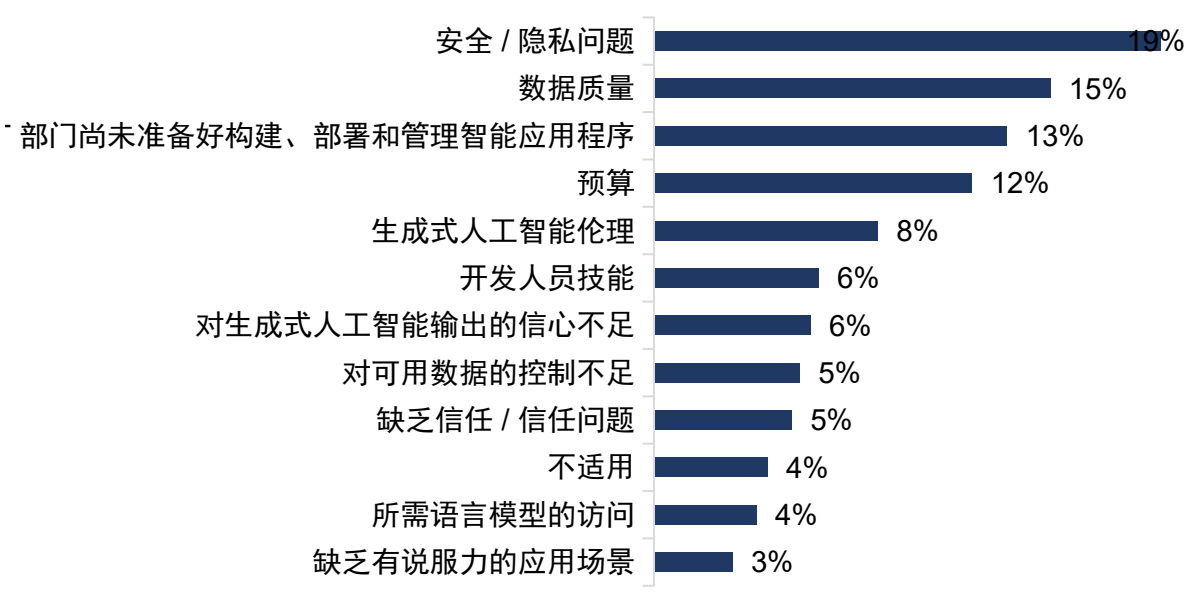
AI解决方案落地的主要阻碍：安全与隐私问题是影响企业AI项目落地最大的阻碍（19%），其次是数据质量（15%）和IT部门能力不足（13%）。另外预算、技术与伦理风险、输出结果可信度不足、数据使用控制不足也是重要挑战。

图：最能影响组织使用PaaS或应用程序平台的功能



资料来源：IDC，国信证券经济研究所整理

图：AI/ML或GenAI整合到组织数字解决方案中的主要障碍



资料来源：IDC，国信证券经济研究所整理

PaaS/Agent平台选择：AI模型为核心，借成熟生态（云、集成商）落地

国信证券 GUOSEN SECURITIES

•多数企业在开发和实施GenAI计划时，资源依赖模式表现得较为明显。尤其在模型调优/适配、评估/测试，方案集成以及运营/监控、基础算力与数据等环节。

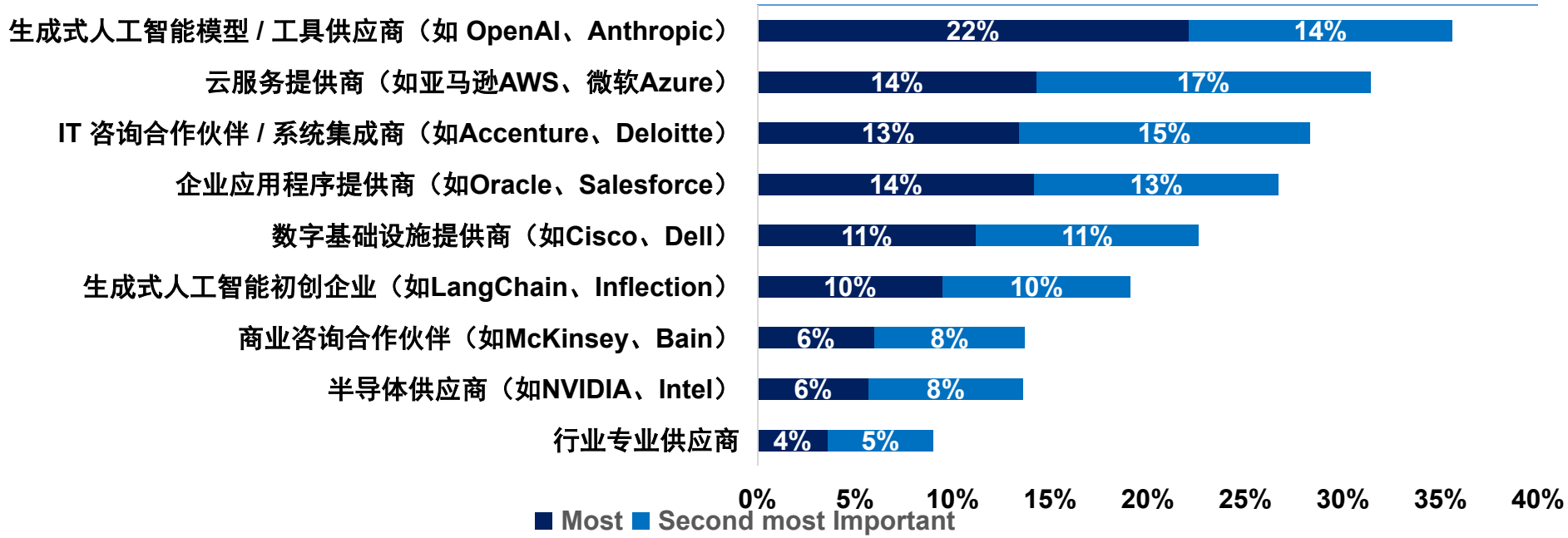
•企业AI战略“重心分层”，核心模型能力是企业AI战略的第一抓手、也是云厂必争环节。

① 第一层（核心）：生成式AI模型供应商+云服务商（算力） → 抢大脑；

② 第二层（支撑）：IT咨询（落地）+企业程序提供商（方案）+ → 保落地；

③ 第三层（补充）：初创企业（工具）、半导体（硬件） → 补细节。

图：最重要和第二重要的战略AI技术合作伙伴



资料来源：IDC，国信证券经济研究所整理

PaaS/Agent平台选择：云为底座、场景落地优先

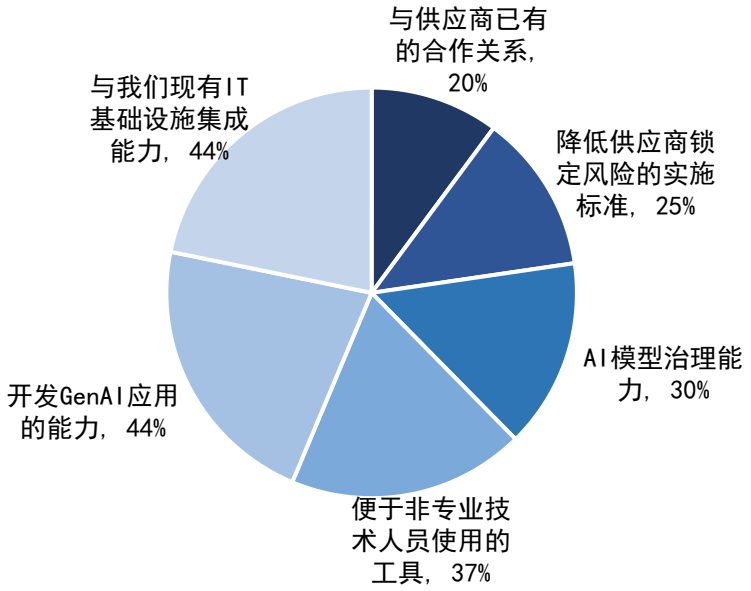
国信证券 GUOSEN SECURITIES

•企业落地AI方案时，最关注“技术整合能力（44%）”和“应用开发能力（44%）”。企业想要快速落地专属 GenAI 应用（如客服、营销），大模型应用是核心环节，同时为了避免AI孤岛，需打通现有系统（如ERP、CRM），让AI真正融入业务流程。

•能帮助企业实现AI目标的供应商中云服务商（49%）碾压式领先，反映出AI应用依赖于AI平台，而AI平台=云算力+开发平台+生态的建设逻辑。

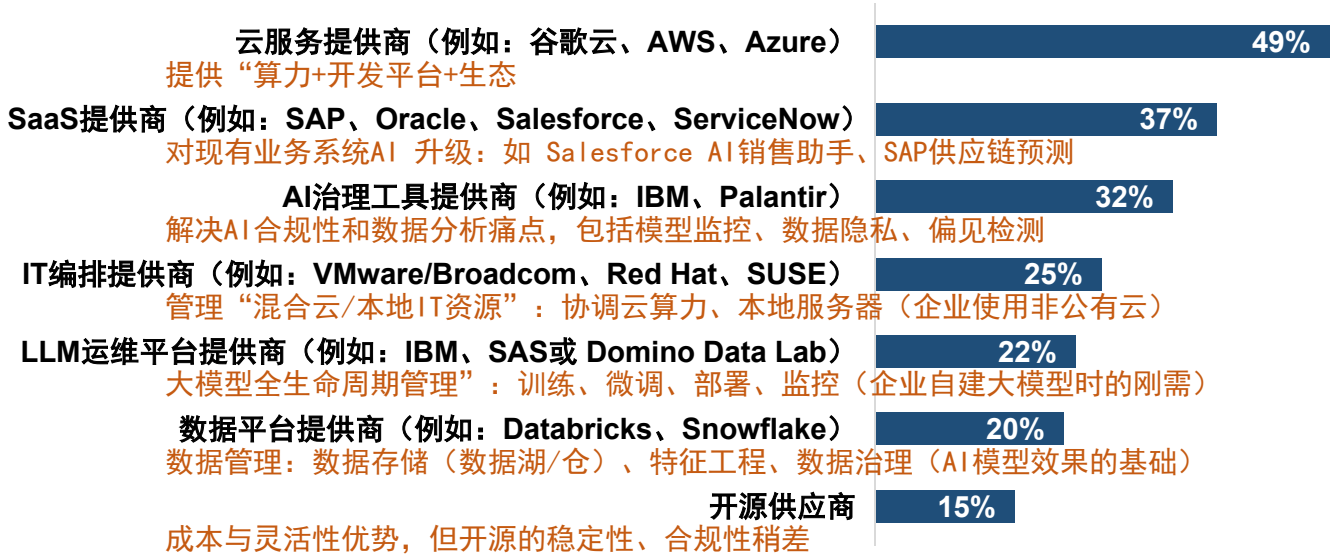
•其次为了使AI技术快速赋能现有业务，SaaS提供商和AI治理工具提供商/解决方案商的重要性也凸显。SaaS提供商可以帮助企业对现有业务系统进行AI化升级，AI治理工具提供商可以解决AI合规性和数据分析的痛点。

图：AI方案落地的关键要素



资料来源：IDC，国信证券经济研究所整理

图：最能助力机构实现 AI目标的供应商



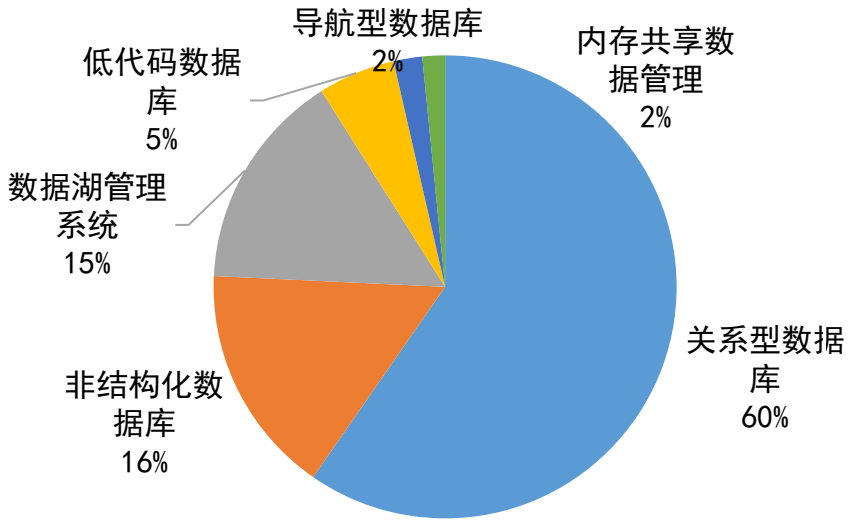
资料来源：IDC，国信证券经济研究所整理

AI Data Infra（数据基础设施）：新需求与数据库产品份额变化

国信证券 GUOSEN SECURITIES

- 根据IDC数据2024年数据库细分市场增速差异明显，增长主力为数据湖管理系统（+27.7%）、低代码DBMS（+26.8%）、非结构化DBMS（+21.6%），而传统的关系型DBMS增速仅6.7%（虽仍占60%份额）。这一差异反映三大需求趋势：
- **①灵活、多模态数据处理需求上升**：AI时代非结构化数据占比超80%，数据湖与非结构化数据库支持海量、多格式数据（如实时流数据、非定义化数据），适配倒逼技术架构向"湖仓一体(LakeHouse)"演进，要求支持多模态数据统一存储、高维矩阵计算、批流一体处理；
- **②低门槛开发需求增长**：低代码DBMS让非专业用户（如知识工作者）无需DBA即可管理数据，提升协作效率；
- **③传统架构需求稳定但放缓**：关系型数据库增速低于灵活型，企业更倾向于“核心系统保留关系型+新兴场景用灵活架构”的混合策略。
- AI对数据基础设施提出的全新要求，也包括**云部署迁移**：公共云服务占比从2023年的56.6%提升至2024年的62.2%；随着大模型在企业级市场应用场景的快速拓展，更需内置AI原生能力，实现"数据治理+模型训推"的无缝衔接。

图：2024年数据库类型与份额



资料来源：IDC，国信证券经济研究所整理

图：各类数据库特点、场景、主要玩家与2024年增速

数据库类型	特点与差异	2024 年增速)	典型使用场景	主要玩家与产品
关系型数据库	基于关系模型（表结构），需预定义schema（明确数据对象的结构（字段/列类型、有效值），用 SQL操作，数据结构化强，适合关联数据处理	**6.70%**	企业交易处理（如银行转账、电商订单）	Microsoft：SQL Server、Azure SQL Database；Oracle：Oracle Database
非结构化数据库	无需预定义数据结构，处理非结构化/半结构化数据（文本、图片等），NoSQL 特性，可无Schema，高扩展性	**21.60%**	AI 语义搜索、社交平台非结构化数据存储	MongoDB；Microsoft：Azure Cosmos DB（文档 API）；Oracle: NoSQL DB；Amazon:DynamoDB（文档存储）；
数据湖管理系统	存储海量多源异构数据（结构化 + 非结构化），保留原始格式，可无Schema，按需处理，成本低	**27.70%**	企业大数据分析（如用户行为挖掘）	Databricks：Lakehouse；Microsoft：Azure Data Lake；Amazon：S3 + Glue；Google：BigQuery+ BigLake； Snowflake;
低代码数据库	可视化界面开发，schema 简单，非数据库管理用户可定义管理数据，集成脚本语言、报表工具	**26.80%**	部门级数据管理（如营销团队客户统计）	Microsoft：Power Apps + Dataverse；Oracle：APEX；
导航型数据库	需 schema，基于层次/网状模型，需预定义导航路径，查询依赖路径，灵活性低	**2.90%**	传统金融核心系统（如老旧账户管理）	Oracle：IMS（历史产品）；
内存共享数据管理	管理内存中瞬态数据，支持多进程共享，高并发.低延迟，部分支持持久化	**15.50%**	实时应用（如高频交易、实时物流调度）	Microsoft：Azure Cache for Redis；Oracle：TimesTen；Amazon：ElastiCache；

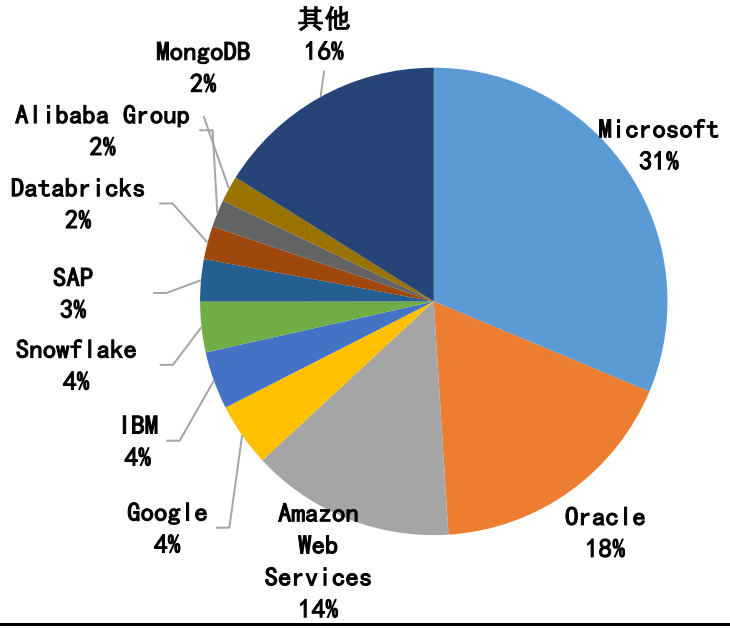
资料来源：IDC，华尔街新闻，各公司官网，国信证券经济研究所整理

AI Data Infra（数据基础设施）：格局壁垒演绎与主要玩家

国信证券 GUOSEN SECURITIES

- **Microsoft**：以30%+的份额领跑（同比+13.4%），凭借SQL Server、Azure SQL、CosmosDB的AI集成提升性能与开发者效率。
- **AWS**：份额14%（同比19.6%），推出无服务器分布式数据库AuroraDSQL，升级Redshift（AI增强）与DynamoDB（一致性/成本优化）；
- **Google**：份额4%（增速42.7%），新产品AlloyDB集成ScaNN向量搜索、Spanner新增图数据库功能（SpannerGraph）。
- **Snowflake**:份额4%（增速27%），升级CortexAI提升AI驱动的数据分析与模型监控能力；Apache Iceberg表正式发布，推出开源跨引擎目录PolarisCatalog，支持多格式数据管理；
- **Databricks**：份额2%(增速45%、头部厂商中最高），开源UnityCatalog、升级DeltaLake4.0，强化湖仓架构AI能力，。
- **MongoDB**:份额2%(增速22%)，作为非结构化数据库（NoSQL）代表，适配企业对动态、多模态数据的存储与检索需求。

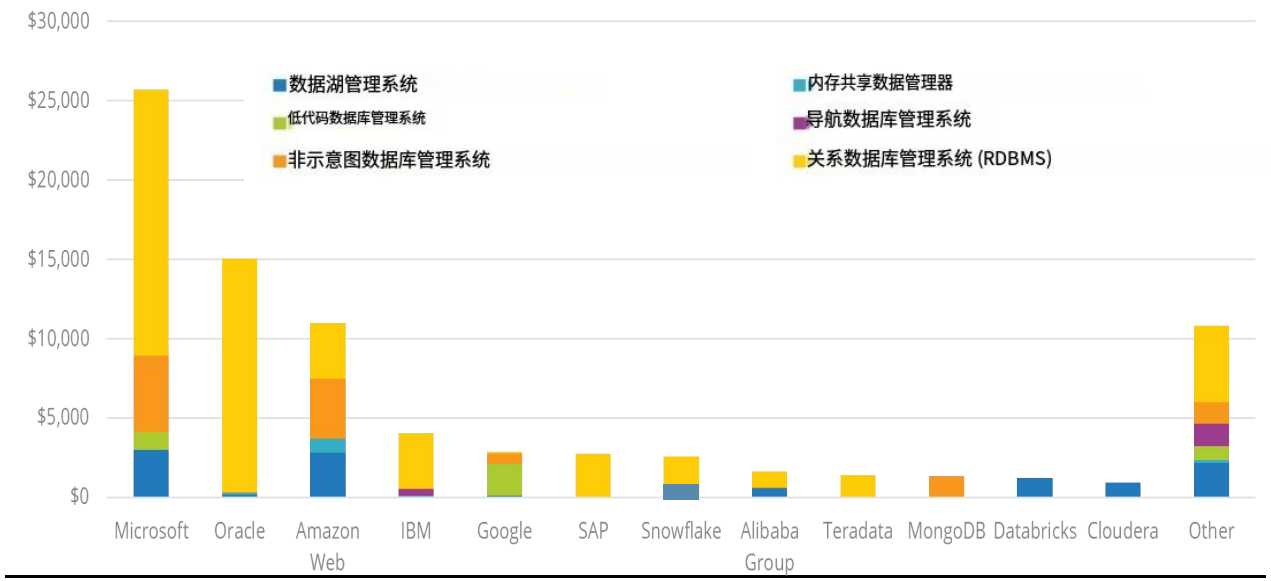
图：2024年数据库厂商份额与增速



公司	2024年收入增速
Microsoft	13.4%
Oracle	4%
Amazon	19.6%
Google	42.7%
IBM	-7.2%
Snowflake	27%
SAP	-0.7%
Databricks	45%
MongoDB	21.7%

资料来源：IDC，国信证券经济研究所整理

图：各个数据库公司细分数据库产品收入组成（百万美元）



资料来源：公司官网、国信证券经济研究所整理

AI Data Infra公司：通过快速收购与推新品完善产品布局

国信证券 GUOSEN SECURITIES

- 2024年至今头部数据库厂商通过收购强化AI原生、云安全能力，提升用户体验与生态，例如：

➢ **Microsoft**：先后收购Fungible（基础设施）+Inflection AI（增强Copilot与自然语言查询），强化云数据库的性能与智能化；

➢ **Snowflake**：收购TruEra补全AI模型监控短板。最新发布Openflow产品基于收购的Datavolo技术构建，支持结构化/非结构化、批处理/流数据接入，瞄准170亿美元数据集成市场。

➢ **SAP**：收购WalkMe优化数据库平台的用户体验，应对云迁移中客户的上手门槛；

➢ **IBM**：收购 HashiCorp，提升多云自动化、简化数据库安全；

图：近两年各数据库厂商收购、投资相关公司情况

资料来源：各公司官网、业绩会，国信证券经济研究所整理

案例：Snowflake 2026财年上半年推出约250项新功能

- Snowflake 26FYQ2业绩表示AI相关产品adoption快速提升（近50%新客户因AI选择Snowflake，尽管收入贡献仍较小），重磅产品如下：

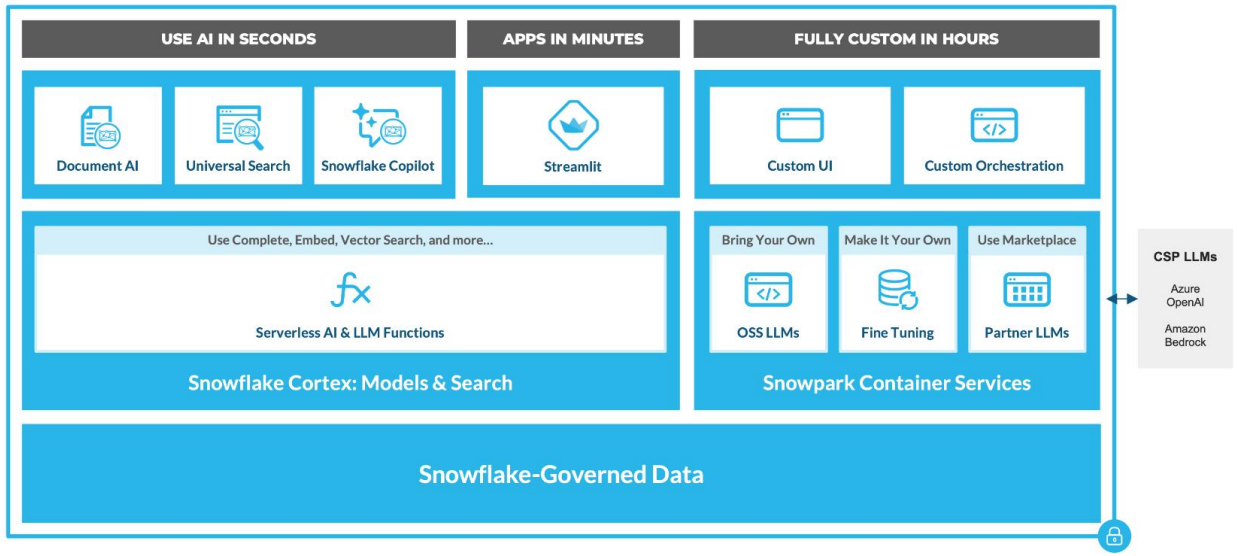
① Iceberg Tables（2024年6月10日宣布正式商用）： Snowflake Iceberg表提供了直接与数据湖中的Iceberg和Parquet数据交互的功能，即可助力模型部署、AI应用开发以通过 Snowflake 无缝安全地共享、协作、处理第三方存储的Iceberg数据。

② Snowflake Cortex（2024年5月GA）：大语言模型推理的托管服务，类似Google的Vertex AI，涵盖从模型数据的抓取整理，到模型训练、验证、部署、扩展、协作、治理、监控、维护等整个 AI/ML 工作流程与模型生命周期。还包括以下功能：

- Snowflake Copilot：是一个由 LLM 提供支持的助手，用于使用自然语言生成和优化 SQL。
- Document AI：适用提取数据，客户可以处理任何文档（pdf、word、txt、屏幕截图）并获得问题的答案。
- Universal Search ：由 LLM 提供支持的搜索，用于数据云/数据库中搜索和发现数据和应用。

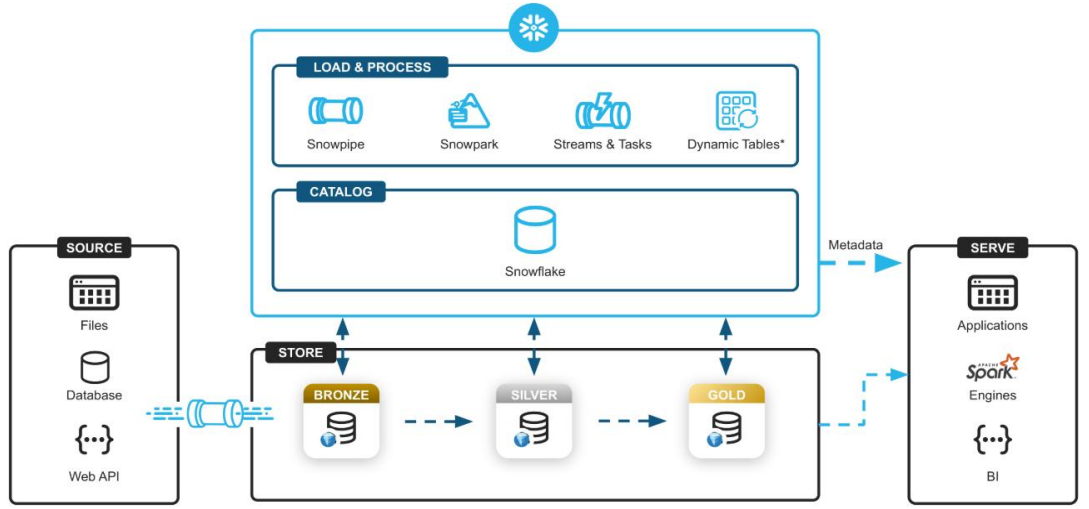
③ Snowflake Intelligence（2025年6月预览）：Agentic AI功能、企业级AI代理平台，整合结构化/非结构化数据，支持用户通过自然语言与企业数据交互（如自动生成采购订单、更新 CRM），已进入公开预览阶段；

图：Snowflake AI产品



资料来源：公司官网，国信证券经济研究所整理

图：Snowflake Iceberg表交互架构

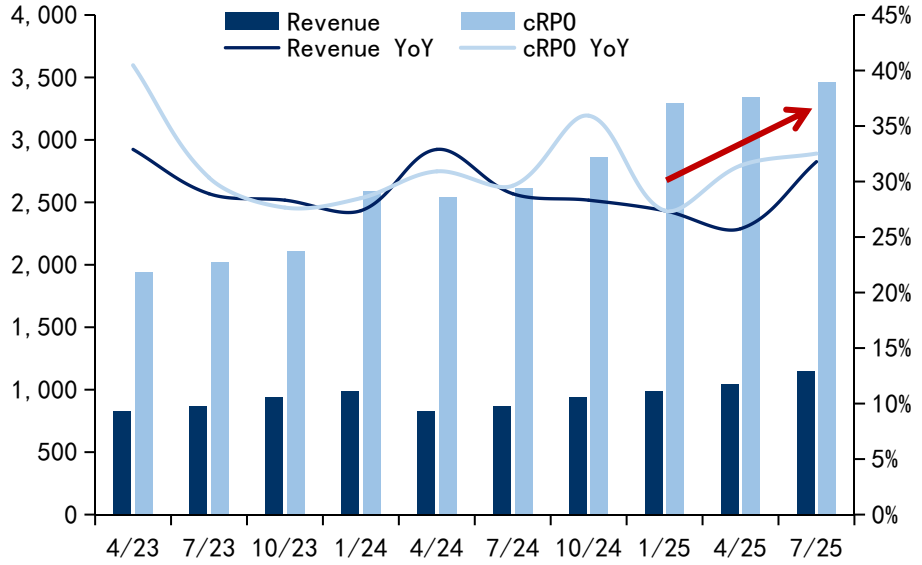


资料来源：公司官网，国信证券经济研究所整理

AI Data Infra：相关SaaS公司二财季业绩数据表现亮眼

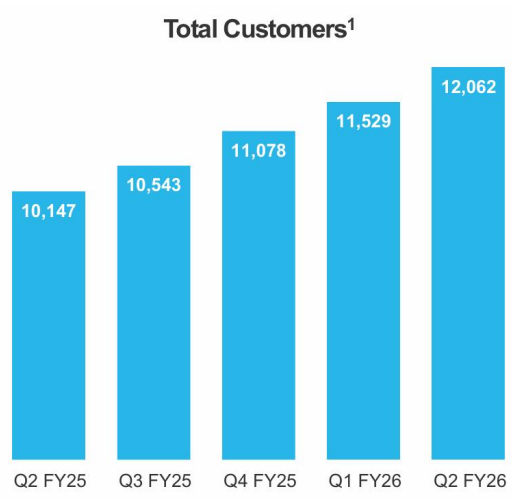
- AI Data Infra公司连续两个季度收入加速明显，二季度业绩超公司指引，纷纷上调全年营收与利润指引。
- **Snowflake 26FYQ2业绩**（25年5月1日-25年7月31日）：产品营收为1.09亿美元（同比+32%），显著超公司25%的同比增长指引，核心业务强劲。Non-GAAP运营利润率为11%，超此前指引8%。上调26财年收入指引至43.95亿美元，同比增长27%（此前指引为同比+25%）。全年上调则基于Q2超预期表现、新功能adoption加速及大客户迁移需求释放；
- **Mongodb 26FYQ2业绩**（25年5月1日-25年7月31日）：营收5.91亿美元（同比+24%），高于指引上限。Non-GAAP运营利润8700万美元（利润率15%，同比提升4个百分点）。上调26财年预计营收23.4-23.6亿美元，同比+16.4-17.4%（原指引22.5-22.9亿美元）。上调26财年Non-GAAP运营利润预计3.21-3.31亿美元（原指引2.67-2.87亿美元），Non-GAAP OPM上限14%，较原指引提升150个基点。

图：Snowflake 收入与cRPO变化(百万美元，%)



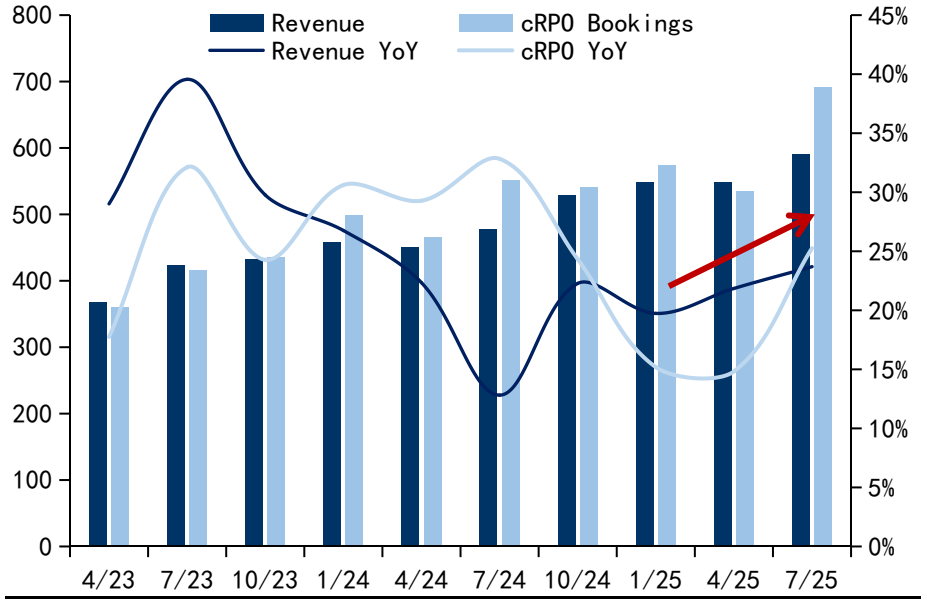
资料来源：公司财报、国信证券经济研究所整理

图：Snowflake用户数变化



资料来源：公司业绩会、国信证券经济研究所整理

图：Mongodb 收入与cRPO变化(百万美元，%)



资料来源：公司财报、国信证券经济研究所整理

Agent平台重要供应商：伴随Agent落地需求股价强劲

图：海外云厂Agent平台对比

	公司名称	公司业务	股价YTD	股价关税底后反弹
云服务商	微软（Microsoft）	Azure 云承载 AI 算力，集成 Copilot+OpenAI 能力，提供企业级 AI 开发与部署平台	19%	42%
	谷歌（Alphabet）	Google Cloud 提供 AI服务，Gemini 大模型 + TPU芯片，赋能搜索、广告等业务AI 化	9%	43%
	亚马逊（Amazon）	AWS 云提供 AI/ML 服务（如 Bedrock），合作Anthopic、自研Nova大模型，支撑电商等 AI应用	4%	33%
	Oracle	提供云技术应用与云基础设施平台Oracle Cloud ，提供数据库、ERP、CRM 等企业级应用	42%	89%
SaaS方案商	ServiceNow	企业服务管理 SaaS，客户服务管理、ITSM、HRSD 流程等产品	-18%	20%
	Salesforce	专注客户关系管理，通过云计算提供销售、营销、客服等CRM 解决方案，拓展应用和平台服务	-26%	1%
	SAP	全球领先的企业应用软件，专注 ERP，提供财务管理、HR 云等多领域软件及行业云解决方案	-2%	10%
AI治理工具	IBM	业务涵盖 AI 平台、云计算服务、服务器等硬件研发制造，以及企业咨询、系统集成等信息技术服务	13%	5%
	Palantir	专注于大数据分析和决策支持平台，为政府和企业提供定制化数据管理与分析解决方案	113%	106%
云平台安全类SaaS	CrowdStrike	网络安全公司，通过软件与服务为企业检测、阻止黑客威胁，利用 AI 提供端点、云等安全防护	24%	30%
	Palo Alto Networks	网络安全 SaaS，提供先进防火墙、云安全、端点防护等方案	2%	22%
云平台数据类SaaS	Snowflake	数据云平台，帮助企业处理海量数据，支持数据工程、仓库、科学等多场景	40%	60%
	MongoDB	专注数据库领域，提供非关系型数据库 MongoDB，助力开发团队构建智能应用	30%	86%
	Datadog	数据观测性SaaS，诊断系统故障，监测服务器、API 等运行状况，助力企业保障 IT 系统稳定高效运行	-12%	43%
AI芯片与硬件	英伟达（Nvidia）	主导 AI 算力芯片（GPU/H100 等），为大模型训练、推理提供核心硬件支撑	34%	87%
	AMD	设计制造微处理器等，业务涵盖数据中心、客户端、游戏、嵌入式产品领域	37%	112%
	Broadcom	提供半导体和基础设施软件产品，涵盖企业软件、宽带网络、数据中心存储等多个领域相关业务。	28%	91%

资料来源：Wind，国信证券经济研究所整理　　按照2025年8月28日收盘数据统计

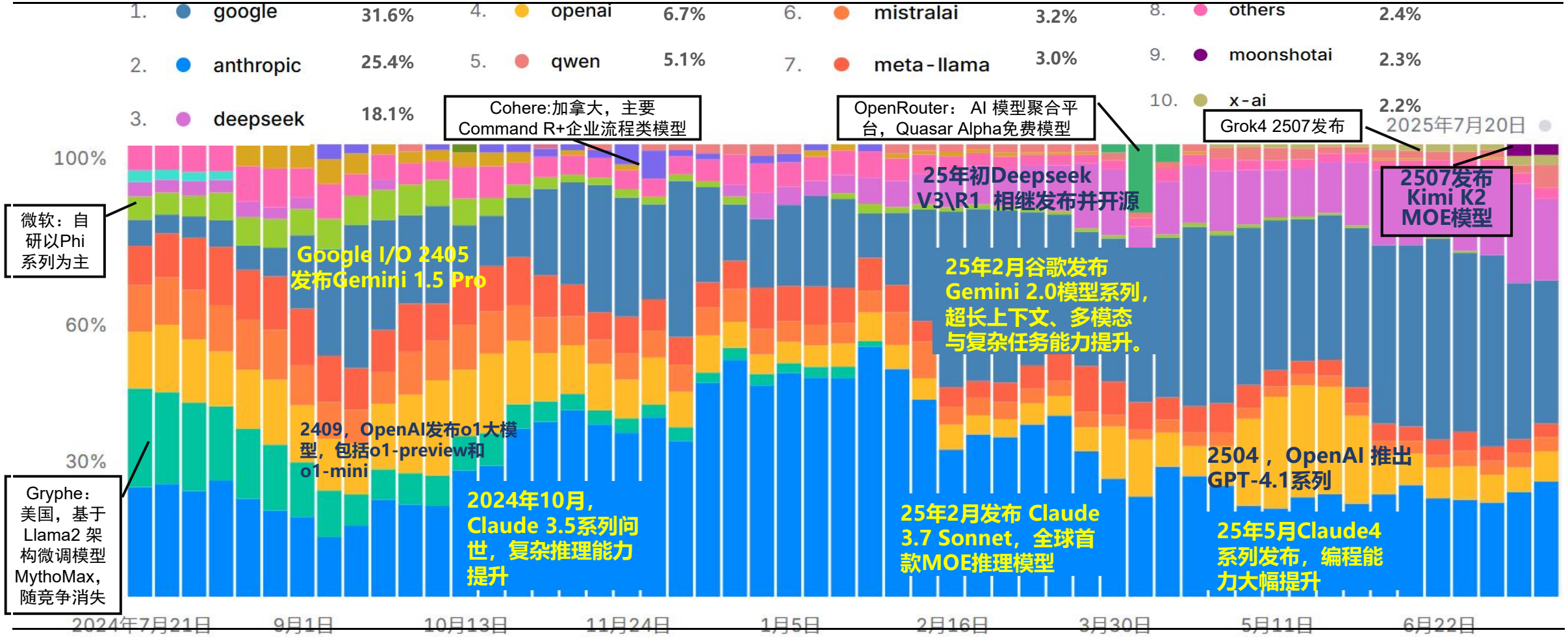
目录

大模型份额变化：谷歌与Anthropic优势明显，国内模型稳健进步

国信证券 GUOSEN SECURITIES

- **根据Openrouter数据谷歌与Anthropic模型份额较高，目前占据模型调用市场半壁以上江山。**两者陆续迭代发布重磅模型版本，如谷歌的Gemini 1.5与Gemini2.0系列，在多模态、长文本领域优势明显。Anthropic发布Claude 3.5\3.7\4.0系列，复杂推理编程能力突出。
- **国内随着Deepseek推理开源V3/R1模型的发布，以开源为特点在模型市场份额稳步提升。**代表如Deepseek系列、QWEN系列、Kimi系列。

图：近一年全球大模型服务商份额变化(根据模型调用量排序)



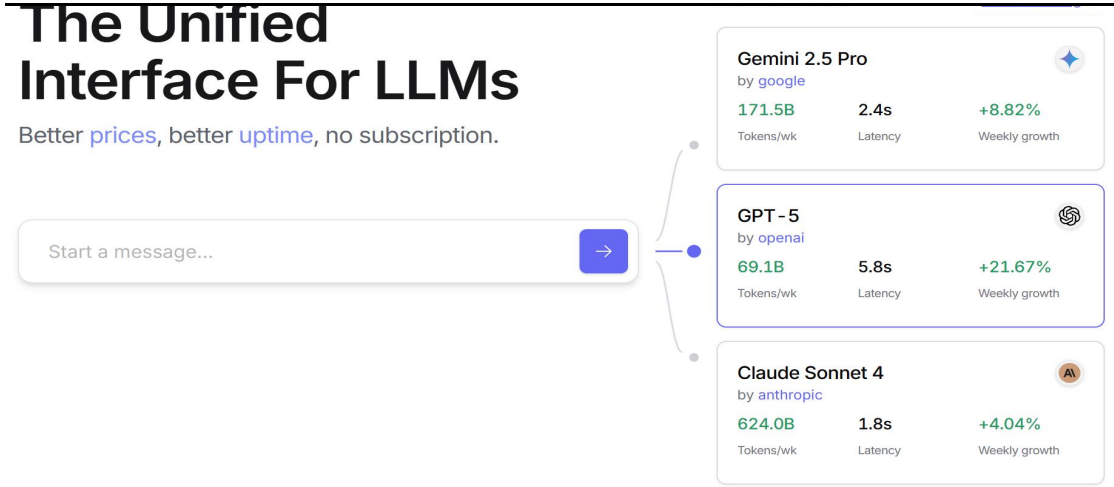
资料来源：Openrouter，国信证券经济研究所整理　注:Openrouter市场份额数据是基于其平台上的第三方API调用量统计

Openrouter API聚合平台与数据口径说明

国信证券 GUOSEN SECURITIES

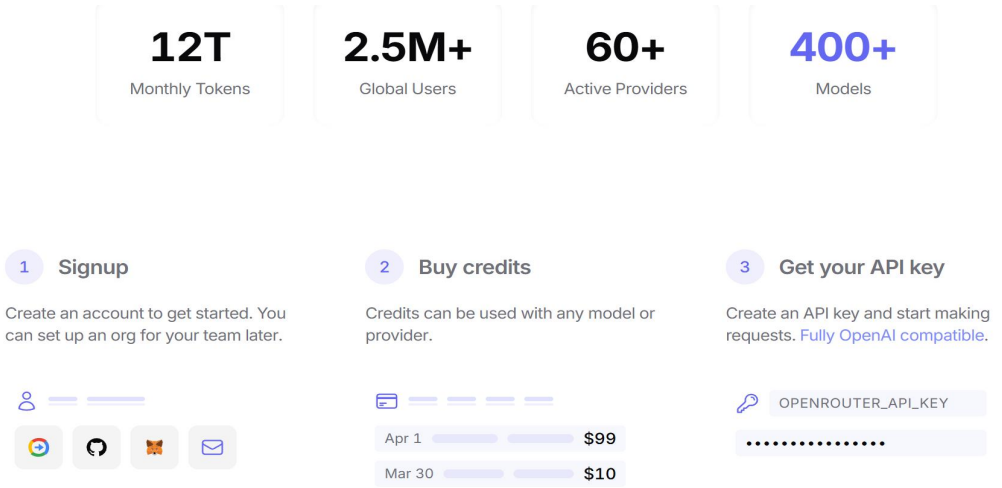
- **OpenRouter是面向大模型的统一API网关/聚合平台**，提供标准化路由、高可用与成本优化，兼容OpenAI生态，支持400+模型与活跃供应商接入。通过兼容接口与SDK，单一APIKey可调用多厂商模型，降低集成成本。收费方法为：充值通用额度，按上游费率消耗，平台收取约5%－5.5%手续费，类似平台包括Hugging Face、AI21 Labs等。
- **OpenRouter模型调用量口径说明：**仅反映“通过OpenRouter平台产生的调用”，不包含用户直接向模型厂商API的调用、通过其他云平台产生的调用量以及企业自建私有化部署大模型量，并非全市场总用量。
- **OpenRouter模型调用量占比全市场总调用量我们测算约为1%。**2025年5月OpenRouter的年化收入约500万美元，对应年化推理支出（GMV）约1亿美元。Polaris MarketResearch预测，全球大语言模型市场规模2025年约为78亿美元，占比约1.2%。考虑到OpenRouter以海外需求为主、而中国区模型调用价格显著低于全球，OpenRouter整体调用量份额占比约为1%。
- **OpenRouter平台用户特点：**主要客户以“开发者/小团队”长尾用户为主，共性是追求多模型统一接入、高可用、成本优化与快速迭代；开发者与小微团队占比高、用量分散，包括有些跨境客户依赖代理中转实现海外模型合规接入，以及学术/研究机构进行统一评测与实验，中大企业与机构占比低。
- **OpenRouter平台行业特点：**模型调用量以编程类应用为主，占比约87%；其他场景（如营销、翻译、角色扮演、金融等）总占比约13%；

图：Openrouter大模型调用说明



资料来源：公司官网，国信证券经济研究所整理

图：Openrouter 接入步骤



资料来源：公司官网，国信证券经济研究所整理

谷歌模型性价比显著，引领市场份额

国信证券 GUOSEN SECURITIES

表：全球调用量领先的大模型参数

模型商	模型名称	发布时间	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
Google	Gemini 2.5 Flash	2025/6/17	**1.05M**	64K	0.30	2.50	**0.51**	Google
Google	Gemini 2.0 Flash	2025/2/5	**1.05M**	8K	0.10	0.40	**0.58**	
Google	Gemini 2.5 Pro	2025/6/17	**1.05M**	66K	1.25	10.00	2.30	
Anthropic	Claude Sonnet 4	2025/5/22	200K	64K	3.00	15.00	2.14	Google、Amazon、Anthropic
Anthropic	Claude 3.7 Sonnet	2025/2/24	200K	64K	3.00	15.00	1.33	
Anthropic	Claude Opus 4	2025/5/22	200K	32K	15.00	75.00	3.09	
OpenAI	GPT-4.1	2025/4/14	**1.05M**	32K	2.00	8.00	0.46	OpenAI
OpenAI	GPT-4.1 Mini	2025/4/14	**1.05M**	32K	0.40	1.60	0.42	OpenAI
OpenAI	GPT-4o	2024/5/13	128K	16K	2.50	10.00	0.49	OpenAI、Azure
Mistral	Mistral Nemo	2024/7/19	32K	16K	0.01	0.00	**0.32**	开源，多数主流云平台（如 AWS、Azure、GCP）
Meta	Llama 4 Maverick	2025/4/5	**1.05M**	16K	0.15	0.60	0.48	开源，多数主流云平台（如 AWS、Azure、GCP）
Meta	Llama 3.3 70B Instruct	2024/12/6	131K	16K	0.04	0.12	**0.22**	
xAI	Grok 4	2025/7/9	256K	**256K**	3.00	15.00	13.03	xAI
DeepSeek	DeepSeek V3 0324 (free)	2025/3/24	33K	164K	**免费**	**免费**	1.14	开源，多数主流云平台（如 AWS、Azure、GCP）
DeepSeek	R1 0528 (free)	2025/5/28	164K	164K	**免费**	**免费**	2.83	
月之暗面	Kimi K2(free)	2025/7/11	33K	32K	**免费**	**免费**	2.71	开源，阿里云、Hugging Face等可下载
阿里	Qwen3 235B A22B 2507(free)	2025/7/21	262K	**262K**	**免费**	**免费**	1.88	开源，阿里云、Hugging Face等可下载
腾讯	Hunyuan A13B Instruct (free)	2025/7/8	33K	32K	**免费**	**免费**	2.04	开源，腾讯云、Hugging Face等可下载

资料来源：各公司官网，Openrouter，国信证券经济研究所整理　注:Openrouter市场份额数据是基于其平台上的第三方API调用量统计

海外大模型差异化发展

- **OpenAI**：**技术路径上依然领先，聚焦强化推理与专业领域能力**，逐步提升性价比，端到端多模态生成能力较谷歌偏弱。市场此前预期GPT5对各模态输入有极强的原生理解并且可以把它转化成可执行的指令/多模态输出，实际并未实现。
- **谷歌：端到端原生多模态领先**（视频、音频、文字）输入理解，输出仅文字。AIGC尤其视频生成方向业内领先。
- **Anthropic：强调实用性、编程场景领先，**高准确度带来了高市场占有率，通用性能与多模态能力相较头部模型稍弱。

表：海外领先的大模型对比

排序	模型商	模型系列	发布时间	模型类型	特点
一梯队	OpenAI	GPT-4.1	2025/4/14	通用推理大模型	支持100 万 token 上下文，编码能力显著提升（SWE-bench 得分 54.6%），多语言处理效率翻倍。
		GPT-5	2025/8/7	通用推理大模型	集成 o系列推理能力，低幻觉率、强大的编码分析和拟人化能力，支持标准版、mini 版和nano版 API，调用成本进一步下降。
	Google	Gemini 2.0 Flash/Pro	2025/2/5	推理MOE模型	集成 Agent 架构，支持实时接收文字、语音、图像、视频信息并进行推理反馈。
		Gemini 2.5 Flash/Pro	2025/6/17	端到端的原生多模态推理模型	支持10M token 超长上下文，推理效率提升，具备代码生成和多模态交互能力。
		Veo 3	2025/5/21	视频生成模型	基于文本和图像提示生成高质量视频，配备 V2A 技术，实现音画完美同步，支持最高 4K 分辨率输出，具备物理效果模拟能力，可执行复杂镜头指令。
二梯队	Anthropic	Claude 3.7 Sonnet	2025/2/24	推理/编程模型	市场上首个混合推理模型，可在普通回答和深度思考模式间切换，编码和前端开发能力突出，API 用户可控制思考预算。
		Claude Sonnet/Opus 4	2025/5/22	推理/编程模型	Opus 4 支持连续 7 小时复杂推理任务，Sonnet 4 在 SWE-bench 编程测试中得分 72.7%。
三梯队	Meta	Llama 4 Scout/Maverick	2025/4/5	多模态 MoE 模型	Scout 支持 1000 万 token 上下文，Maverick 以 402B 总参数超越 GPT-4o，支持图像理解和代码生成。
	xAI	Grok 3 Beta	2025/2/18	通用大模型	支持 131K token 上下文，数学推理（AIME 2025 得分 93.3%）和代码生成能力领先，集成 DeepSearch 实时数据抓取功能。
		Grok 4	2025/7/9	推理模型	分为标准版本和多代理版本，上下文窗口最高支持 256K tokens，号称 “世界上最强 AI 模型”，处理学术问题达博士级别。

资料来源：各公司官网，AGI-Eval 榜单、FlagEval 榜单、SEAL 榜单，国信证券经济研究所整理

国内大模型并未拉开显著差距，领先模型以自研为主

国信证券 GUOSEN SECURITIES

- **DeepSeek：技术研究领先**，采用混合注意力机制、动态路由MoE等架构创新，降低计算复杂度和通信开销、提升数据利用率，在代码生成、数学计算等专业领域表现出色，主打开源、支持超100种语言。
- **阿里：自研能力与综合能力强**，模型参数与种类丰富，Qwen3.0在多模态和对话交互方面表现突出，支持超长文本处理。
- **其他：**字节豆包大模型各模态表现较均衡，虽多数模态不领先，但总分较高。百度文心大模型在中文场景深度优化，长文本理解能力突出。腾讯混元大语言模型基于DeepSeek改造，自研投入较少。

表：国内领先的大模型对比

排序	模型商	模型系列	发布时间	模型类型	特点
一梯队（自研能力强）	DeepSeek	DeepSeek V3 0324	2025/3/24	开源推理模型	推理任务表现提高，在数学、代码类相关评测集上得分超过 GPT-4.5，中文写作、搜索等能力优化
		R1 0528	2025/5/28	推理模型	基于 V3 架构，采用多头潜在注意力（MLA）和混合专家（MoE）技术，推理效率比传统架构提升 30%。
	阿里	Qwen3	2025/7/21	开源推理模型	包含 2 个 MOE 模型和 6 个 Dense 模型，0.6B-235B 参数版本，支持轻量级设备部署和复杂推理任务，开源基座模型适配多场景。
二梯队	月之暗面	Kimi K2	2025/7/11	MoE基础模型	1万亿参数，支持128Ktoken上下文，代码生成（SWE-bench得分72.5）和Agent任务能力突出，API定价低至4元/百万输入tokens。
	字节	豆包大模型1.6	2025/6/11	多模态通用模型	推理、数学、指令遵循、Agent 等能力有较大提升，首创按输入长度区间定价（0-32K 区间输入成本0.8元 / 百万tokens），支持视频生成和企业级 Agent 开发。
	百度	文心大模型4.5	2025/3/16	原生多模态推理模型	采用 FlashMask 动态注意力掩码和多模态异构专家技术，长文处理和多轮交互能力显著提升。文心大模型 4.5 Turbo 具备多模态、强推理、低成本三大特性。
三梯队	腾讯	混元 Turbo S	2025/5/22	语言模型	在 Chatbot Arena 排名全球前八，国内仅次于 DeepSeek，代码、数学等理科能力进入全球前十。
		混元 T1	2025/6/25	MOE深度推理模型	基于 TurboS 快思考基座，扩展推理能力，全面搭建模型文理科能力，长文本信息捕捉能力强。

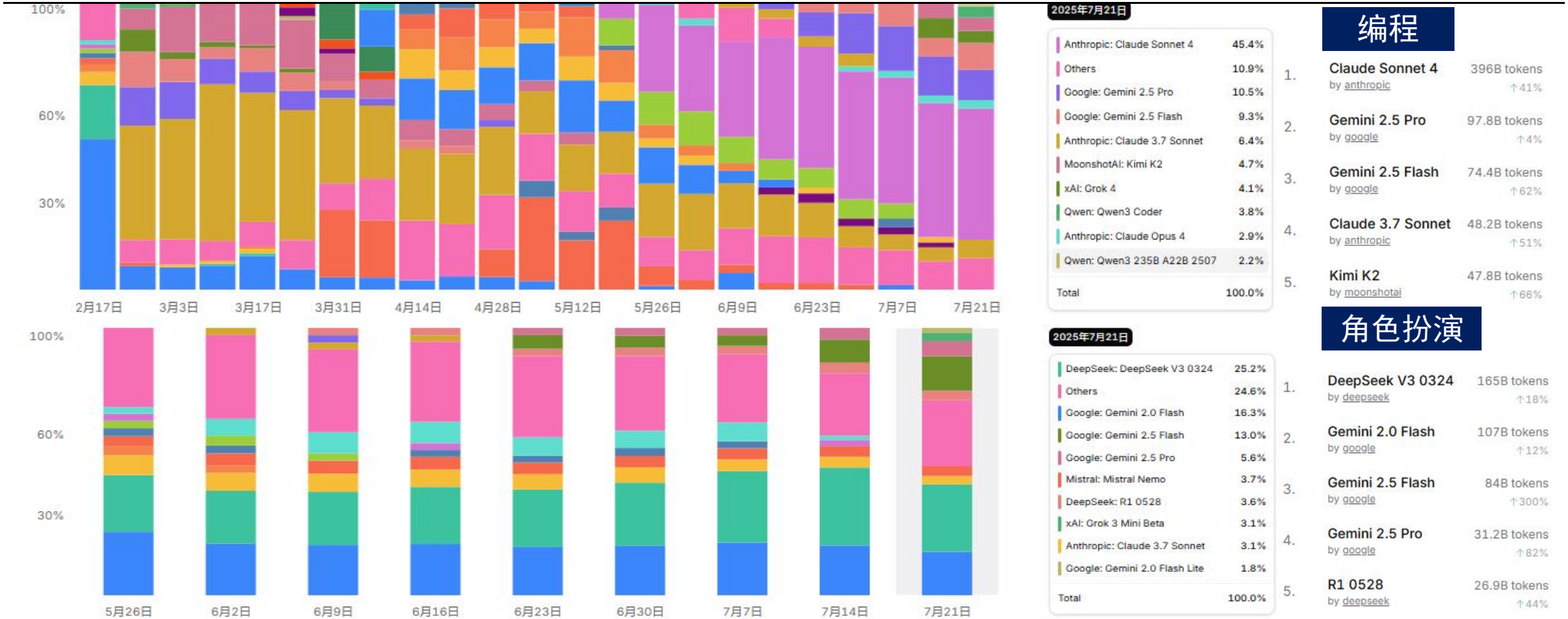
资料来源：各公司官网，AGI-Eval 榜单、FlagEval 榜单、SEAL 榜单，国信证券经济研究所整理

模型分场景对比：编程场景Claude Sonnet4占据近半调用量

国信证券 GUOSEN SECURITIES

- **编程场景Claude Sonnet4占据近半调用量**:超高编程准确率，同时在代码规范性、安全性、长文本逻辑连贯性表现突出。Gemini 2.5 Pro与Flash则主打“百万Tokens超长上下文+ 高性能”在大型代码库分析、多文件联动开发中具备优势。
- **DeepSeekV3擅长角色扮演场景：**训练强化了对“角色核心特征”的捕捉，严格遵循用户设定，开源允许用户基于自身需求微调角色模板。

图：大模型在编程与角色扮演场景份额变化



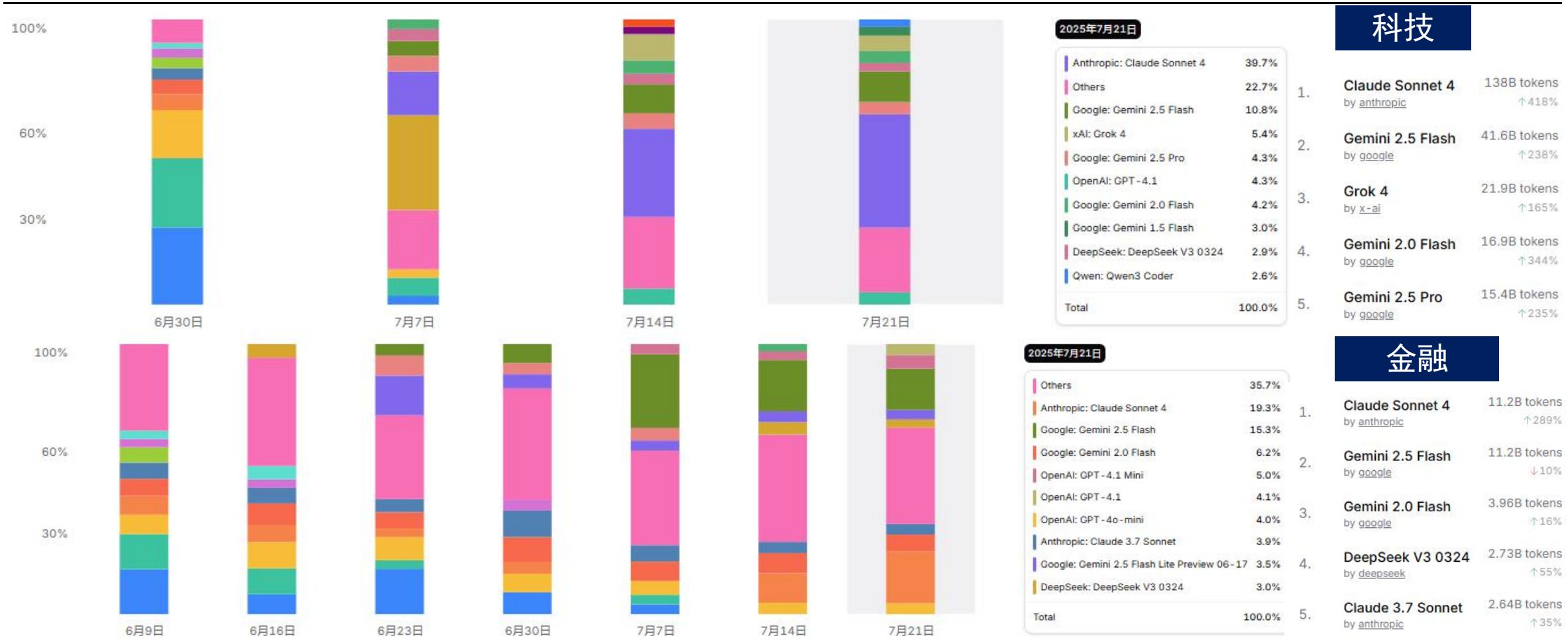
资料来源：Openrouter，国信证券经济研究所整理　注:Openrouter市场份额数据是基于其平台上的第三方API调用量统计

模型分场景对比：严谨科学领域Claude Sonnet4与Gemini Flash占优

国信证券 GUOSEN SECURITIES

- Claude Sonnet4综合性能与准确率较高、尤其在复杂严谨使用场景，安全合规与隐私保护方面满足行业严格法规要求。
- Gemini 2.5/2.0 Flash具备“百万Tokens超长上下文+低时延”优势可快速响应与高效处理，同时多模态交互能力强。

图：大模型在科技与金融场景份额变化



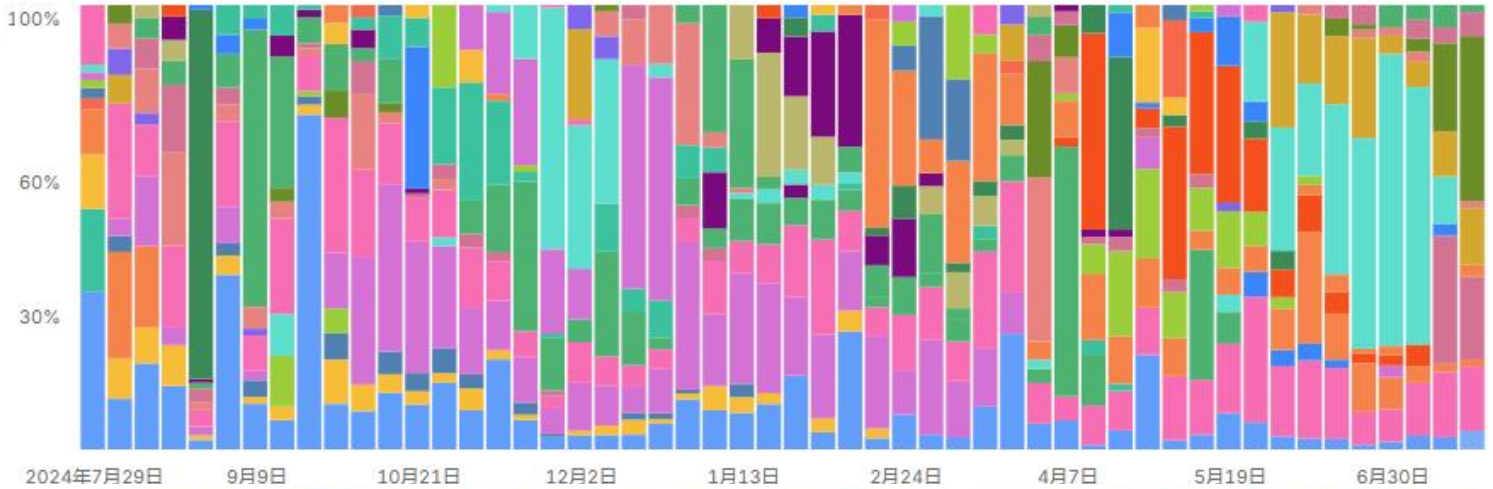
资料来源：Openrouter，国信证券经济研究所整理 注:Openrouter市场份额数据是基于其平台上的第三方API调用量统计

模型分场景对比：Gemini Flash在长上下文与低时延场景优势明显

国信证券 GUOSEN SECURITIES

- 营销搜索与翻译场景对上下文长度、响应速度以及多模态的交互要求较高，Gemini Flash支持百万tokens超长上下文处理、毫秒级响应与原生多模态生成，且作为轻量级模型、资源占用少。

图：大模型在营销搜索与翻译场景份额变化

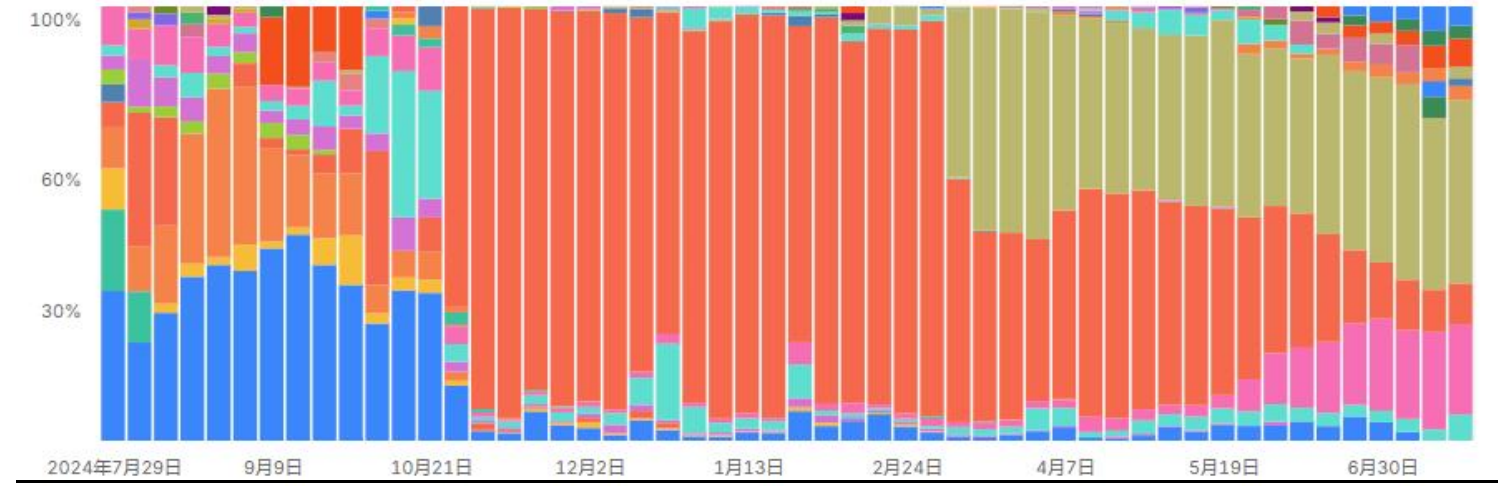


2025年7月21日

Google: Gemini 2.5 Flash	37.1%
Google: Gemini 2.0 Flash Lite	18.6%
Others	14.5%
Qwen2.5 7B Instruct	12.8%
Google: Gemini 2.5 Flash Lite Preview 06-17	5.4%
OpenAI: GPT-4o-mini	4.2%
OpenAI: GPT-4.1	2.6%
Anthropic: Claude Sonnet 4	1.6%
Google: Gemini 2.5 Pro	1.6%
Google: Gemini 2.0 Flash	1.5%
Total	100.0%

营销搜索

1.	Gemini 2.5 Flash by google	6.18B tokens	↑858%
2.	Gemini 2.0 Flash Lite by google	4.38B tokens	↑147%
3.	Qwen2.5 7B Instruct by qwen	1.99B tokens	↑96%
4.	Gemini 2.5 Flash Lite Preview by google	759M tokens	↑16%
5.	GPT-4o-mini by openai	481M tokens	↑31%



2025年7月21日

Google: Gemini 2.0 Flash	42.2%
Others	20.7%
Google: Gemini 1.5 Flash 8B	9.4%
Google: Gemini 2.5 Flash	6.5%
Google: Gemini 1.5 Flash	6.0%
Google: Gemini 2.5 Flash Lite Preview 06-17	4.5%
Anthropic: Claude 3.7 Sonnet	3.2%
Google: Gemini 2.5 Pro	3.0%
Anthropic: Claude Sonnet 4	2.7%
DeepSeek: DeepSeek V3 0324	1.7%
Total	100.0%

翻译

1.	Gemini 2.0 Flash by google	21.6B tokens	↑8%
2.	Gemini 1.5 Flash 8B by google	4.98B tokens	↑4%
3.	Gemini 2.5 Flash Lite Preview by google	3.44B tokens	↑92%
4.	Gemini 2.5 Flash by google	3.21B tokens	↑50%
5.	Gemini 1.5 Flash by google	3.17B tokens	↑140%

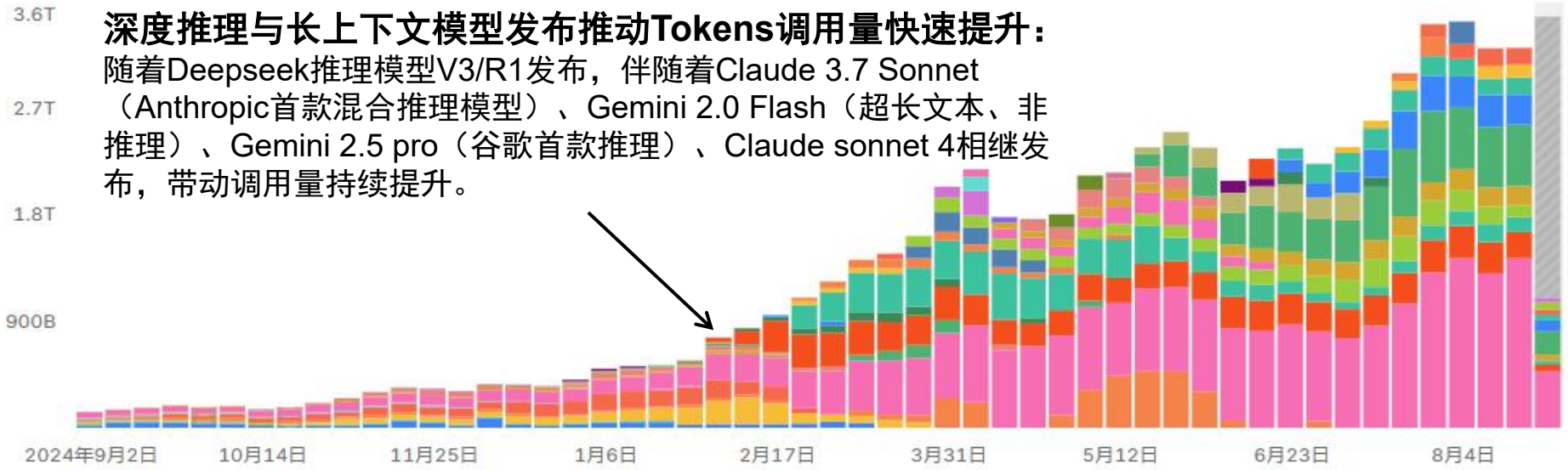
资料来源：Openrouter，国信证券经济研究所整理　注:Openrouter市场份额数据是基于其平台上的第三方API调用量统计

大模型Tokens调用量变化：过去半年模型周Tokens消耗量增长4.7倍

国信证券 GUOSEN SECURITIES

图：近一年大模型Tokens调用量周度变化

深度推理与长上下文模型发布推动Tokens调用量快速提升：
随着Deepseek推理模型V3/R1发布，伴随着Claude 3.7 Sonnet（Anthropic首款混合推理模型）、Gemini 2.0 Flash（超长文本、非推理）、Gemini 2.5 pro（谷歌首款推理）、Claude sonnet 4相继发布，带动调用量持续提升。



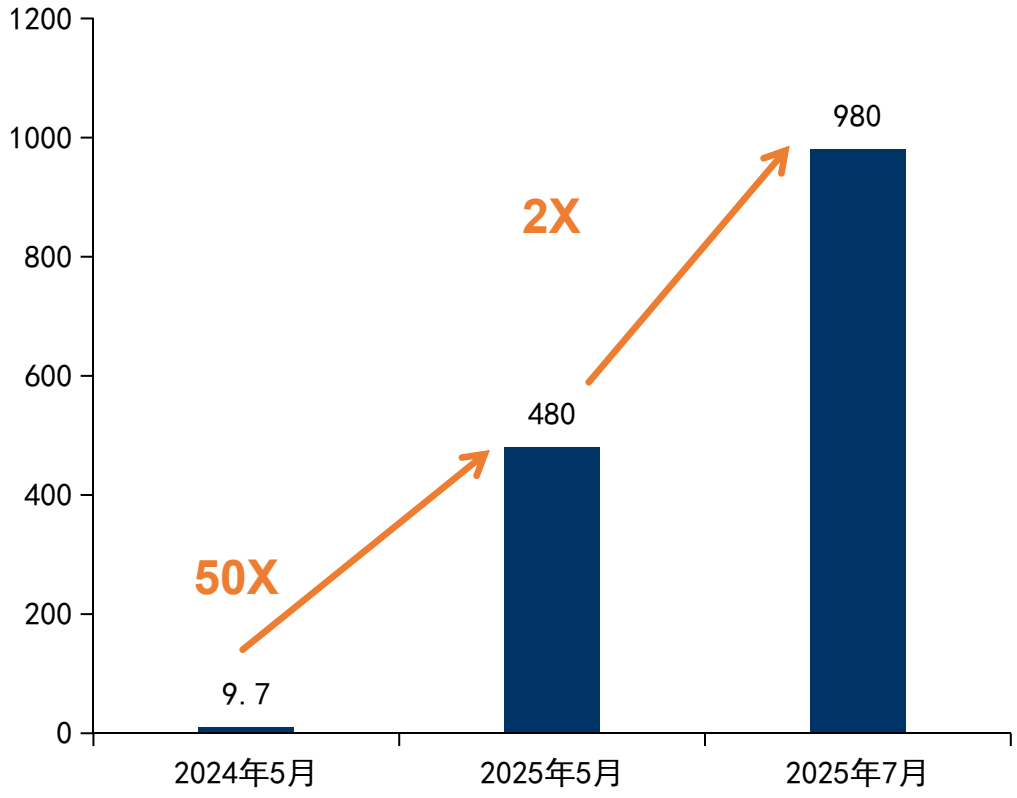
1.	Claude Sonnet 4 by anthropic	537B tokens ↑4%	6.	DeepSeek V3.1 by deepseek	143B tokens new
2.	Gemini 2.5 Flash by google	254B tokens ↓5%	7.	Qwen3 Coder by qwen	138B tokens ↓6%
3.	Gemini 2.0 Flash by google	183B tokens ↓33%	8.	Claude 3.7 Sonnet by anthropic	129B tokens ↑0%
4.	DeepSeek V3 0324 by deepseek	162B tokens ↑2%	9.	R1 0528 (free) by deepseek	102B tokens ↓6%
5.	Gemini 2.5 Pro by google	145B tokens ↑7%	10.	Grok 4 by x-ai	96.2B tokens ↑396%

资料来源：Openrouter，国信证券经济研究所整理　截止2025年8与25日数据，注:Openrouter市场份额数据是基于其平台上的第三方API调用量统计

谷歌Tokens使用量分析：24年5月至今使用量增长近百倍

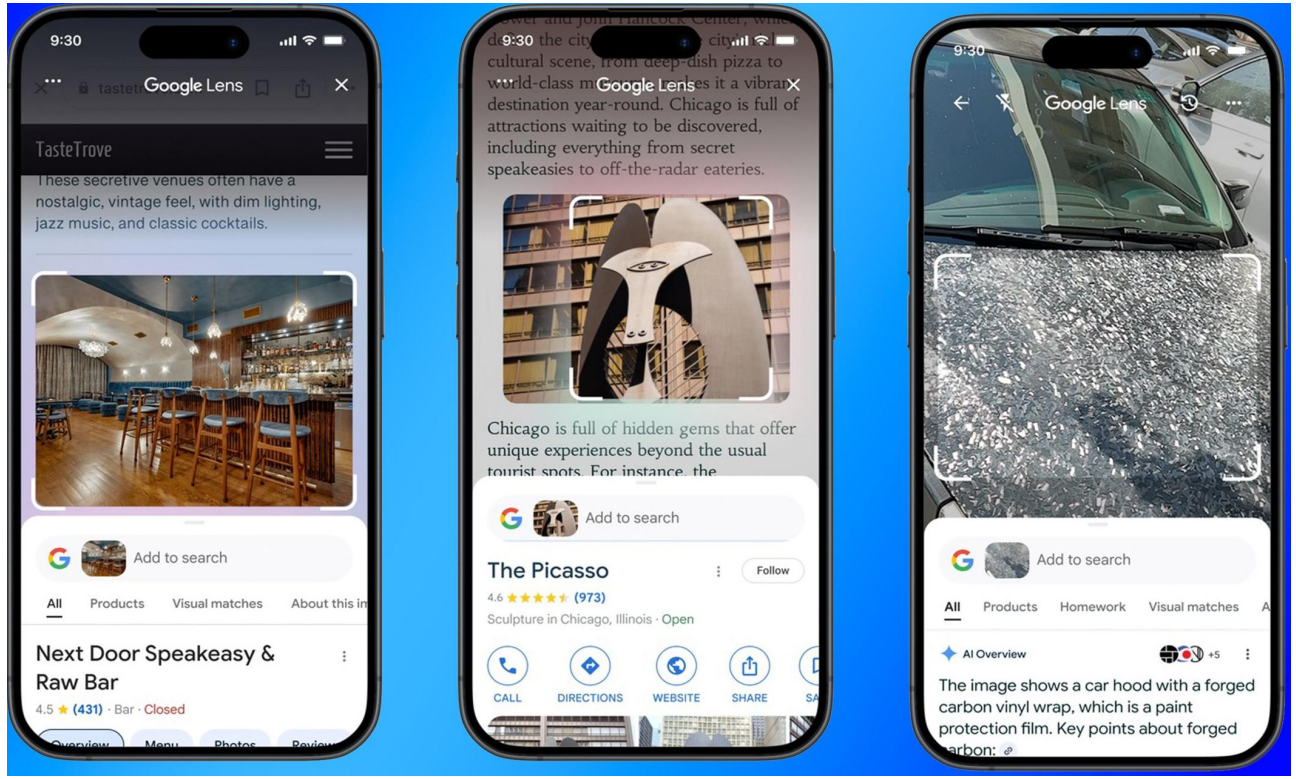
- **25年7月谷歌Token使用量高达980万亿（Trillion），相比去年五月的9.7万亿增长了近100倍。**巨大的增幅主要来自搜索业务中引入的AI Overviews（24年10月推出、25Q2 MAU达到20亿）、Lens（拍照搜索，25Q2 同比+70%）、Circle to Search（画圈搜索，已在3亿多部安卓设备上推出）等新功能，以及Workspace等产品线对Gemini模型的深度集成。

图：谷歌月度tokens调用量变化（万亿tokens）



资料来源：谷歌财报会，谷歌IO大会，国信证券经济研究所整理

图：谷歌AI Overviews与Lens搜索



资料来源：谷歌官网，国信证券经济研究所整理

谷歌Tokens使用量分析：外部B端与C端调用量占比约11.6%

国信证券 GUOSEN SECURITIES

- **外部Gemini C端调用量：Gemimi聊天助手用户数与调用量仅次于ChatGPT与MetaAI，25年3月Gemini月token消耗约为4.2T，财报会披露7月Gemini APP MAU 4.5亿、则token消耗量约5.4T（占比0.6%）。**（假设一次ChatBot对话Token消耗为1k，1个英文字符≈0.3个Token，1个中文字符≈0.6个Token。）
- **外部Gemini系列B端调用量（企业级API）：我们测算7月B端tokens消耗量约为108万亿（Trillion）、占比总消耗量980万亿的11%。**其中包括根据企业直接API调用（主要为云厂中大客户）消耗量占比约10.4%以及通过聚合平台分发（主要为长尾客户）占比约0.6%。
- 计算谷歌7月tokens消耗量外部占比总计约11.6%。

表：2025年3月 AI聊天产品用户与调用量相关数据测算

公司	产品	DAU（百万）	每日查询次数（百万）	每DAU查询次数	用户占比	查询量占比	测算月tokens消耗量（T）
OpenAI	ChatGPT	160	1,200	**7.5**	49%	71%	36
Meta	MetaAI	100	200	2	31%	12%	6
Google	Gemini	**35**	**140**	4	11%	8%	**4.2**
xAI	Grok	15	75	5	5%	4%	2.25
DeepSeek	DeepSeek	10	50	5	3%	3%	1.5
Perplexity	Perplexity	7	30	4.3	2%	2%	0.9

资料来源： Google反垄断调查证物，国信证券经济研究所整理测算（2025年3月28日数据）

图：谷歌B端tokens调用量占比测算

使用途径	具体数据	Token消耗量估算
企业直接API调用（主要为云厂中大客户）	公司二季度财报会披露Gemini系列B端企业用户85,000家（含LVMH、Salesforce等），假设平均月消耗1.2万亿Token（参考头部企业DBS Bank公开数据）	**102万亿Token（占比10.4%）**
聚合平台分发（主要为长尾客户）	OpenRouter平台7月Gemini占比43%，平台月调用量13.72万亿Token（周3.43T×4）	**5.9万亿Token（占比0.6%）**
B端总计		**107.9万亿Token（11.0%）**

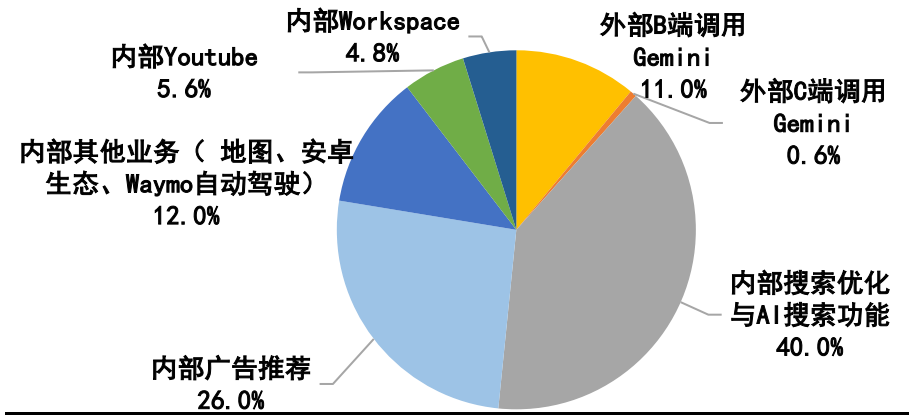
资料来源：官方财报会、DBS bank、OpenRounter，国信证券经济研究所整理测算

谷歌Tokens使用量分析：外部调用以搜索和广告场景为主

国信证券 GUOSEN SECURITIES

- **谷歌内部场景中搜索算法优化、AI搜索功能与广告推荐系统占到总token消耗量的66%**，主要用于提升搜索结果相关性、索引更新、出价策略优化、关键词智能匹配、广告素材优化、反作弊与流量过滤、模型算法迭代训练。
- 谷歌生态产品丰富、拥有七款月活超20亿的重磅产品，其他内部场景如Youtube、Workspace、地图、安卓生态、Waymo自动驾驶凭借着大MAU、高用户活跃度也有较高的Token消耗量，用于提升用户体验、交互效率。

图：谷歌月度tokens调用量场景占比



表：谷歌内部场景AI功能与token消耗测算

内部应用场景	相关数据与假设	Token 消耗量估算
搜索算法优化与AI搜索功能	**功能**：2025年6月，谷歌搜索核心AI团队负责人RajanPatel在《麻省理工科技评论》提及“为提升搜索结果相关性与多模态理解能力，Gemini模型在搜索算法优化中的月度Token消耗量已占谷歌内部AI总Token消耗的40%，主要用于索引更新、意图识别模型迭代及反作弊算法训练。” **产品数据**：AI Overviews 25Q2 MAU达到20亿、Lens（拍照搜索，25Q2同比+70%）、CircletoSearch（画圈搜索，3亿多部安卓设备上推出）	**392万亿 Token（40%）**
广告推荐系统	**功能**：出价策略优化、关键词智能匹配、广告素材优化、反作弊与流量过滤 **产品数据**：根据eMarketer 2025Q2《全球数字广告实时竞价（RTB）市场报告》谷歌广告每日约50亿日活广告请求、2.3万亿月均搜索词、1.8亿组广告素材、120亿日活广告展示	**255万亿 Token（26%）**
Youtube	**功能**：视频摘要、评论摘要、字幕翻译等； **产品数据**：月活用户30亿、其中18%使用AI视频摘要等功能（Statista 2025年7月），假设1W Token/视频、月均10个视频。	**54万亿token（5.6%）**
Workspace	**功能**：Gmail智能撰写、GoogleDocs摘要生成、GoogleSheets数据分析、幻灯片自动排版 **产品数据**：Workspace企业用户3000万（谷歌官方数据），假设其中25%订阅AI增值服务（约750万用户）、月均消耗约100万Token/人，共计7.5万亿token/月； 个人用户约10亿，假设AI功能渗透率20%、渗透用户月度使用次数20次、单次消耗1W token，约40万亿token/月；	**47.5万亿 Token（4.8%）**
其他业务（ 地图、安卓生态、Waymo自动驾驶）	**场景、功能与数据；** **谷歌地图**：实时路况预测、POI智能推荐等，全球月活用户15亿（Statista2025年数据） **安卓生态（Android）**：Google Assistant语音交互、行为预测以及谷歌照片编辑。全球活跃设备35亿台，其中搭载Android14及以上系统（支持Gemini集成）的设备占比60%（约21亿台）； **Waymo自动驾驶**:数据分析、场景生成。每周100万英里测试，每英里需100万Token环境模拟	**117.6 万亿 Token（12%）**
内部合计		**866万亿 Token（88.4%）**

资料来源：官方财报会，eMarketer，Statista，similarweb，《麻省理工科技评论》，国信证券经济研究所整理测算

谷歌Tokens使用量分析：所需推理TPU占比总AI服务器支出约10%

国信证券 GUOSEN SECURITIES

- **根据测算谷歌在25Q2 需要约27万颗TPUv6芯片来满足其推理Tokens消耗，对应新增TPU芯片支出达7亿美元，TPU服务器占比总AI服务器支出约10%、且以季度约翻倍的速度增长。**Google的AI推理需求已成为其未来基础设施建设的核心动力之一，25Q2也进一步上修全年CAPEX至850亿，26年将进一步增加。

- **25Q2 谷歌推理Token的成本占当季运营支出（不含流量获取成本TAC）的3%，占比Google搜索收入约1.4%**（传统搜索毛利率约80%、不含流量获取成本），相比AI功能带来的流量与体验改善，其成本负担尚可控。

表：谷歌tokens消耗量与所需推理TPU测算

月度推理FLOPS消耗计算	2Q24	3Q24	4Q24	1Q25	2Q25	备注
月度推理Tokens消耗 (T)	50	140	365	840	2180	
% q/q	-	180%	161%	130%	160%	
Gemini Flash TFLOPs (T)	1.5	4.3	11.2	25.7	66.7	浮点运算FLOP的数量，推理FLOP≈2*Token*模型参数
Parameters (B)	17	17	17	17	17	
% of Tokens	90%	90%	90%	90%	90%	
Gemini Pro TFLOPs (T)	2.9	8.1	21.0	48.4	125.6	1 TFLOP = 10^12 FLOPs
Parameters (B)	288	288	288	288	288	
% of Tokens	10%	10%	10%	10%	10%	假设Flash版本与Pro版本tokens负载量为9：1
总消耗TFLOPs (T)	4	12	32	74	192	
% q/q	-	180%	161%	130%	159%	
所需TPU与Capex计算						
TFLOPs /片	918	918	918	918	918	假设需要BF16 270k TPU v6 芯片
推理芯片峰值FLOPS利用率	10%	10%	10%	10%	10%	10%的模型FLOPS利用率（MFU）估计值来自GOOGL 2022年的高效扩展Transformer推理白皮书
季度在线时间 (s)	7,776,000	7,776,000	7,776,000	7,776,000	7,776,000	
需要TPU数量	**6,178**	**17,298**	**45,099**	**103,788**	**268,969**	
% q/q	-	180%	161%	130%	159%	
新增TPU数量	6,178	11,120	27,800	58,690	165,181	
% q/q	**-**	**80%**	**150%**	**111%**	**181%**	
TPU 单价	$4,500	$4,500	$4,500	$4,500	$4,500	
新增TPU Capex ($M)	**$28**	**$50**	**$125**	**$264**	**$743**	
% q/q	-	80%	150%	111%	181%	
谷歌季度Capex ($M)	13,186	13,061	14,276	17,197	22,446	
假设AI服务器占比总Capex	45%	50%	55%	60%	66%	
推理TPU服务器占总AI服务器Capex	**0.9%**	**1.5%**	**3.2%**	**5.1%**	**10.0%**	25Q2：绝大部分资本支出用于基础设施，其中约2/3服务器,1/3数据中心和网络设备。
推理TPU服务器占比总Capex	**0.2%**	**0.4%**	**0.9%**	**1.5%**	**3.3%**	

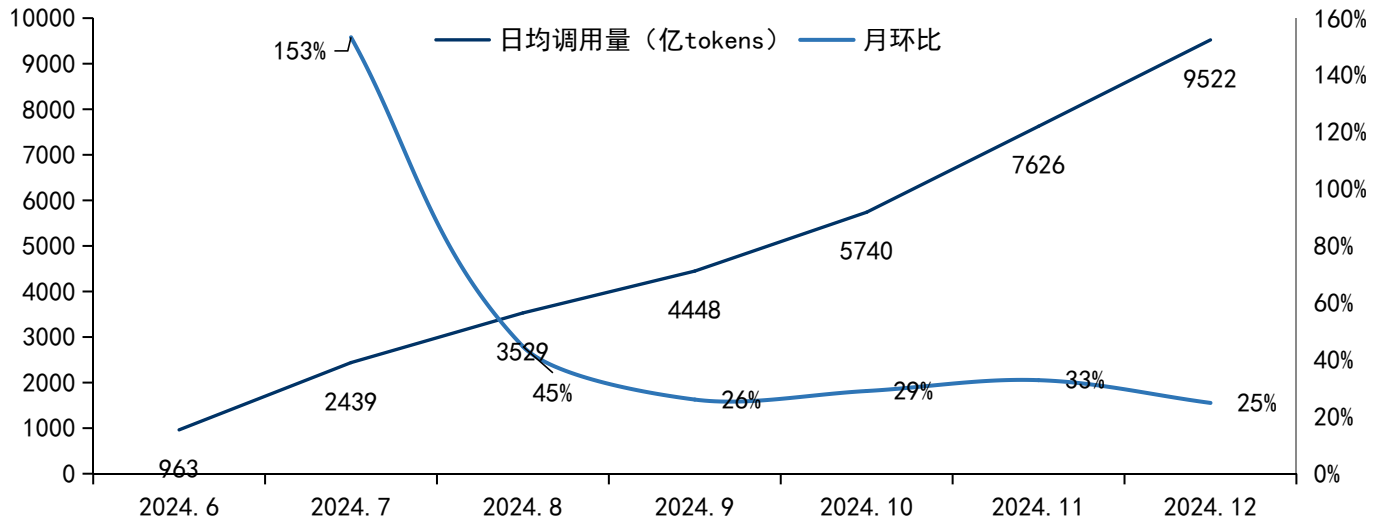
资料来源： Google反垄断调查证物，谷歌官网，SEMI Analyse,国信证券经济研究所整理测算

国内大模型Tokens调用量变化：24年下半年十倍增长，字节占比一半

国信证券 GUOSEN SECURITIES

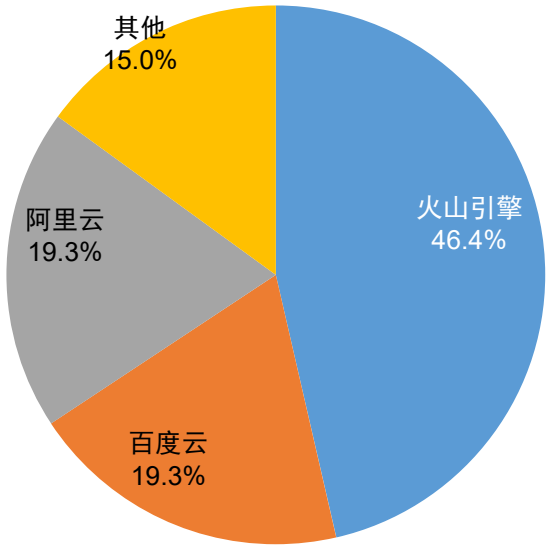
- **2024年下半年国内公有云大模型日均调用量实现十倍的增长。一方面**随着云厂商大模型能力的快速迭代，另一方面大模型厂商采用降价策略的推行MaaS平台，大模型调用量进入高速增长阶段。24年中国公有云大模型调用量累计达114.2万亿Tokens（不含出海群体使用的海外MaaS平台调用量）。
- 模态分布方面，当前调用量仍以大语言模型及文本类能力为主导。24Q4开始，语音类、图像、视频类大模型调用量已显现增长趋势。25年伴随深度思考模型出现，模型调用量进一步攀升。
- **根据IDC数据，24年公有云大模型调用量占比中，火山引擎以46%的份额位居第一。**百度智能云、阿里云紧随其后。主要是字节在国内AI公有云市场占比较大，公有云tokens量虽高，但单价低、赠送量大。BAT特别是讯飞、阿里、智谱、百度以私有化客户为主（如银行、政府），私有云tokens消耗比较难统计、并未包含其中。

图：2024.6-2024.12中国公有云大模型日均调用量（纯外部调用）



资料来源：IDC，国信证券经济研究所整理

图：2024中国公有云大模型服务调用量分布

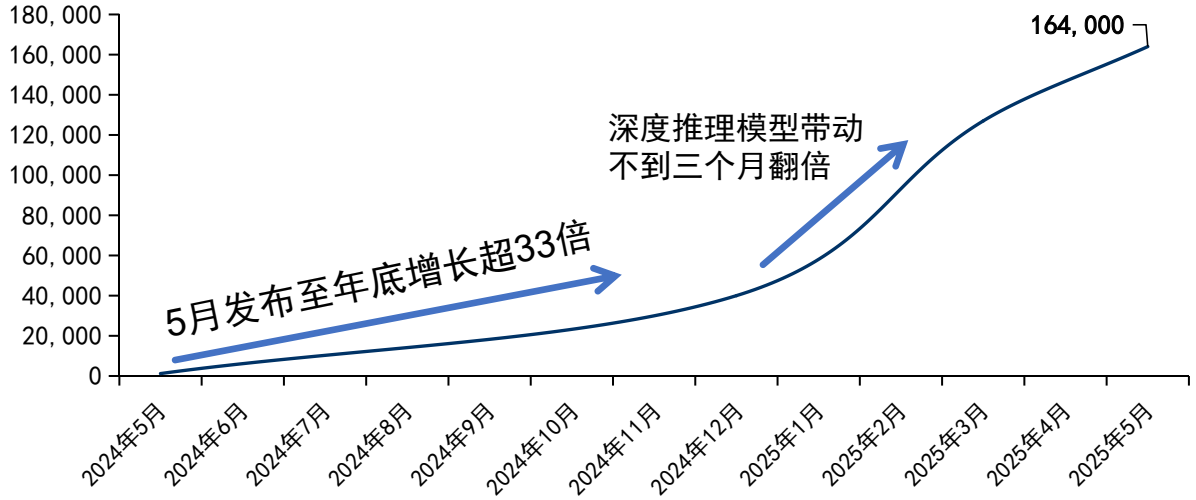


资料来源：IDC，国信证券经济研究所整理

国内豆包Tokens调用量变化与典型场景

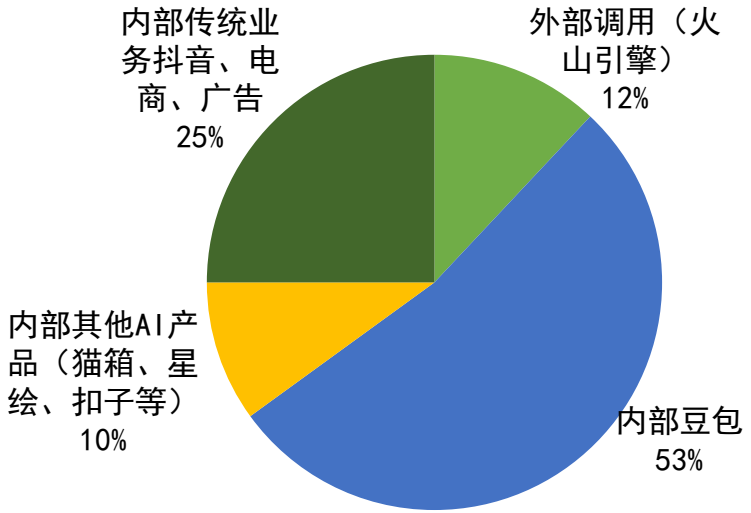
- **2025年5月，豆包大模型日均tokens调用量达到16.4万亿（包含内部与外部使用），相比2024年5月增长137倍。**2024年5月豆包大模型发布时日均tokens使用量约为1200亿。
- 伴随着模型调用量快速增长，我们测算25年底有望实现日均50万亿Tokens调用量，明年在此有望继续实现十倍增长。
- **测算豆包大模型Tokens调用量字节内部调用占比约88%，其中以豆包系产品（包含豆包、猫箱、星绘等）为主。**豆包系AI产品占63%的使用量，如豆包APP（3000W DAU）、猫箱、星绘等。字节传统业务场景调用量占比约25%，主要是抖音内电商与广告场景的AI使用。而外部调用占比12%（日均2万亿Tokens），主要集中在教育功能（拍照解题）、工具功能（文生图/视频）、服务类功能（汉得、泛微合作）。

图：2024.6-2025.5豆包大模型日均调用量变化（内外部调用），亿tokens



资料来源：官方公众号，Force大会，国信证券经济研究所整理

图：豆包大模型各场景调用量占比



资料来源：官方公众号，Force大会，国信证券经济研究所整理

国内阿里千问大模型Tokens调用量变化与典型场景

国信证券 GUOSEN SECURITIES

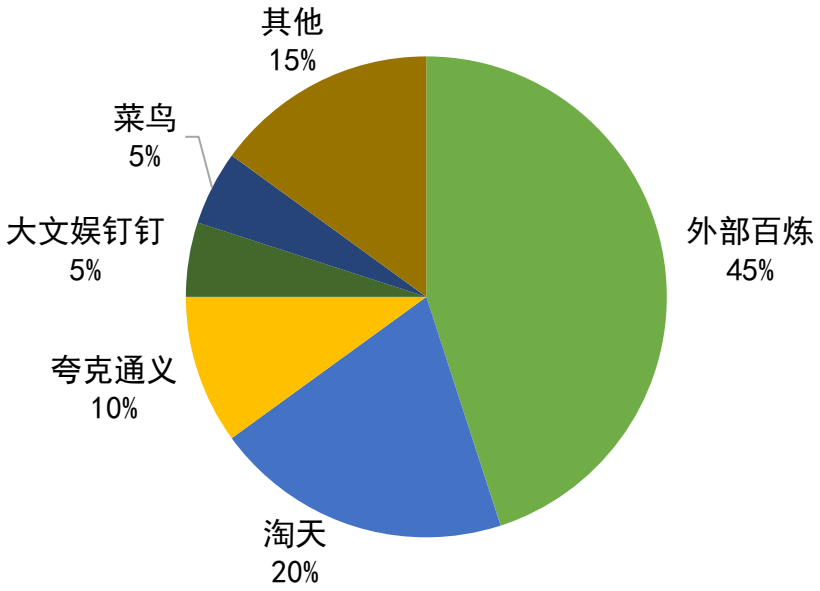
- **我们测算千问大模型2025年5月日均约9万亿tokens，其中内部业务占比约55%。**阿里以中大型B端客户为主，Q2处于示范项目构建早期、tokens消耗量较少，三季度测算tokens消耗有望达日均20万亿tokens，年维度有望呈10倍增长。
- **阿里Tokens调用量增长驱动力：**

① 客户渗透：测算云上客户中AI渗透率26财年有望从9%-10%提升至15%-20%，26财年结束AI客户数有望增至60-80万。

② 技术升级：从“调用模型”转向“智能体应用”，单次对话tokens消耗从2000-3000跃升至2万-10万。

③ 内部业务全覆盖：2025年下半年全面推进AI作为基础组件，内部调用量从5万亿向10万亿迈进。

表：阿里各场景日均tokens调用量占比



资料来源：官方公众号、国信证券经济研究所测算整理测算

表：阿里各产品落地Agent功能

Agent落地产品	Agent功能
C端夸克	拥有1.52亿月活（约4,000万日活）。重点Agent包括智能辅导、作业批改、深度研究、高考志愿填报和AI PPT生成等，通过智能体交互重构用户服务。
本地生活：飞猪	7-8月（三季度）推出“问一问”智能体，整合规划、酒店、预算和旅游攻略助手，10-20分钟生成旅行方案供客户选择。
汽车：智己	等新能源车将车载助手与本地生活、地图打通（路上订餐、到家取货）。
教育：精准学、新东方	打造AI老师智能体针对用户弱项提供个性化讲解。

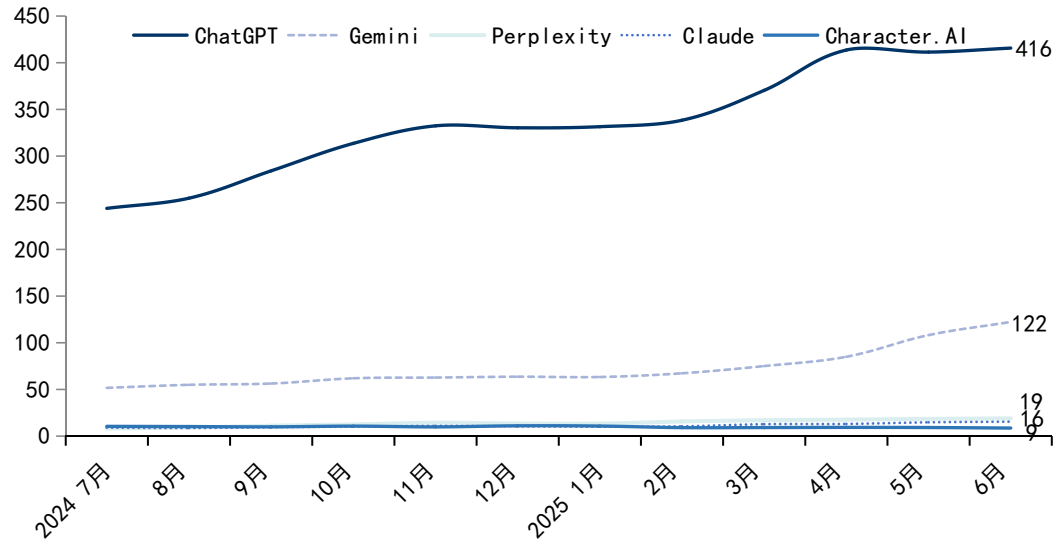
资料来源：官方公众号、国信证券经济研究所测算整理测算

目录

重磅C端搜索产品访问量：主要依赖模型能力与生态导流

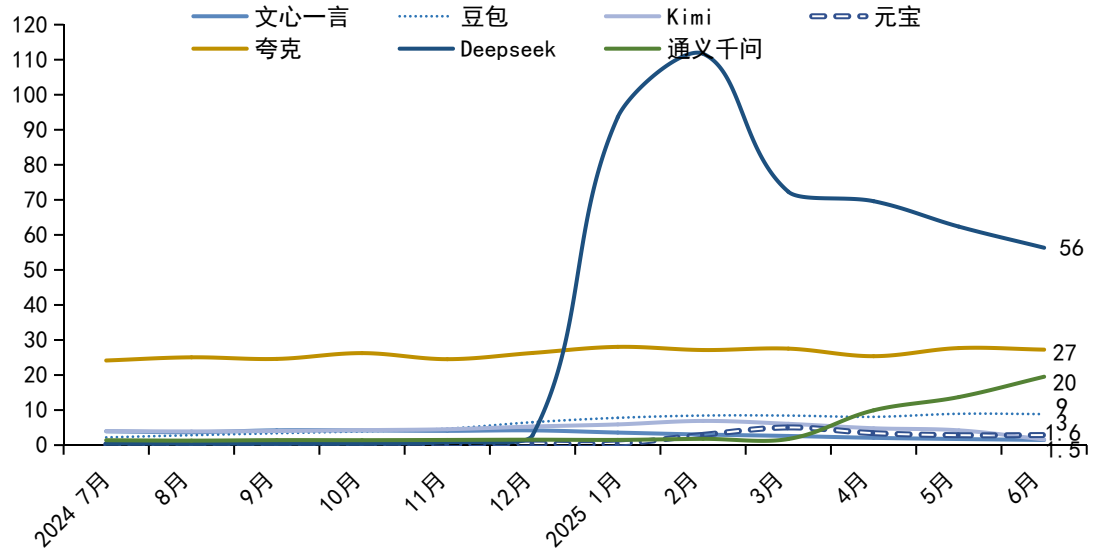
- **国外AI搜索产品：**ChatGPT在整个观测期内始终保持显著领先地位，其用户规模远超第二名Gemini，显示出较强的用户粘性和品牌优势。这一领先地位可能源于其早期市场进入、强大的模型性能以及广泛的应用集成。Gemini伴随着2.0系列模型发布与使用加速增长。其相比之下，Claude、Perplexity和Character.AI等尚未形成显著竞争力。
- **国内AI搜索产品：**DeepSeek自2024年年底起持续领先，得益于其模型能力与春季热度、峰值超1亿，近期MAU出现较大回落，系其R2模型发布计划延后、响应速度慢、性能不稳定等因素影响。夸克凭借浏览器和搜索入口的流量支撑，MAU也保持在较高水平。通义千问自3月起MAU显著上升，反映出阿里生态的有效导流和在电商、智能助手等多场景应用的推动效果。相比之下，Kimi的MAU在2025年后回落明显，可能因长文本处理功能面临竞争且缺乏强生态支撑，随着K2模型发布流量有所改善。文心一言访问量呈现持续下行态势，系技术迭代节奏及市场竞争环境等因素综合影响。

图：近一年AI应用网站月访问量变化（M）



资料来源：Similarweb，国信证券经济研究所整理

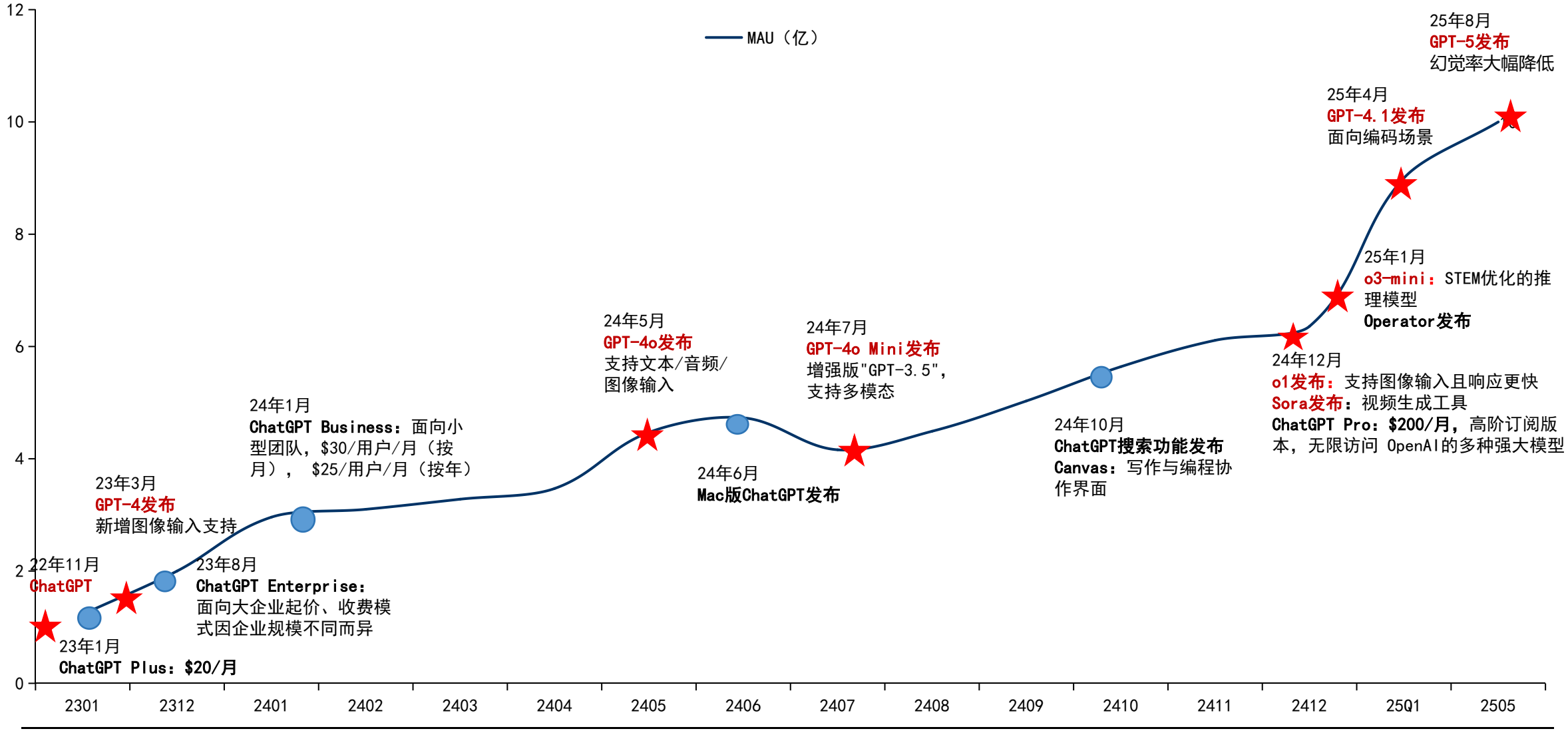
图：近一年AI应用网站月访问量变化（M）



资料来源：Similarweb，国信证券经济研究所整理

ChatGPT发展历程

图：ChatGPT月活与产品迭代变化



数据来源：ChatGPT官网、华尔街新闻、Xsignal、国信证券经济研究所整理

海外重磅C端产品介绍：ChatGPT、Gemini、Claude、Perplexity、Manus

国信证券 GUOSEN SECURITIES

- 海外科技大厂扶持的C端Agent/搜索产品发展迅猛，两年头部产品MAU增长至5亿/10亿量级、年化收入破百亿、估值达千亿美金。
- 目前产品发展仍以用户增长与产品体验提升、模型技术迭代为主，商业化变现主要通过订阅与API调用实现。

表：海外重磅C端产品数据汇总

产品	OpenAI/ChatGPT	Anthropic/Claude	谷歌/Gemini	Perplexity	Manus
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	2025.5：1000 万美元 ARR 2025.8：9000万美元ARR（官方披露）
DAU	2023：约2500W 2024：3860W 2502：1.23亿 2508：约1.8亿	2023：约500W 2024：800W 2505：1847W	2024：超1500W 2503：3500W		
MAU	2307：1亿 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	2503：月访问量峰值超2000万 2506：5月开放注册后月访问量回落至约1500万
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）	原为2美元/任务，后改为分级订阅（202508披露收入结构:19 美元 / 月基础版占比 35%，99 美元 / 月专业版占比 50%，199 美元 / 月企业版占比 15%）
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	多模态自主决策框架 +Claude3.5/阿里Qwen等
主要合作伙伴	微软Azure	亚马逊AWS	谷歌GCP	亚马逊AWS	阿里通义、微软Azure
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	2025年5月完成了7500万美元融资，估值金5亿美元

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

国内重磅C端产品介绍：豆包、夸克、Deepseek、元宝、通义千问、Kimi

国信证券 GUOSEN SECURITIES

- **产品定位：豆包、夸克、元宝依托巨头生态的C端超级应用，侧重场景整合与用户规模扩张。**通义千问、DeepSeek面向企业与开发者的大模型平台，以技术性能与开源策略吸引B端客户，通义千问B端收入占比超80%。Kimi通过专业场景优化与高成本投流争夺细分市场。
- **技术路线:豆包、夸克、通义千问以自研模型为主，**结合多模态与生态联动。**元宝、DeepSeek采用“自研+开源”双轨策略，**平衡技术自主性与商业化效率。Kimi坚持自研路线，聚焦专业场景技术突破。
- **变现：豆包、夸克、DeepSeek基本形成“基础免费+增值付费”的模式。**元宝、Kimi仍处于变现探索期，依赖生态流量与投流驱动增长。

表：国内重磅C端产品数据汇总

产品	豆包	夸克	Deepseek	元宝	通义千问	Kimi
DAU	2024年底：900W 202502：超1000W 202505：APP 2700W	202502：3430W	202502：2000W 202506：1200W	202502：APP 350W	202505：70W	2024年底：300W
MAU	2024年底：7523W 202502: APP 8198W 202505：1.15亿 202506：APP 1.26亿	202502：超2亿 202506：APP 1.6亿	202502：APP 6181W 202505：2.6亿 202506：APP 9400W	2024年底：291W 202502：APP 1312W 202504：APP 2636W 202506：APP 4000W	202505：5800W	2024年底：2100W 202506：APP 2350W
变现模式	API授权、B端合作、广告分成，计划通过订阅服务与企业解决方案拓展收入；强化订阅服务（如 Pro 版本），探索硬件生态（如 AI 耳机、智能家电）	网盘会员订阅（年费 99-158 元）、AI 文档处理增值服务；扩展AI Agent 服务，探索教育、办公场景付费解决方案	API授权、订阅服务、广告分成，通过开源策略吸引开发者与中小企业，强化企业定制服务，探索内容创作与知识付费生态	探索会员服务与企业级 API 授权，尚未形成清晰变现路径	企业合作、API 授权、云服务集成，通过降价策略扩大市场份额；企业级解决方案、开发者授权，探索 C 端基础服务免费 + 增值服务付费模式。	广告、会员订阅，测试“打赏” 功能（高峰期优先使用权），探索 B 端定制服务
底层模型	自研豆包 1.5・深度思考模型（MoE 架构，200B参数）	通义千问 Qwen2.5-Max、QwQ-32B 模型	DeepSeek-V3（6850亿参数开源版本）	腾讯混元大模型 + DeepSeek-R1 双模型架构	Qwen2.5-Max、Qwen3-235B	自研模型，支持多模态交互与深度推理
主要合作伙伴	字节	阿里	Deepseek	腾讯	阿里	月之暗面

数据来源：公司官网、Similarweb、AI产品榜、QuestionMobile、国信证券经济研究所整理

Manus和Perplexity案例：初创AI公司警惕通用Agent陷阱

国信证券 GUOSEN SECURITIES

- **能力边界与市场需求的错配**：近期Manus与Perplexity的用户增长、商业化变现遇到一些风波，由于“全场景覆盖”的产品定位，导致核心功能相比其他头部产品较弱、成本结构承压及用户留存相对较低。
- **初创AI公司的核心竞争力并非覆盖范围，而应是场景穿透力**。在OpenAI、谷歌等巨头已垄断通用场景以及AI模型技术的背景下，唯有聚焦垂直场景、控制成本、建立用户依赖，才能避开陷阱，正如斯坦福研究所说：“真正的AI创业机会，藏在员工渴望但未被满足的细分任务里，而非看似宏大的全场景蓝图中。”

表：Manus和 Perplexity的困境

困境	Manus	Perplexity
全场景定位导致核心能力较头部产品弱	2025年3月6日正式发布，以"通用智能体"为定位，强调自主执行复杂任务的能力，依托“AI Agent全场景（编程生成-数据爬取-视频制作-金融建模）”概念营销。 吸引大量尝鲜用户，但实际功能并未达到较优效果，如视频生成卡顿、数据爬取失败率约60%、非结构化任务崩溃率达17%，相比ChatGPT基础任务准确率达92%。	Perplexity聚焦于深度研究与企业级搜索解决方案，从实时搜索切入，逐步扩展至医疗、教育、企业服务领域。 各场景均存在浅尝辄止问题：医疗合规模块虽通过HIPAA认证，但医生渗透率仅3%（缺乏临床数据训练），教育版学术数据库覆盖量仅为ChatGPT Academic的40%。
用户留存较头部产品较低	近期用户增速较6-8月明显放缓，30天付费用户留存率约为50%+，6个月留存率约30%+。	Perplexity Pro 6个月留存约50%，低于ChatGPT Plus的70%+与Claude、Gemini的60%+，直接制约LTV与增长质量。
成本结构承压	Manus单次任务调用多模型API，算力成本是ChatGPT的3倍，而平均ARPU约18美元/月、不足后者的80%。2025Q2不得已关闭多模型API调用（改用单一模型）降低70%算力成本。	Perplexity 2024 年三方模型支出 5700 万美元（超年度总收入约 3400 万），2025Q2 通过与 Azure 合作降低 30% 云成本，才实现 ARR 盈利。Perplexity 接入 Elsevier 学术数据库后，内容授权成本占营收比达 18%，高于行业平均 12%。
变现困境	2025Q2 从 “2美元/任务” 改为 “16美元 /月订阅制”，涨价幅度较大，直接导致大量付费用户流失；	最初强调 "不受广告驱动"，到 2024年4月宣布将引入广告，探索混合变现路径，但其2024Q4 广告收入仅 2 万美元，负责广告与购物业务拓展负责人Taz Patel与2025年8月离职，AI搜索广告模式遭遇困难。
合规与版权问题	美国出口管制导致虚拟化引擎技术无法升级，多模态功能延迟上线3个月	2024年7月多家媒体（如《纽约时报》、《福布斯》等）就版权问题发出

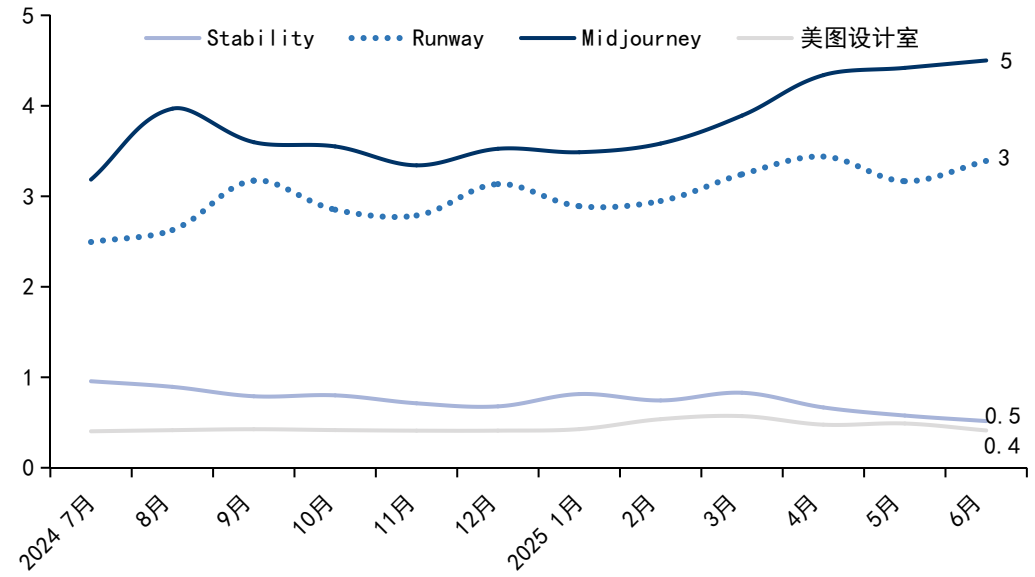
数据来源：公司官网、Xsignal、Similarweb、彭博社、国信证券经济研究所整理

图像类AI应用网页访问量：Midjourney与可灵流量领先

① **国外：**Midjourney从网页访问量上保持领先，反映出其在生成质量和创意社区方面的持续吸引力；Runway和Stability AI则依托视频生成和开源模型策略占据稳定市场空间，但增长相对平缓。

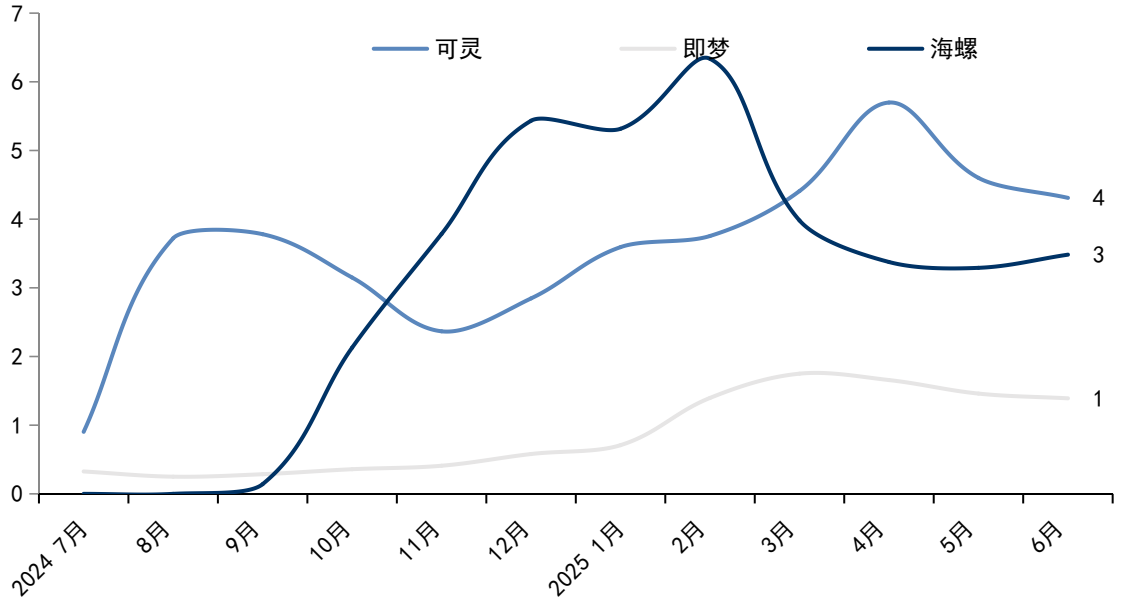
② **国内：**可灵上线后用户增长显著，自3月以来MAU持续领先，受益于其视频功能的快速迭代与进步以及与快手生态系统深度整合；海螺AI前期增长迅速，但自2025年起MAU逐步下滑至低位，可能系其功能同质化及用户使用频率有限；即梦MAU始终处于较低水平，未见显著突破，反映其产品认知度和应用场景仍待加强。美图网站月访问量处于较低水平，因其重点产品在APP移动端、网页用户较少导致。

图：近一年AI应用网站月访问量变化（M）



资料来源：Similarweb，国信证券经济研究所整理

图：近一年AI应用网站月访问量变化（M）



资料来源：Similarweb，国信证券经济研究所整理

图像类AI产品：可灵 、即梦、美图设计室、Midjourney、StableDiffusion

表：图像类AI产品数据汇总

产品	Stable Diffusion/Stability AI	Runway	可灵	美图设计室	Midjourney	即梦
年化收入	202506：ARR 1亿美元	2024：8400万美元 25年底：2.65亿美元	202503：年化1亿+美元 202506：月收入1.5亿元	2024：2亿元（+100%）	2025：ARR 5亿美元	2508：月收入2200W元
MAU	202506：Web 95万	2024：数百万 2506：Web 340万	2024年底：用户数600W 202506：Web MAU 430W 202507：用户数4500W（70%专业P端、30%B端客户；70%海外，30%国内）	2024年底：1890万，付费110W 2506：Web 40万 25H1：订阅用户数同比+56%	2506：Web 450万	2506：Web 139W 2508：C端DAU 600w，月付费用户20W+未来三年目标：DAU 4500-5000w
变现模式	- 会员：Pro $10/月、Max $20/月 - 企业定制化方案	- 会员：$15-95/月 - 企业定制化方案 - 按credits计费（$0.05/秒）	- 会员： 66-1314元/月 - 企业API服务	- 个人版： （48元/月、128元/季、358元/年） - 团队版：776/年	- 会员：基本$8/月、标准$24/月、Pro$48/月、Mega$96/月	- 会员：基础79元/月、标准239元/月、高级649元/月
产品特点	- 开源灵活，可自由调整参数适配多元场景 - 质量高，应用广泛，基于扩散模型生成高分辨率、逼真图像，细节强	- 整合Gen-4电影级视频生成、图像生成及Adobe插件，互动性强，质量高，强调专业级影视制作能力	- 支持 1080p、多种宽高比，能模拟物理特性，实现复杂概念组合 - 多模态创作，文生视频、图生视频、虚拟试穿等功能集成	- 功能全面，集合了AI商拍、图像处理和AI设计等工具，一站式解决电商设计需求 - 提供商用版权素材，确保作品合法性和原创性	- 可创作高精度艺术作品 - 支持自定义色彩、纹理与构图参数，实现个性化表达	- 依托字节云端算力集群，覆盖全品类的艺术创作矩阵，内置百款场景化提示词模板，日均衍生作品 10W+
场景	工业设计/游戏开发公司	专业影视创作/广告公司	短视频创作者/中小企业	电商卖家/社交媒体运营	艺术家/创意工作者	普通用户/创作者
底层模型	Stable Diffusion 3.0/Stable DiffusionXL	Gen-3 Alpha、 Gen-4 Turbo	可灵 2.0/可图 2.0	Miracle Vision 4.0、DeepSeek R1	Version 6.1、Niji 6	Seedream 3.0
主要合作伙伴	Stability AI	谷歌、英伟达	快手	美图	Meta	字节
估值	2024：10亿美元	24年7月融资4.5亿美元，估值30-40亿美元	60亿美元		2023：$100亿	

数据来源：公司官网、AI产品榜、Similarweb、Sacra、国信证券经济研究所整理

编程类AI产品：Cursor、GitHub Copilot、Codeium、通义灵码

编程类AI产品场景优势极高，所需技术与大模型技术能力高度吻合、面向高同质化、客群付费意愿强且市场前景广阔。产品思路成熟，已有经下游验证的成熟模式，后续随模型性能提升将逐步拓展广泛的潜在客群。

头部企业增长较快，通过微调模型算法、代码嵌入，能提升产品可用性且精准把握程序员需求痛点，形成技术与设计层面的优势。

表：编程类AI产品数据汇总

产品	Cursor	GitHub Copilot	Codeium（更名为Windsurf）	通义灵码
年化收入	2411：6500万美元（+6400%） 24年底：1亿美元（9900%）ARR 2505: ARR 5亿美元	2024：3亿+美元	2024：1200万美元，估计25年2月达4000万美元	
DAU/MAU	2503：超200万 2506：Web MAU 670W	2506：GithubWeb MAU 1.2亿 2507：Copilot用户数超2000万	25年初MAU：数十万 2506：Web MAU 170W	
变现模式	- 个人订阅：Hobby 免费、Pro $20/月、Pro+ $60/月、Ultra $200/月 - 企业订阅：Teams $40/月、Enterprise 定制计费	- 个人订阅：Pro $10/月、Pro +$39/月 - 企业订阅：Business $19/月、Enterprise $39/月	- 个人订阅：基础功能免费、Pro $15/月 - 企业订阅：Teams $15/月、Enterprise $60/月起 - 两百人以上企业定制化方案	- 个人订阅：基础版免费、专业版￥59/月 - 企业订阅：标准版￥79/月、专属版￥159/月 - 企业私有化部署方案
产品特点	- 灵活性高，调用GPT及Claude模型，并自定义模型Cursor-small - 互动性强，交互式编辑、支持团队协作、自然语言对话代码库 - 用户友好，代码质量高、简化操作流程、维护流程简单	- 原生嵌入GitHub生态，支持IDE实时补全、多文件编辑及自动化测试修复 - 支持多模型切换，适配不同编码场景- ISO 27001认证 + 企业级安全审计，支持Workspace全仓库级修改（CI/CD自动化）	- 支持70+编程语言，集成多模型，自定义选项丰富 - 交互友好，集成40+IDE,提供高质量代码建议和实时聊天支持 - 强调代码数据隐私保护，支持本地化部署	- 内置Qwen大模型，支持MCP服务，满足不同场景需求 - 智能体协同，多智能体框架自主拆解任务，实现工程级多文件修改 - 企业定制化，私域知识增强、VPC专属部署，适配企业研发标准
场景	复杂工程开发	企业团队协作（标准化开发流程）	个人开发者与低成本需求	中文环境下的开发需求（金融政务项目）
底层模型	Claude Sonnet 4、o3-pro、GPT-4.1及自定义模型Cursor-small等	GPT-4.1、Claude 3.5/3.7/4 Sonnet、Gemini 2.5 Pro等	OpenAI、Claude、Gemini、xAI 等	Qwen3系列模型、DeepSeek系列模型
主要合作伙伴	OpenAI	微软	OpenAI	阿里云
估值	25年6月融资9亿美元，估值99亿美元		25年2月以 28.5 亿美元的估值进行新一轮融资	

数据来源：公司官网、微软业绩会、Similarweb、Sacra、Crunchbase、TechCrunch、国信证券经济研究所整理

从通用到垂类：对技术要求越低、产品要求越高，越容易商业化

国信证券 GUOSEN SECURITIES

- **为什么Cursor发展和商业化进展显著更快？通用搜索助手、AIGC Agent、编程类AI助手将怎样发展与演绎？**
- 从通用搜索助手→AIGC Agent→编程类 AI Agent，典型呈现从通用AI产品→垂类（模型）AI产品→更垂类场景AI产品，**我们判断越从通用到垂类，对产品技术与门槛要求越低，对场景理解、产品功能与操作交互体验要求会越高，竞争格局会越分散。同时，越向垂类场景/功能，目标用户（范围小）对功能感知会越深入、越容易商业化闭环（形成付费意愿），越容易直接变现。**

表：通用搜索助手、AIGC Agent、编程类 AI助手比较

	通用搜索助手 Agent	AIGC Agent	编程类 AI Agent
对AI大模型需求	极高，需自研多模态大模型，对模型综合性能如准确性、上下文长度、时延有较高要求（领先的模型厂商才具备通用Agent助手的资质）	高，需训练/微调垂类多模态大模型，对模型跨模态（文本、图像、音频、视频）能力、一致连贯性要求高	一般，可调用开源或通用模型，侧重代码生成/优化/检测
其他技术难点	• 需具备各个渠道（网页、社交媒体、API ）信息源、信息源部分决定了搜索结果质量，需要实时处理海量动态数据。 • 对模型工程化部署要求较高，要求高并发与低时延，同时要求合规与数据隐私。 • 需要丰富的API和开发者生态，决定了能力边界。	• 需对生成内容控制强、交互性好，能生成高质量、连贯且符合用户需求的内容，保证内容多样性和一致性。 • 能深入理解用户意图理解，激发用户创造力； • 与现有AIGC使用场景有较好的融合打通； • 模型推理计算资源与效率优化；	• 生成语法正确、逻辑严密且符合最佳实践的代码，适应各种代码规范要求，避免安全漏洞和潜在风险； • 多语言支持、多环境适配、跨平台兼容，与开发环境和工具链深度集成； • 响应快，低操作门槛和高易用性；
产品付费率	ChatGPT：约4%（25年4月：5亿MAU、付费2000万） Claude：约2%（2024年底用户数1.89亿，付费417W）	美图设计室：约5%+（2024年底用户数1890万，付费110W） 可灵：约2%+（用户数4500W，付费用户测算约200万+）	Cursor：约33%（2025年初MAU约200万，66万付费） GitHub Copilot：约12%（ 2025年初用户超1500万，付费用户约180万）
竞争格局趋势判断	模型厂商以及其背后科技巨头形成垄断或寡头格局	准寡头格局，现有平台内容科技厂商与拥有垂类AIGC场景的软件厂商为主要玩家	科技大厂与AI初创公司等较多竞争者，能和现有开发生态集成的均具备能力
商业模式判断	不直接向用户收费，广告或交易撮合抽成	效果付费、订阅制	订阅制为主

数据来源：公司官网、Windows Central、Similarweb、AI产品榜、国信证券经济研究所整理

硬件类Agent产品类型介绍

国信证券 GUOSEN SECURITIES

- 未来的交互将软硬一体，硬件终端变成人类感知的延伸，与外界交互、感知并处理信息。
- **短期AI手机与PC仍为核心交互载体**:手机为高频个人设备，随AI发展和数据积累，能深化用户理解、强化个人助理属性，提升体验；PC作为核心生产力工具，硬件可承载更强AI模型，有望成为Agent早期落地载体，赋能高效办公与创作场景。
- **长期多终端构建全域交互生态：**交互载体将拓展为“AI手机+多元可穿戴设备”的组合形态，向无感化演进，新形态无屏设备、语音或取代触控成主流交互。智能体（Agent）可深度嵌入各类终端，形成“持续状态监控+实时需求反馈”的闭环，实现从“被动响应”到“主动服务”的转变，无缝融入用户日常生活场景。

表：硬件类AI产品汇总

类型	产品形态	目前进展
AI 手机	作为核心Agent与流量入口，软硬件深度一体，借 AI优化交互、服务体验，多家厂商推出AI机型，AI功能聚焦文本、图像、音频编辑生成（如文生图、图像去杂物）	战略重点，高频功能获正反馈，提升高端机型竞争力，如Apple Intelligence提升支付意愿11%，成54%换机用户决策因素。正推进协议适配如MCP和底层UI技术完善，未来摄像头或成视觉感知系统；目前主要调用外部模型，逐步以自研替代
AI PC	承载更强 AI 模型，助力生产力场景，像提升办公效率、内容创作辅助等，微软等推动系统与 AI 融合	逐步从概念到商用，成为 PC 产业升级方向，硬件适配与软件生态协同完善
AI 汽车	实现智能座舱交互（语音控制、场景化服务）、智能驾驶（L2 - L4 级辅助 / 自动驾驶）	车企广泛布局，以智能座舱为主，智驾优先用成熟方案，暂未将大模型与智驾直接结合。世界模型研发中，未来或成“超级个人助手”（需L4及安全保障）
AI 智能家居	涵盖智能音箱（语音交互中枢）、智能家电（联动控制、自主场景执行 ）、中控设备（统筹家居系统）	设备互联互通更成熟，优先级低于手机和汽车，多数设备无本地算力，AI 化程度低，以嚎头为主；待手机、车端成熟后复用技术
AI 眼镜	可穿戴，支持 AR/VR 交互、实时信息推送、环境感知与辅助（如翻译、导航）	处于技术迭代与场景拓展期，消费级产品逐步丰富，在特定领域（工业、教育）试点应用
AI玩具	面向成人的萌宠型、面向儿童的聊天陪伴型多种类型	市场目前处于早期阶段，但发展迅速。Ropet自2024年12月开启众筹以来，3 个月内已在 Kickstarter 平台众筹超 20.2万美金。
AI学习机	内置丰富的学习资源，包括点读教材、题库、同步学科视频等，还具备语音交互、图文批注、AI拍照搜题等功能。	主要厂商科大讯飞、学而思、步步高、小猿学练机等，产品升级迭代中。

数据来源：公司官网、Sacra、Crunchbase、TechCrunch、国信证券经济研究所整理

硬件类Agent产品发展路线

国信证券 GUOSEN SECURITIES

- **AI硬件实现：上下游加速整合，模型厂商与芯片商合作优化适配。**硬件厂商采用端云结合，端侧处理高频隐私任务，云端应对复杂需求，轻量化模型成主流。
- AI原生程度上，短期AI提升现有应用效率，长期将成“Agentic”应用入口，用户借自然语言交互，掌握数据与生态的企业更具优势，最终模型、硬件与生态的融合度，将决定下一代智能终端主导者。

表：AI消费硬件三大发展路线

路线类型	核心代表	核心特点	典型产品案例	商业模式	核心瓶颈	现状与前景
强化原生设备渐进派	**苹果、Meta、小米、华为**	现有终端为基础，系统性引入AI能力，实现渐进式重构，不彻底颠覆原有形态	- 苹果 Apple Intelligence：本地大模型（M系列芯片支持）集成至手机、平板、电脑 - Meta Ray-Ban智能眼镜 - 小米/华为AI眼镜：集成语音助手、图像问答 - Pixel Buds、AirPods：引入大模型实现实时对话、语音操作	- **硬件销售**：带动成熟硬件销售； - **增值订阅**：提供深度健康分析、专业翻译、专属AI功能等增值服务；	- **伪AI功能争议**：用户对无真实价值的浅AI功能缺乏付费意愿，无刚需场景； - **硬件同质化**：产品形态趋同，差异化需依赖更深层次AI能力与生态协同；	以上一代硬件为基础，增强AI交互体验，路线稳健，是当前最具规模化潜力的路线
模型为核心赋能路径	**OpenAI、阿里、谷歌**	不自建硬件，通过API、SDK等接口，将AI能力注入第三方设备，做“通用智能”底层支撑	- GPT-4o：嵌入 Ray-Ban 智能眼镜、Be My Eyes，提供实时对话与视觉感知 - 阿里“通义听悟道”耳机、豆包耳机：以语音助手为核心，打造“随时唤醒、持续响应”体验	- **API/SDK收费**：按调用量（Token消耗、请求次数）向开发者、企业、硬件厂商收费； - **解决方案服务**：提供企业订阅、私有部署、安全定制等高附加值服务；	- **成本高**：模型推理成本高，无法像安卓系统低成本大规模授权 - **适配难度大**：模型与终端适配存在技术门槛，本地端时延高、功能缩水 -**平台控制权弱**：缺乏统一生态标准与分发渠道，用户侧话语权被硬件厂商隔离	灵活度高、迭代快，能快速渗透各类终端，核心是构建“模型即平台”的智能生态，让模型能力无处不在
AI原生设备探索范式	**Rabbit、Humane、GROOVE X**	跳出智能手机交互范式，弱化App、菜单与窗口，依赖大模型驱动语义理解与任务执行，探索“意图即操作”	- Rabbit R1：实体屏幕 + 滚轮交互，199 美元 - Humane AI Pin：无屏投影，2025 年 2 月 28 日停止所有在线功能 - Loona、Ema 陪伴机器人：语音交互、情绪识别	- **硬件高溢价**； - **生态订阅**：通过云服务、AI功能更新、专属交互/内容包等订阅； - **生态延伸**：探索配件销售、IP授权；	- **功能性不足**：实用性与体验不及手机等成熟硬件； - **陪伴类产品缺陷**：如续航短、互动单一；	交互与系统结构具范式创新（类似早期 Palm OS），但受体验成熟度、稳定性、用户认知限制，不确定性高

数据来源：腾讯研究院、国信证券经济研究所整理

目录

B端Agent/Copilot产品介绍和市场规模预期

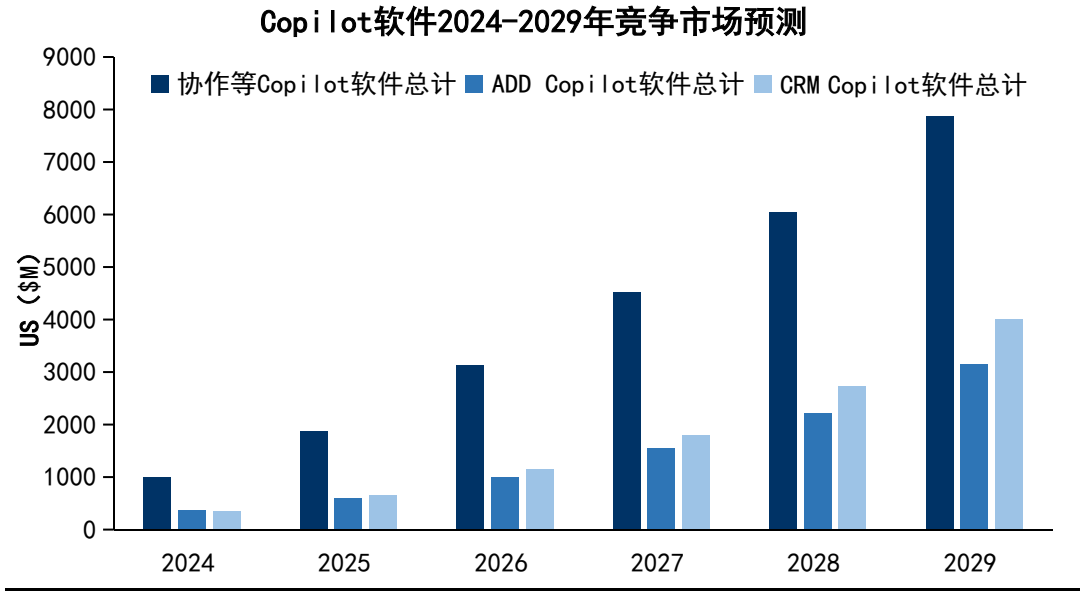
根据IDC对新兴Copilot市场的分析和估计，主要将Copilot应用分为以下三类。其预计Copilot将深度融入办公、开发与客户管理等场景，2029年市场规模有望达到近200亿美金，2024-2029年CAGR为54.3%。

① **协作、内容工作流和管理应用程序：**通过共享信息、通信和频繁的流程，协同团队工作。典型产品包括：Microsoft（Office）365的Copilot、GoogleWorkspace的Gemini、AmazonQBusiness等；

② **应用程序开发和软件质量应用程序（ADD）：**开发人员主要用于构建、测试和部署软件以及处理、集成、治理和分析数据的工具和平台。典型产品包括：GitHubCopilot、GeminiCodeAssist等

③ **客户关系管理（CRM）应用程序：**在组织内自动执行面向客户的业务流程，而不受行业特性（即广告、营销、数字商务、销售、客户服务和联络中心）的影响。典型产品包括：Salesforce Einstein Copilot、Amazon QinConnect等。

图：Copilot软件收入发展预测



资料来源：IDC，cbinsights，国信证券经济研究所整理

表：各公司主要Agent产品

供应商	产品	供应商	产品
Microsoft	Copilot系列（GitHub Copilot、Copilot for Azure、Copilot for Dynamics 365）	**Salesforce**	Einstein Copilot
Google	Gemini for Google Workspace	**Salesforce**	Slack AI
Google	Google Duet AI（Gemini for Workspace）	**SAP**	Joule
IBM	watsonx Orchestrate	**ServiceNow**	NowAssist
AWS	Amazon Q for Business	**UIPath**	UIPath AutoPilot
NICE	NICE Enlighten Copilot	**Yellow.ai**	Agent Copilot、AI Builder Copilot
Cisco	AI Assistant for Webex	**Zoom**	Zoom AI Companion
Sage	Sage Copilot	**Automation Anywhere**	Co-Pilot

资料来源：各公司官网，国信证券经济研究所整理

Copilot与Agent功能的对比

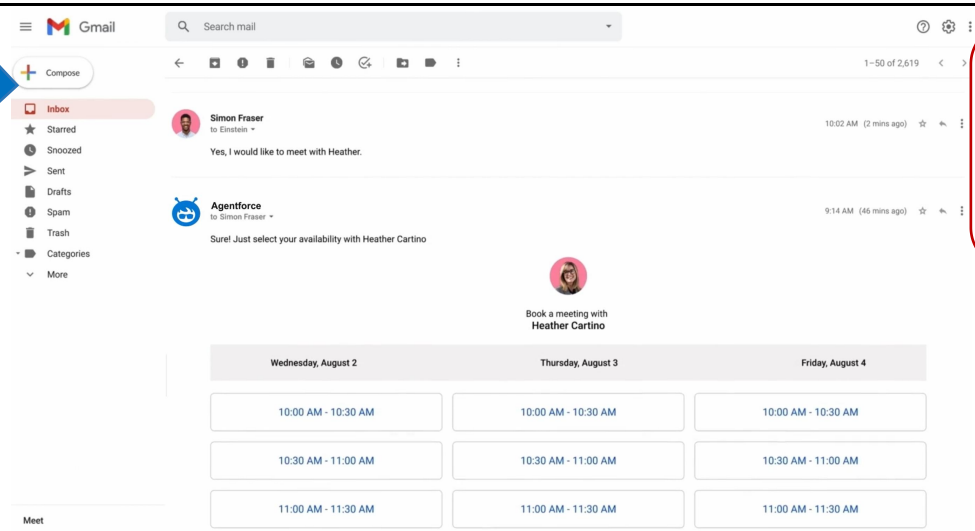
图：使用Salesforce Agent使用案例

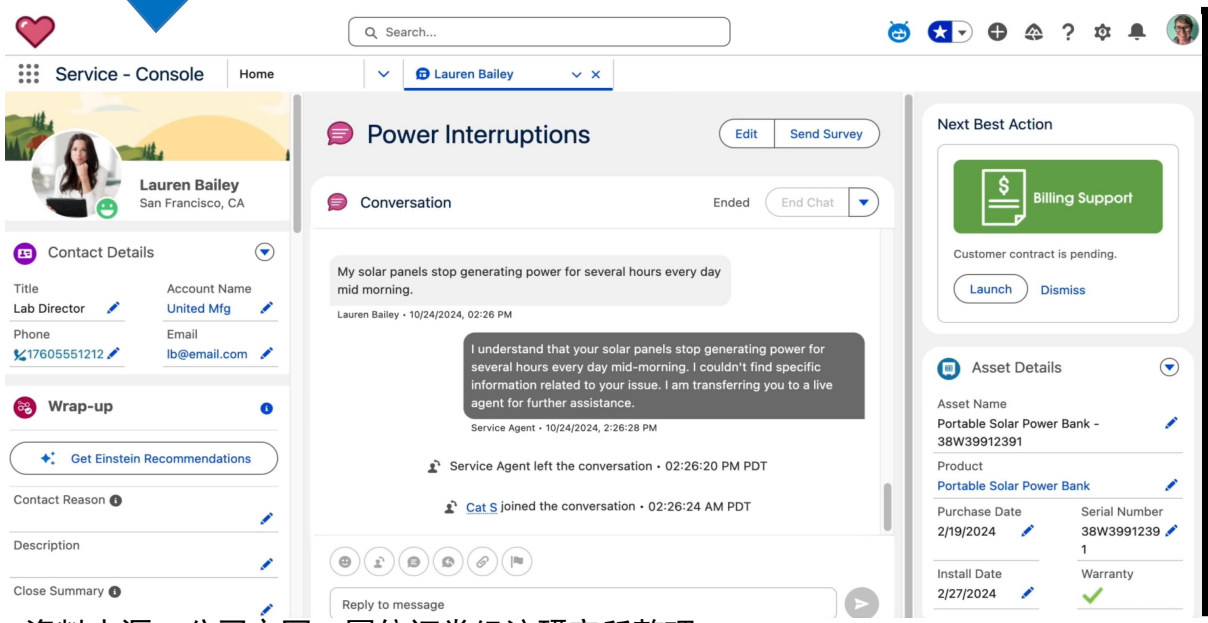
Agent作为会议管理，可以查询所有员工空闲时间，根据开会者需求预定会议并邮件通知。

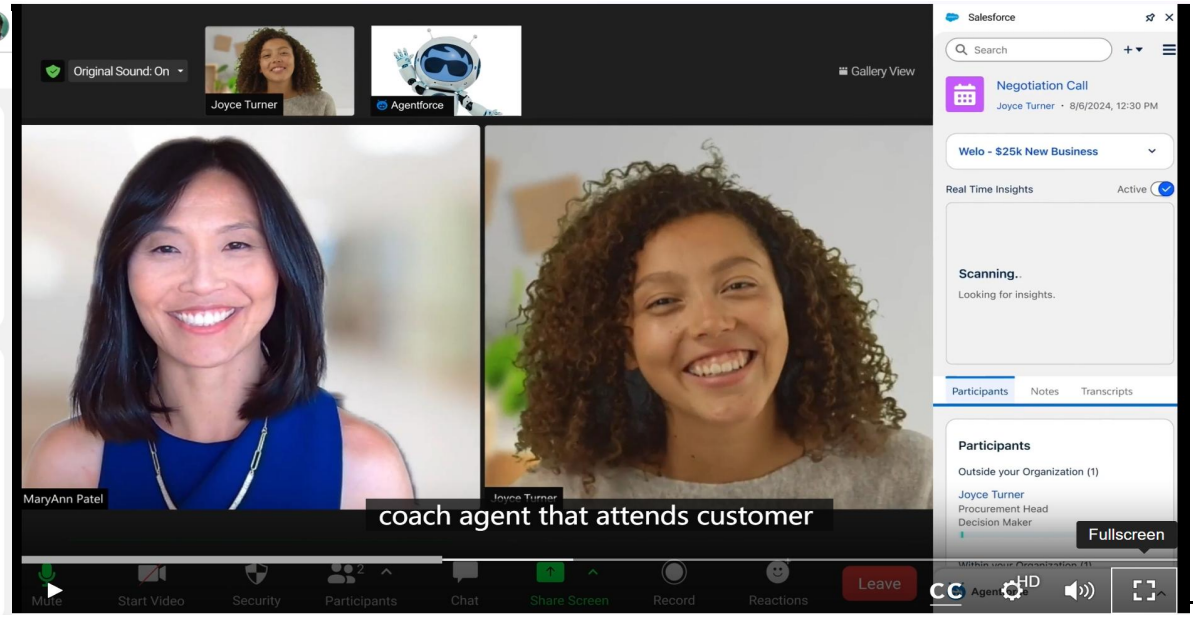
对比：GPT/Copilot为通用产品，Agent提供定制化能力（类似GPTs），可用操作端口影响现有工作流，从辅助人到代替人。

Agent作为售后客服，可以查询客户订单、解答产品问题、提供故障售后解决方案、预约线下维修时间，如不能解决可转接对口工程师。

Agent作为会议助手，可以提前通知会议、同步参会者信息，在参会过程中实时生成总结摘要。







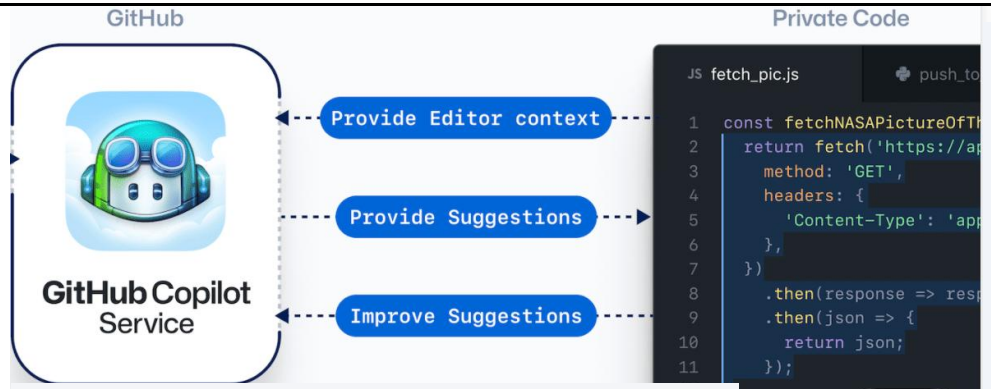
资料来源：公司官网，国信证券经济研究所整理

微软Copilot全景图

- Copilot家族月活跃用户已超1亿，若涵盖所有产品AI功能则月活跃用户超8亿。
- M365 Copilot:目前已有超1亿月活用户（商业和消费者合计），25Q2新增席位数量创发布以来新高。
- GitHub Copilot：用户规模达2000万，环比+33%。

图:微软Copilot产品一览

支持在 Word、Excel、PPT、Outlook 和 OneNote 等应用中使用Copilot，定价30美元/月。



GitHub Copilot依托OpenAICodex模型，能实现代码自动补全、代码生成/检查，费用从免费到39美元/月不等。

Copilot Pro面向个人用户的AI助手订阅服务，月费20美元。

DynamicCopilot嵌入Dynamics 365套件，在销售、营销、客户服务提供建议。

Windows Copilot在Windows10/11中推出，以侧边栏形式存在，可帮助用户控制PC设置、启动应用程序、回答查询等。

Security Copilot安全解决方案，提高防御效率和功能，支持事件响应、威胁搜寻支持。

资料来源：公司官网，国信证券经济研究所整理

Google各类Agent助手介绍

- **Google I/O 2025 大会宣布正式推出：**

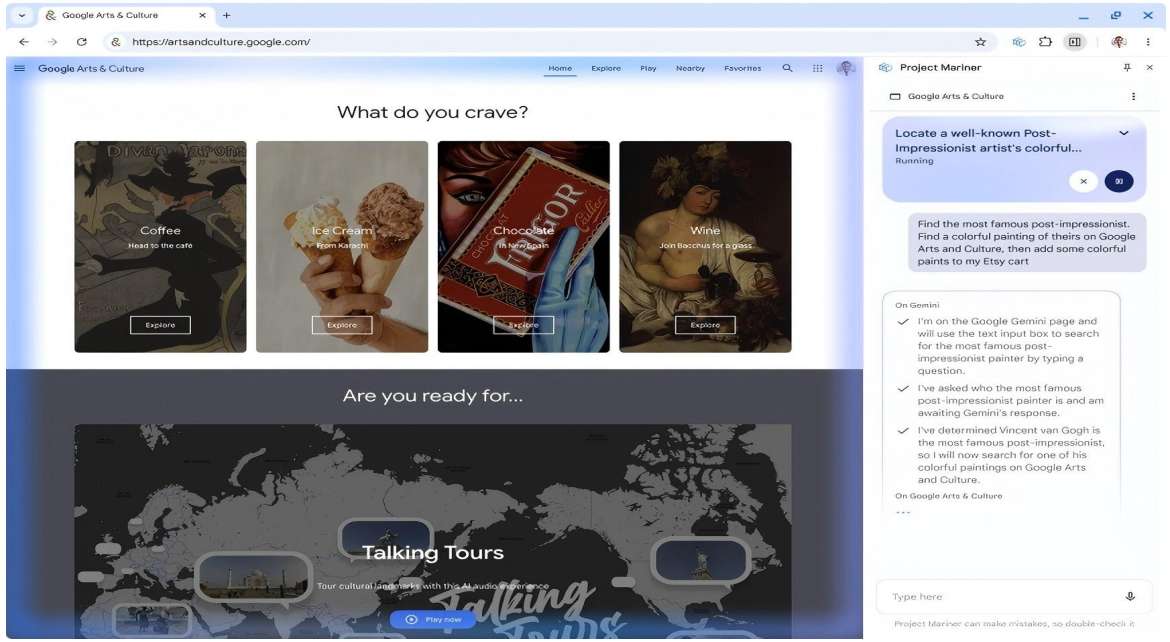
① 通用大模型助手 Project Astra：多模态AI助手，能实时理解物理环境、有长期记忆能力，可记住用户喜好信息并调取。

② 浏览器助手 Project Mariner：多任务处理能力出众，升级后的 Project Mariner 代理系统可同时处理10项任务，像房产交易时筛选信息、预约看房，购物中比价推荐 。

③ 编程助手 Jules：集成于 GitHub 工作流程，深度解析编程问题，生成解决方案，自动编写代码，助力开发者攻克复杂逻辑 。

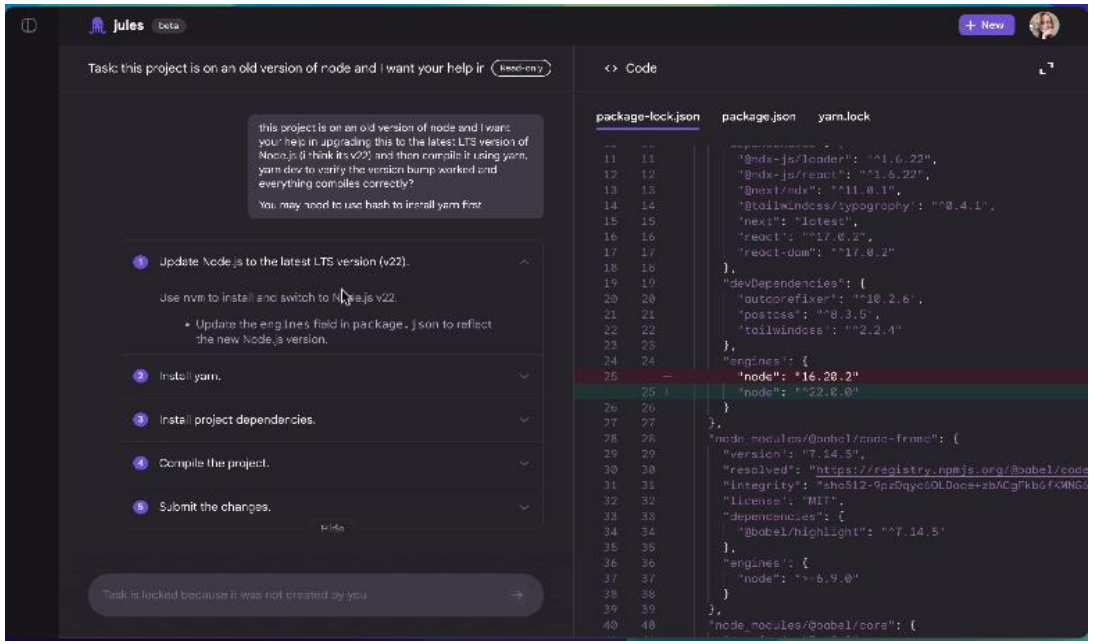
- **智能体模式将多平台上线：**将在Chrome、搜索、Gemini App 中推出，智能体可以与浏览器和其他软件进行交互和操作。

图：Project Mariner 使用示例



资料来源：公司官网，国信证券经济研究所整理

图：Jules 使用界面



资料来源：公司官网，国信证券经济研究所整理

B端GenAI渗透：开始使用的企业比例较高，但大规模落地仍有难点

国信证券 GUOSEN SECURITIES

- 根据IDC数据，企业对于使用GenAI产品积极性较高，当前77%的企业已在数字解决方案中融入AI/ML和/或GenAI，21%的企业计划在未来一年内实施，仅2%尚无计划。但企业大规模使用Agent产品仍面临一些难点：

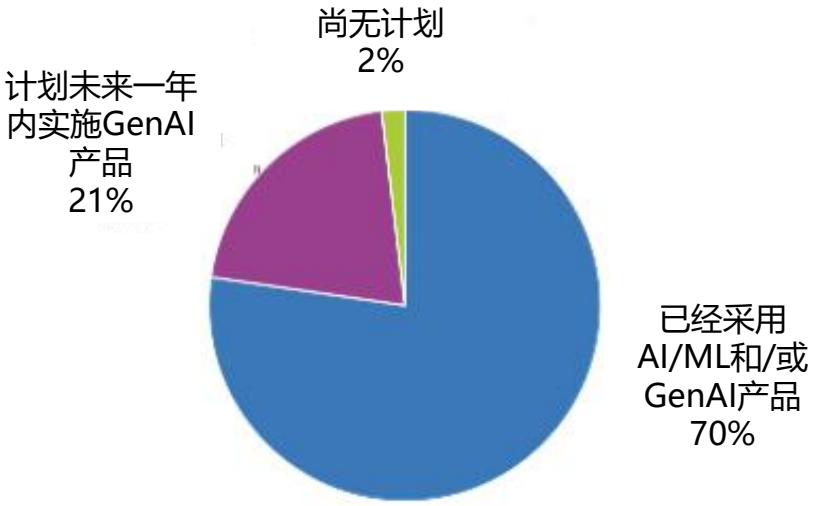
① 幻觉问题：AI输出准确性难保障，需依赖原文校验等辅助手段降低风险。

② 数据安全：企业敏感信息需严格分级隔离，防止数据泄露以及低权限员工越权访问。

③ 人才缺口：开发需大量人才，培养周期长，拉长部署时间。

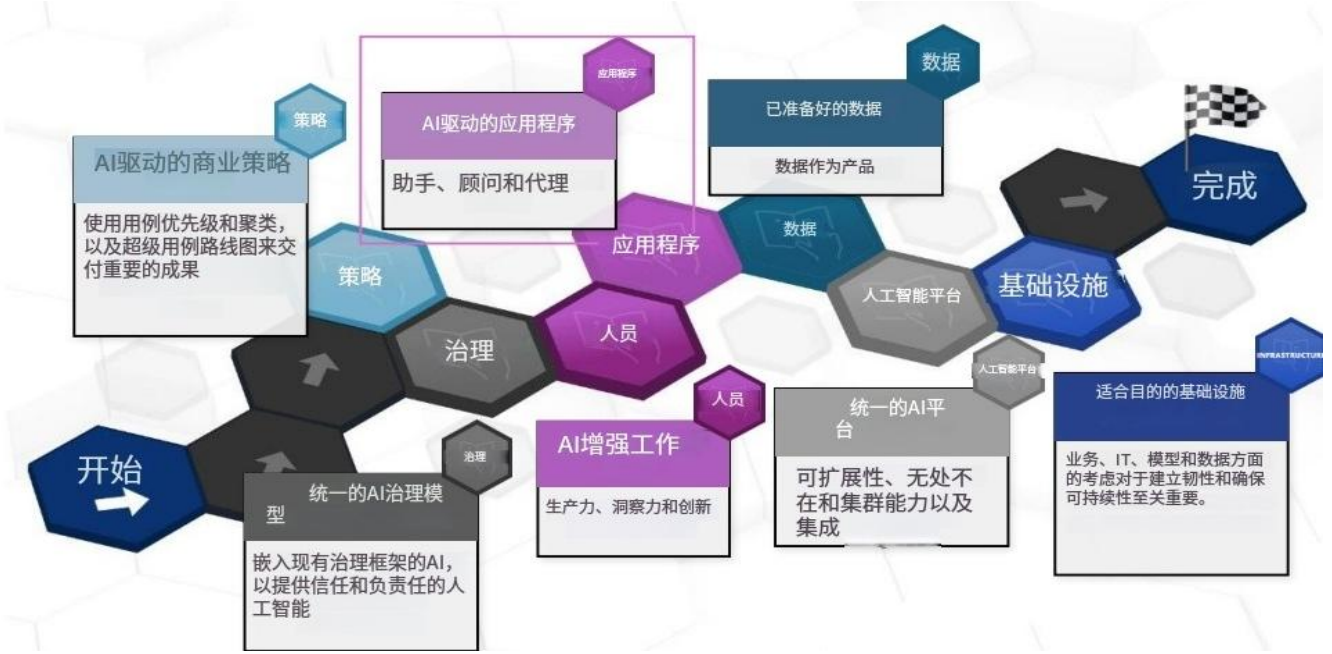
④ 成本过高：Agent调用大语言模型（LLM）的成本约为直接调用LLM 的15倍，依赖模型降价推动规模化。

图：企业采用GenAI的计划



资料来源：IDC，国信证券经济研究所整理

图：企业落地AI功能的步骤



资料来源：IDC，国信证券经济研究所整理

B端各行业的GenAI支出：显著增加，酒店/餐饮/旅游行业领先

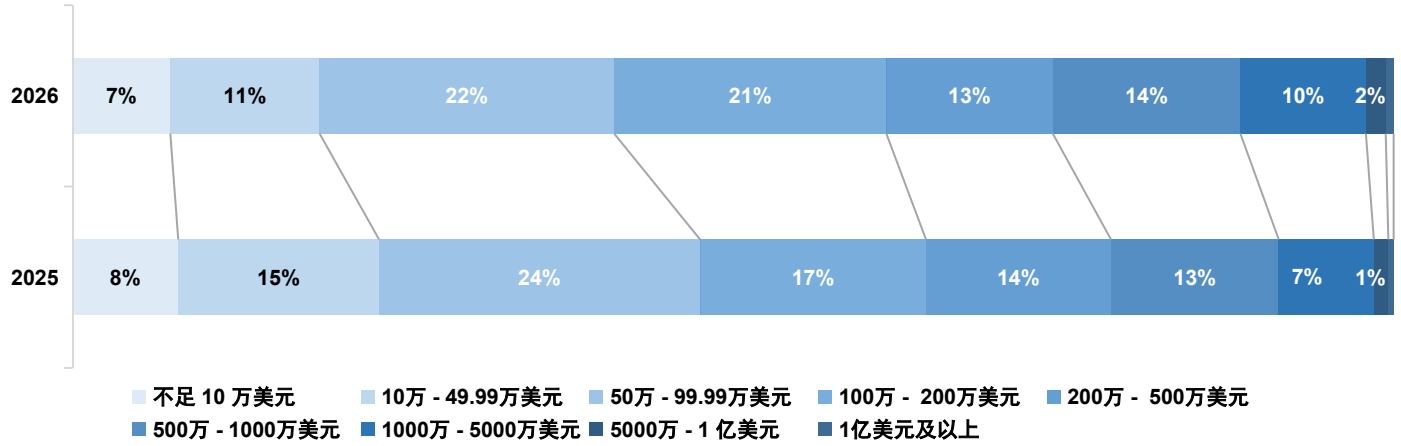
从IDC数据看出2025至2026年，企业在GenAI方面的支出将稳步增加，尤其100万美元以上投入企业比例从25年的53%提升至60%。

从行业投入来看酒店/餐饮/旅游行业、制造业以及媒零售行业在GenAI的投资处于领先地位，而智能城市、K12和国防医疗行业的投入相对滞后。

① 2025年，酒店、餐饮和旅游行业的支出遥遥领先，平均为1110万美元，其次是零售业为550万美元；

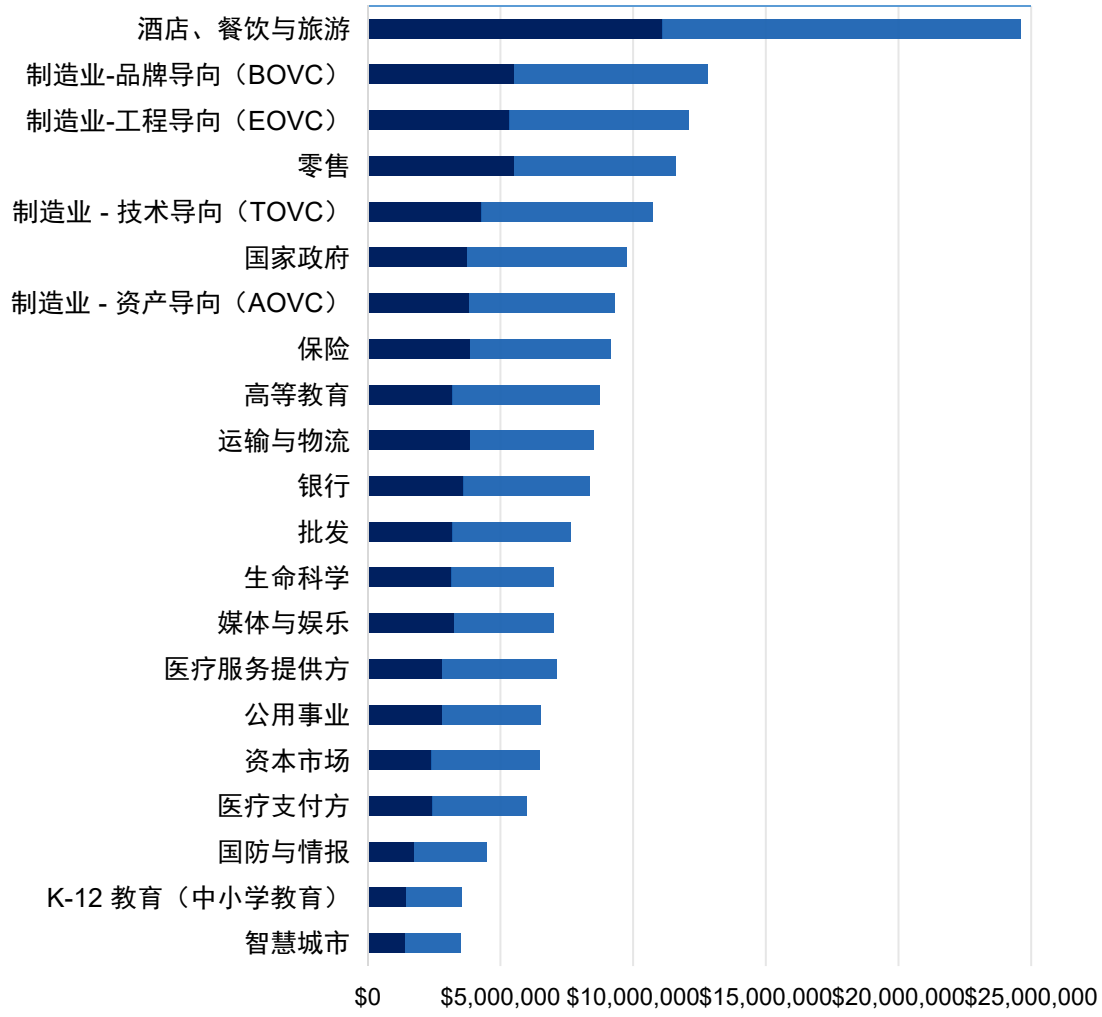
② 2026年，酒店、餐饮和旅游行业的支出预计将达到1250万美元，品牌导向制造商以740万美元位居第二。

图：2025 & 2026年组织计划在行业业务线专属GenAI工作的投入



资料来源：IDC，国信证券经济研究所整理

图：2025 & 2026年组织计划在行业业务线专属GenAI工作的投入



资料来源：IDC，国信证券经济研究所整理

B端各行业Agent案例：差异显著，赋能企业流程提效与个性化服务

国信证券 GUOSEN SECURITIES

不同行业的GenAI用例具备显著差别，从进度来看酒店/餐饮/旅游行业、健康保险公司以及媒体与娱乐行业在GenAI的全面实施上处于领先地位，而智能城市、公用事业和保险行业的实施进度相对滞后。

① **聚焦效率提升**：助力企业优化流程，如消费者反馈分析、制造业预测维护，通过智能分析和决策、降本增效。

② **强化个性化服务**：行程智能规划、内容推荐、保险个性化产品，贴合用户需求、提升服务体验与精准度。

③ **依赖数据与场景**：各行业用例基于自身业务数据和场景需求，需专业数据支撑，实现场景化落地。

图：行业主要GenAI用例

行业	主要 GenAI 用例	行业	主要 GenAI 用例
酒店、餐饮与旅游	行程智能规划，动态定价与服务优化	**公用事业**	对话式应用（如聊天机器人）、智能预测
零售	客户支持聊天机器人、选品推荐、虚拟试穿	**媒体与娱乐**	内容流媒体个性化推荐、影视剧情AIGC创作
资产导向型价值链制造业	预测性维护、故障模拟诊断、能耗智能优化	**智能城市**	安防预警、智能调度
品牌导向型价值链制造业	消费者产品反馈分析、品牌营销	**国家政府**	自动化税务合规、智能客服
工程导向型价值链制造业	预测需求与供应链优化、智能预警、方案比选	**医疗服务提供商**	医疗联络中心与虚拟助手、辅助医疗影像诊断
技术导向型价值链制造业	成本优化、研发路径规划与辅助设计	**健康保险公司**	医疗联络中心与虚拟助手、健康风险智能评估
保险	个性化保险产品定制、智能理赔审核	**生命科学**	药物安全平台、临床试验设计
银行	支付数据分析、定制理财组合、反欺诈	**K-12 教育**	自动化评分与反馈、定制学习计划
资本市场	客户细分与精准营销、标的筛选、市场情绪分析	**高等教育**	AI 研究与写作助手、智能查重、数据分析
运输与物流	资产与司机状态报告生成、路径优化	**国防与情报**	财务监督与分析、情报智能分析研判
批发	服务响应优化、智能预测与供应商评估		

资料来源：IDC，cbinsights，国信证券经济研究所整理

各个SaaS公司传统业务与AI发展阶段

表：各个SaaS公司传统业务与AI发展阶段

	公司	传统业务阶段	25年收入增速预期	24Q4 AI进展与赋能
AI驱动公司整体收入快速增长	Palantir	商业端与政府端软件采购需求均加速	45%+	AIP 推动客户数同比+43%，25Q2完成总合同价值22.7亿美元，同比+140%
	Duolingo	语言学习MAU与付费率稳健提升	35%+	DuolingoMax订阅占总订阅的7%（Q1为5%），预计中国未来会上线。ARPU同比增长约6%，毛利率超预期因AI成本下降（tokens成本降低）。
	Applovin	游戏广告持续强劲，拓展电商广告第二曲线	30%+	AXON 2.0凭借公司全链路数据积累显著提升营销效率，10月1日面向国际市场开放AXON，并在26H1全球公开发布
	Shopify	商家GMV强劲叠加PLUS计划保持高增速	25%+	推出Shopify Catalog授权AI伙伴访问shopify数百万产品，推出Checkout Kit结账的Copilot工具，以及Sidekick数据分析工具
	Roblox	主要受《Grow a Garden》等内容驱动	20%+	1）AI升级推荐算法，让更多新优质内容获得曝光，25Q1Top100游戏有24款是过去一年内推出的；2）3D基础模型Cube 3D已投入使用，生成了超过100万个3D模型。
	Microsoft	AI拉动云持续加速与SaaS ARPU提升	18%+	25Q2 Azure同比+39%，Azure年收入超过750亿美元，AI 驱动作用显著。
AI产品收入增长较快，传统业务与AI影响需要观察	Snowflake	本地数据迁移到云端趋势持续，传统工作负载需求亦有所提升	20%+	2024年Snowpark贡献3%的产品收入，面临非结构化数据存储需求的增加与竞争格局恶化、技术变迁同时作用的影响
	Salesforce	收入降速，指引今年高个位数增长。	8%+	数据云和AI ARR达到9亿美元，同比+120%。Agentforce 上线仅90天后，已经有3,000 名付费客户，预计26 财年对收入的贡献不大。
	ServiceNow	ITSM传统需求稳健，格局较好	20%+	ProPlus交易数量环比增长超50%，计划在2026年实现150亿美元订阅收入及10亿美元NowAssist ACV目标。
	Gitlab	客户订阅需求企稳，OPM提升显著	25%+	26FYQ1 AI相关收入占比达15%，2026财年目标提升至25%。GitLab Duo 的首购客户数环比+35%。7月，AI协作平台GitLab Duo正式开启公测。
	Crowdstrike	安全需求稳健，OPM将逐年提高	20%+	24FYQ4云安全业务增长超45%，年末ARR超6亿美元；客户多模块采用率同比提升五个点以上。
AI冲击传统业务	Adobe	三年增速持平约10%，AI贡献被原核心业务新增ARR下滑所抵消	9%+	AI独立产品FY25Q1 ARR达到1.25亿美元，本财年末有望翻倍
	C3.A1	企业压缩IT预算并对解决方案AI需求提升，传统订阅收入增长承压	25%	AI收入构成核心驱动，AgenticAI业务年化ARR约6000万美元、生成式AI收入同比增长超100%

资料来源：各公司财报会，国信证券经济研究所整理

美股SaaS AI产品进展阶段

图：各个SaaS公司传统业务与AI发展阶段

资料来源：各公司财报会，国信证券经济研究所整理

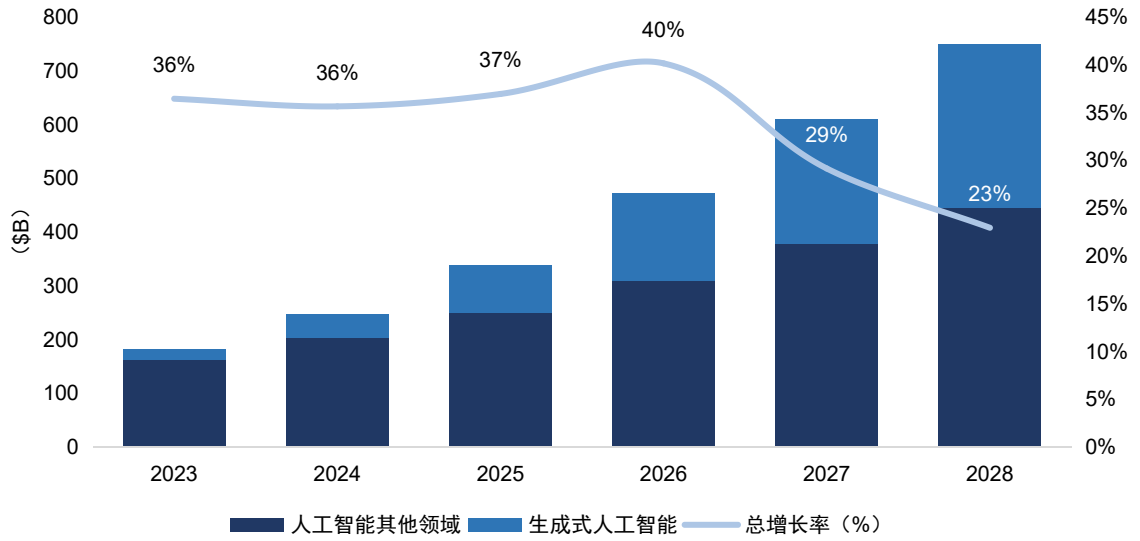
目录

全球人工智能IT支出预测：五年复合年增长率（CAGR）为22.3%

IDC发布的2024-2028年全球人工智能IT支出预测显示，**2023-2028年全球人工智能IT支出呈现持续增长态势，总市场的复合年增长率（CAGR）为22.3%。**过去两年AI已从“概念”进入“真金白银投入”，且呈现爆发式增长。其中GenAI的CAGR为73.5%（GenAI信息创建，如文本、音频、视频、图像和代码生成相关），其余AI的CAGR为22.3%（Rest of AI，主要为解释性AI以及预测性/规范性AI）。

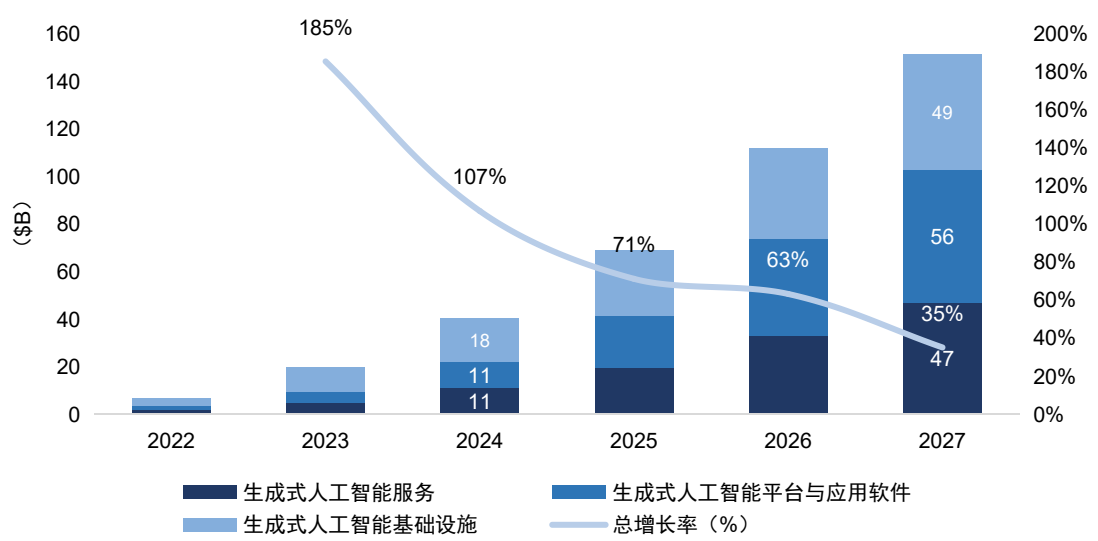
到2028年，AI技术支出将占总IT支出的16.4%，其中生成式AI（GenAI）自身将占支出的6.7%。在GenAI领域 “基建、服务、平台/应用” 分工和投入逐渐清晰，从2024年到2028 年将增加三倍。各个组织正转向更具战略性的人工智能方法，整合应用程序、平台、数据和基础设施方面的投资，旨在通过先进自动化、更高效的模型和数据复用率以及高效的推理来提升AI价值。

图：全球人工智能支出概况



资料来源：IDC，国信证券经济研究所整理

图：Gen AI的全球核心IT支出概况

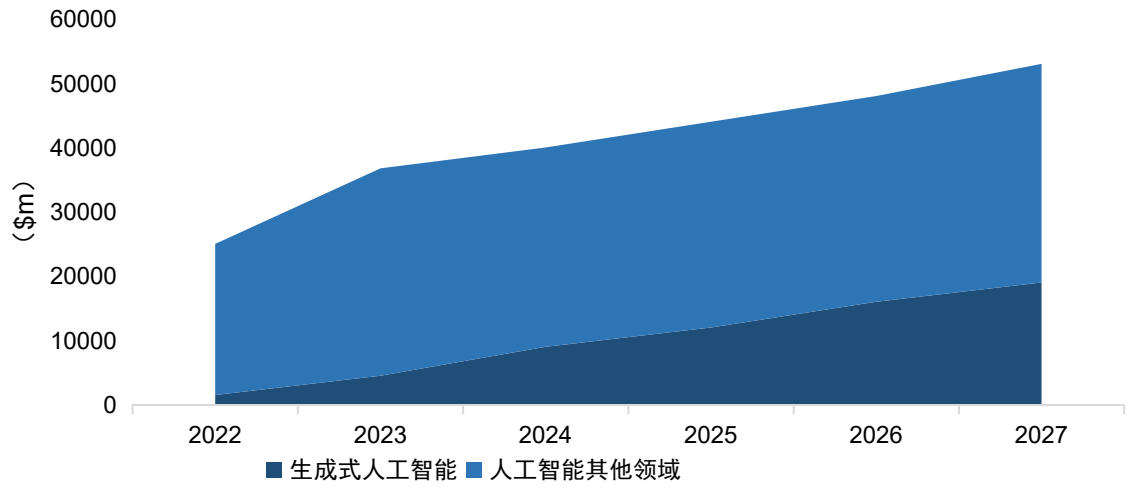


资料来源：IDC，国信证券经济研究所整理

GenAI支出分配：软件成为最大支出领域，其次是AI硬件

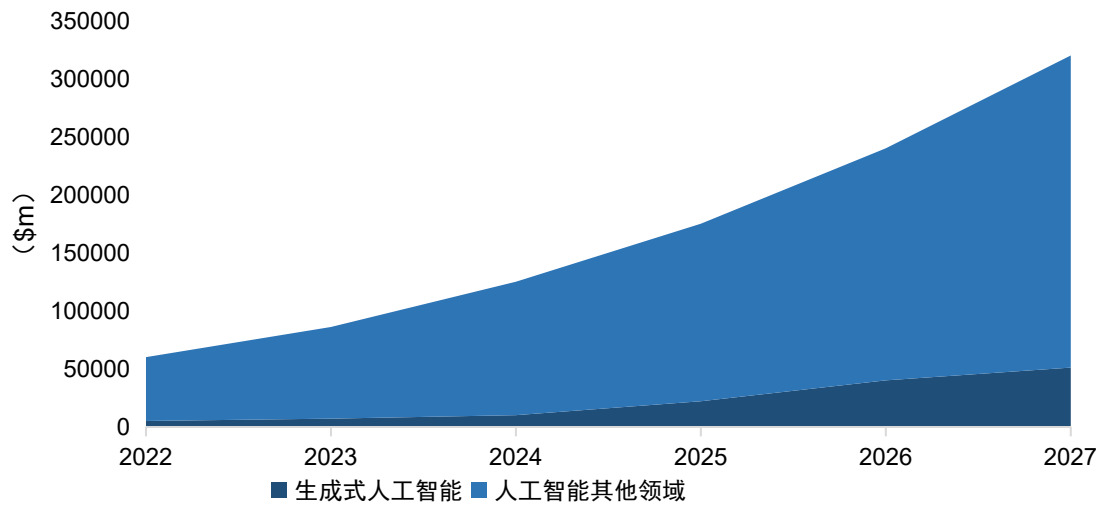
- **IDC预测软件将成为最大的技术支出类别，在大部分预测中占整个AI市场的一半以上，AI软件的五年复合年增长率为33.9%。**生成式AI的热潮正推动部分新的 IT 支出和预算重新分配，其中有部分支出被重新分配到平台即服务（PaaS），以积极支持包括 AI 开发在内的云平台。所有软件支出的三分之二将用于AI支持的应用程序和人工智能平台，而其余的将用于AI应用开发和部署以及AI系统基础设施软件。
- **AI硬件（包括服务器、存储和基础设施即服务IaaS）上的支出将成为第二大技术支出类别。**其中近几年服务器增长快于预期，AI工作负载部署从2023年下半年推动服务器价值上升。另外供应链改善也推动了网络设备支出的增长，但传统PC与手机等设备增长相对疲软。
- **IT服务的增长率将略快于硬件，五年复合年增长率为24.3%。**为了实现业务流程优化、提升客户体验、增强数据驱动决策能力等目标，企业会采购大量与数字化转型相关的IT服务，包括IT 服务团队进行数据标注、算法优化等工作，以及在模型部署和运维阶段IT服务来保障系统的稳定运行和性能优化。

图：服务器/存储支出



资料来源：IDC，国信证券经济研究所整理

图：软件支出



资料来源：IDC，国信证券经济研究所整理

AI Agent市场：近两年赛道公司营收与融资飙升，2032年超千亿

- **市场规模：**CBINSIGHTS统计预测2024年营收超50亿美元，预计2025年底突破100亿美元，2032年达到1036亿美元，2024-2032年CAGR 44.9%。超半数AI Agent公司2023年成立，2024年赛道融资近乎翻3倍；
- **Plivo预测2025四大趋势：**

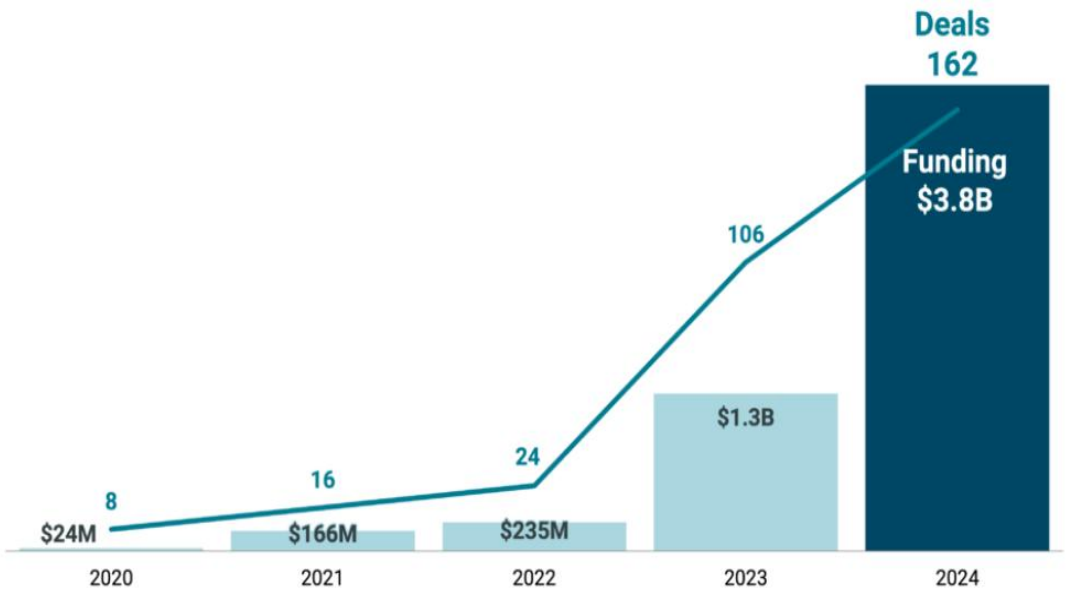
① 巨头垄断通用场景：科技巨头借海量用户（如OpenAI周活4亿）、企业信任，推动通用Agent更优、更普及，挤压初创。

② 初创深耕垂直细分：横向市场占半壁股权交易，企业借客服、编码等场景深度数据集成突围，行业专属方案加速落地。

③ 基建体系结构化：开发流程走向清晰，数据治理、全栈平台等细分赛道成型，全栈平台成最大基建品类。

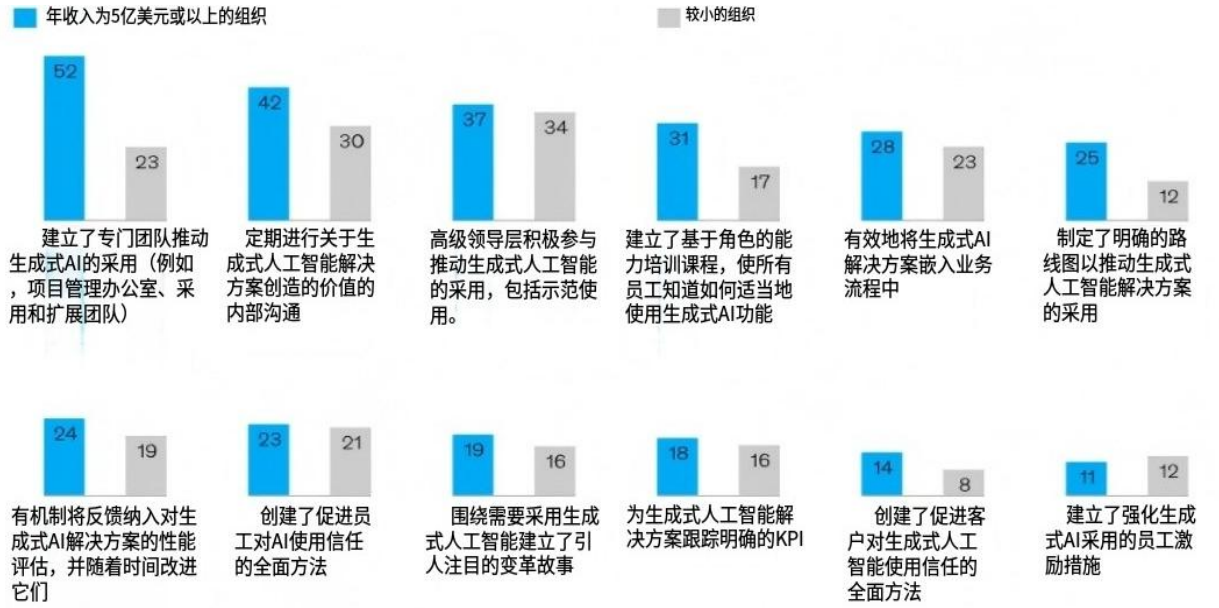
④ 企业从“试”到“用”：63%企业视Agent为未来12个月关键布局，可靠性、人才等挑战仍待解决，人机协同、数据基建可破局。

图：AI Agent初创公司的股权交易和融资



资料来源：CBINSIGHTS据(截至2025年2月27日)，国信证券经济研究所整理

图：大型组织更多地参与 GenAI落地实践



资料来源：Plivo，国信证券经济研究所整理

AI Agent的技术演进与发展：终极形态是认知共生的人类助手

国信证券 GUOSEN SECURITIES

- **发展阶段时间线：**

表：Agent应用架构演进

阶段	2025（短期）	2027（中期）	2030+（长期）
特征	助手增强(Assistant-Enhanced)	代理主导(Agent-Led)	代理及应用 (Agent as apps)
应用栈	聚焦SaaS平台功能优化，发展无代码与低代码能力	AI Agent 的感知与交互能力显著升级，智能体协作系统兴起	Agent 自主决策能力突破，能完成复杂任务，能与真实世界交互
商业模式	按用户收费（API/计算使用量）	新增基于成果计费	基于 API 或计算的使用量计费；按用户 / 员工基于成果计费
竞争格局	AI原生产品大量出现	传统SaaS与AI厂商迭代、并购、整合	少数平台主导代理生态，部分SaaS现有厂商并存

资料来源：Gartner、华尔街新闻，国信证券经济研究所整理

AI Agent渗透率/采用率变化预测

商业价值驱动下，Agent的企业采用率将大幅提升。Capgemini预测，2025年25%使用生成式AI的企业将部署Agent，82%的组织计划在2026年前集成 Agent，用于邮件生成、编码、数据分析等任务。

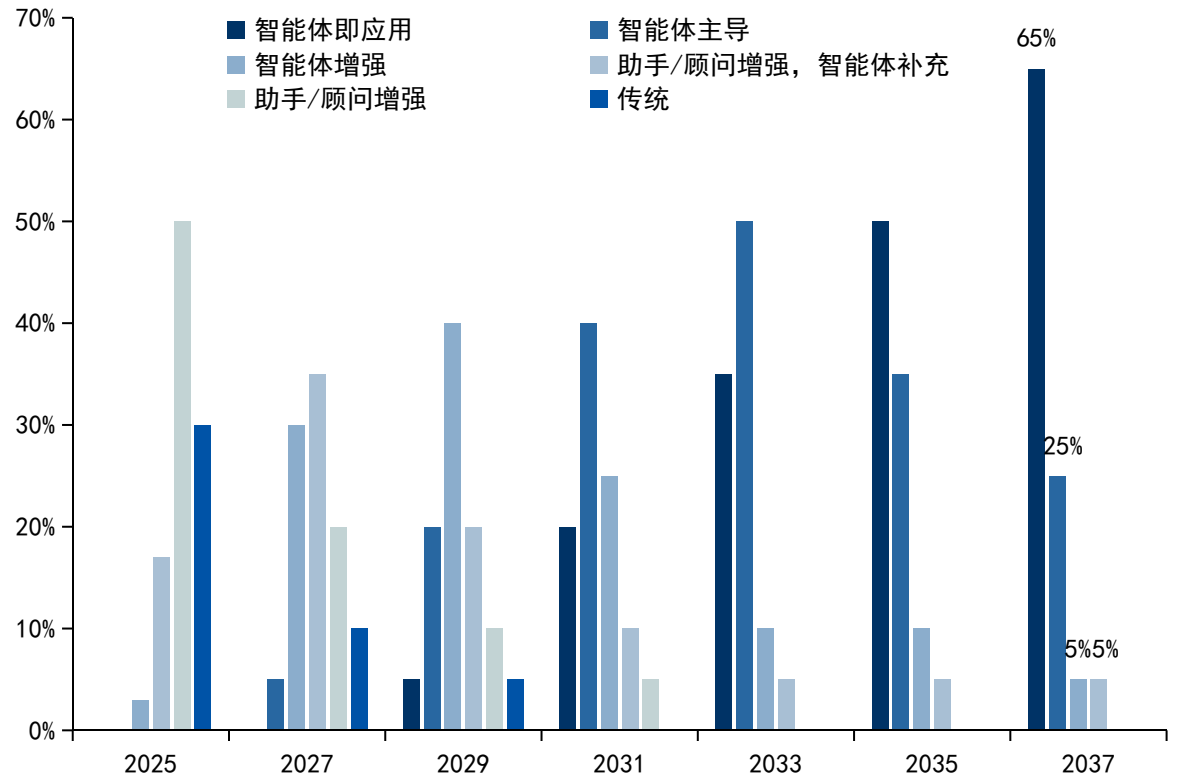
未来10年，企业应用智能体化演进将从功能增强逐步迈向由智能体主导，2037年“智能体即应用”将成为主流。

表：企业应用智能体演化

演进阶段	核心特征
智能体即应用	• 智能体取代整个应用（如公司将会拥有CRM智能体/智能体群等）； • 传统用户界面极少被使用。
智能体主导	• 智能体取代应用内的整个功能模块（如供应链管理系统中，一个智能体负责库存管理，另一个负责物流等）； • 智能体交互界面（文本提示和语音）占主导地位，传统界面很少使用。
智能体增强	• 智能体驱动功能显著增强应用能力，提升生产效率； • 智能体交互界面日益占据主导地位，对传统界面的依赖逐渐减弱；
助手/顾问增强，智能体补充	• 大部分功能由助手(As)/顾问(Ad)增强，嵌入式智能体开始增长； • 传统界面占大多数，部分被智能体交互界面取代。
助手/顾问增强	• 助手/顾问增强的功能逐渐增多； • 传统用户界面保持不变。
传统	• 传统的功能与用户界面。

资料来源：IDC，国信证券经济研究所整理

图：企业应用智能体演化各阶段市占率



资料来源：IDC，国信证券经济研究所整理

风险提示

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。

免责声明

国信证券 GUOSEN SECURITIES

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