上海证券
SHANGHAI SECURITIES

证券研究报告
2025年2月19日
行业：计算机
增持（维持）

# 关注信创国产化和AI商业化两大主线

## ——计算机行业2025年度投资策略

分析师：吴婷婷　　SAC编号：S0870523080001

## 主要观点

**政策与生态共振，自主可控全面加速。**国资委79号文要求2027年底前实现所有中央企业的信息化系统安可信创替代。万亿国债重点聚焦高水平科技自立自强，有望支撑信创产业发展。据艾媒咨询测算，2023年中国信创产业规模将达20961.9亿元，2027年有望达到37011.3亿元。外部不确定性增强，自主可控大势所趋，国产算力迎发展新机遇，华为原生鸿蒙开启国产OS新篇章，信创国产替代加速。节奏上，党政信创持续引领，行业信创新政频出，“2+8+N”战略正在逐步推进，新一轮信创招标或已开启。

**AI商业化拐点已至，Agent助力大模型落地。**B端，微软、谷歌、Salesforce、百度、腾讯、字节等持续加码，推动AI Agent商业化落地。C端，国内外Agent惊艳涌现，OpenAI发布AI代理Operator、Deep Research，智谱发布AutoGLM和GLM-PC两大系统，覆盖移动设备和桌面端，Anthropic发布Computer use功能。根据Markets and Markets预测，全球AI Agent市场将从2024年的51亿美元增长到2030年的471亿美元，年复合增长率达44.8%。根据头豹研究院，2023年中国AI Agent市场规模为554亿元，预计2028年中国AI Agent市场规模将达到8520亿元，2023-2028年均复合增长率达72.7%。

**推理算力需求爆发，国产算力迎新机遇。**豆包应用全球火热，截止2025年1月底，豆包日活已达千万级，凭借庞大的用户基础和高活跃度，成为行业发展的引领者。同时，豆包实时语音大模型、豆包大模型1.5 Pro持续发布，有望推动推理算力需求持续提升。DeepSeek火热出圈，通过技术创新有效实现大模型训练成本降低，大约为Meta 的 1/10，OpenAI 的 1/20，并通过开源协议和蒸馏等工程方法为业界带来低成本的端侧模型商品。我们认为，DeepSeek的成本创新，有望降低技术门槛，带来AI平权，加速AI应用的爆发；随着强化学习成为后训练阶段的标配，推理计算占比将逐步提升，ASIC 以及国产芯片有望逐步抢占英伟达GPU的份额，迎来发展机遇。

**风险提示：**AI应用落地不及预期；AI需求不及预期；行业竞争加剧。

# 目录

Content

# 一、信创产业已进入“真替真用”阶段

◆ 信创经历了预研、发展、试点、应用各发展阶段，2020年经过多轮试点后进入规模化推广阶段，在党政信创的引领下，2022-2023年逐渐向国计民生行业加速渗透，中石化、中交集团、中国稀土、中储粮等央国企及行业信创大单频现，**信创产业产品采购步入常态化阶段，信创产业“真替真用”开启。**

**图1 信创产业发展历程**


*资料来源：亿欧智库，上海证券研究所*

# 一、政策高度关注，加速信创国产化进程



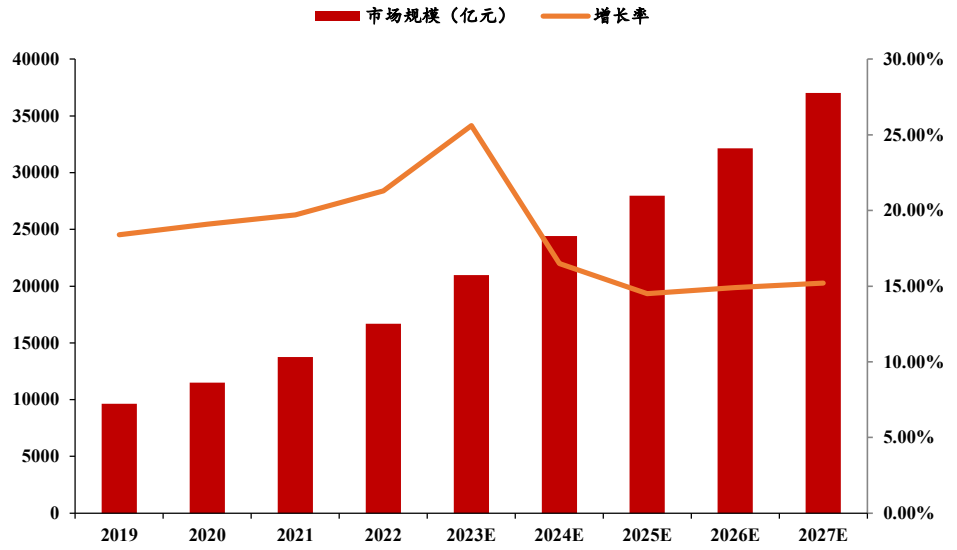
◆ **国家高度重视信创产业。**“十四五”规划中明确指出，到2025年行政办公及电子政务系统要全部完成国产化替代。2022年9月底，国资委发布79号文，要求2027年底前实现所有中央企业的信息化系统安可信创替代。党的二十大报告再定增强国家安全主基调，重申发展信创产业，实现关键领域信息技术自主可控的重要性。2024年8月6日，国务院国资委印发了《关于规范中央企业采购管理工作的指导意见》，文件提出，在卫星导航、芯片、高端数控机床、工业机器人、先进医疗设备等科技创新重点领域，充分发挥中央企业采购使用的主力军作用，带头使用创新产品。

◆ **各部委信创政策陆续发布，支撑信创产业市场化规范化发展。**2023年7月，安全可靠测评工作指南发布，截止24年9月，已公布安全可靠测评结果公告2023年第1号、2024年第1号及2024年第2号，为用户产品选型提供参考依据。2023年底，财政部会同工信部发布7项基础软硬件政府采购标准；2024年，央采网公布中央国家机关台式机、便携式计算机采购标准；2024年9月，工信部发文要求，到2027年，完成约200万套工业软件和80万台套工业操作系统更新换代任务。我们认为，随着信创行业采购标准的推出，信创产业有望向市场化、规范化方向发展。

# 一、万亿国债支撑信创产业景气度提升

◆ 2024年政府工作报告提出，从2024年开始拟连续几年发行超长期特别国债，专项用于国家重大战略实施和重点领域安全能力建设。5月13日，财政部发布1万亿超长期特别国债发行安排。我们认为，超长期特别国债重点聚焦加快实现高水平科技自立自强，有望支撑信创产业发展。据艾媒咨询测算，2023年中国信创产业规模将达20961.9亿元，2027年有望达到37011.3亿元。

图2 2019-2027年中国信创产业规模及预测




*资料来源：艾媒咨询，上海证券研究所*

# 一、中美科技竞争升级，自主可控大势所趋

上海证券 SHANGHAI SECURITIES

◆ **美国芯片限制加码，外部环境变化加速自主可控。**2024年12月2日，美国商务部工业与安全局（BIS）公布了对华半导体出口管制措施新规，包括对24种半导体制造设备和3种用于开发或生产半导体的软件工具的新控制，以及新增140个实体列表和14个实体列表，涵盖半导体相关工具制造商、半导体晶圆厂和投资公司。2025年1月13日，美国拜登政府正式公布针对AI的临时最终出口管制规则，进一步限制中国大陆等国家和地区获得美国AI技术的能力。2025年1月15日，美国发布新规，对华限制16nm及以下先进制程代工服务。此次美国出台的出口管制政策将主要限制台积电、三星等海外晶圆代工厂为中国大陆客户提供16nm及以下先进制程芯片代工服务。我们认为，美国对AI芯片的限制不断加码，外部环境的变化有望加速自主可控，国产算力产业迎来新机遇。

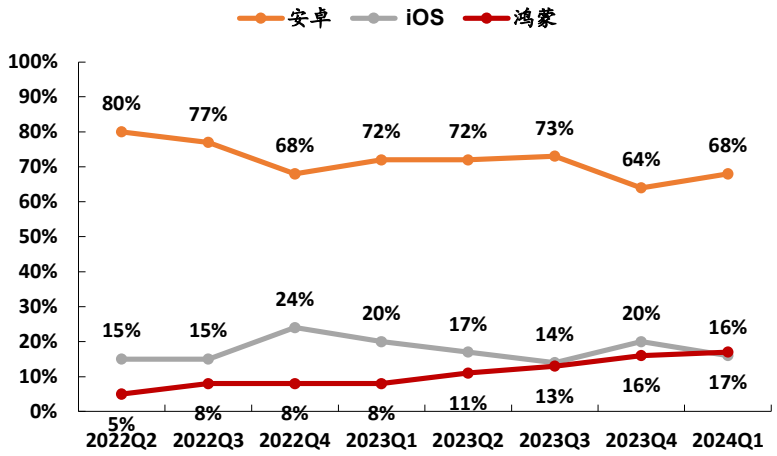
◆ **2024年以来，微软蓝屏事件、黎巴嫩寻呼机爆炸事件、美国科技制裁和限制加码凸显自主可控重要性。**1月20日，特朗普宣誓就职，其上一任期对中国科技企业采取了强硬态度，发起了一系列“贸易战”和科技制裁。我们认为，随着外部不确定性增强，自主可控紧迫性提升，信创产业将迎来发展机遇。

# 一、大国博弈背景下，信创产业内涵延伸



- 近年来，我国信创产业产品生态体系已初步成型，未来发力重点在于构建国产化信息技术软硬件底层架构体系和全生命周期生态体系，解决关键环节“卡脖子”问题。近期，我国对谷歌公司展开反垄断调查，表明信创产业的国家战略高度，国产替代需求将不断提升。
- **华为原生鸿蒙发布，开启国产OS新篇章。**2024年10月22日，华为发布全新的HarmonyOS NEXT，实现从内到外的全栈自研，标志着华为自研自主可控操作系统的成熟。据Counterpoint Research，HarmonyOS在中国的市场份额已由2023年一季度的8%上涨至2024年一季度的17%，超越iOS，成为中国第二大操作系统。据Counterpoint，2024年第四季度，华为以18.1%的市场份额跃居中国智能手机市场榜首，2024年全年的市场份额为16.3%，位居第二。据StatCounter，截至2024年11月，全球智能手机市场中苹果和安卓的市场份额分别为27.93%和71.42%，鸿蒙仍有很大的提升空间。截至2025年1月，已有超20000+鸿蒙原生应用及元服务上架鸿蒙原生应用市场。

**图3 中国智能手机市场安卓、苹果、鸿蒙操作系统市场份额**




*资料来源：Counterpoint Research，上海证券研究所*

# 一、“2+8+N”逐步深入，全面自主可控可期



- **党政信创持续深化：从省市一级纵向下沉至区县乡镇，从电子公文系统横向拓展至电子政务系统。**从2020年党政信创三期开始招标以来，党政信创不断深化。2022年，根据“十四五”数字经济发展规划中对电子政务的国产化替代目标，党政信创的重心逐渐从电子公文偏移到电子政务领域。2023年底，随着安全可靠测评结果公布、财政部会同工信部发布7项基础软硬件政府采购标准，信创产品在政府部门进一步深入。2024年，中央国家机关政府采购中心公布中央国家机关台式机、便携式计算机采购标准，要求乡镇以上党政机关，以及乡镇以上党委和政府直属事业单位及部门所属为机关提供支持保障的事业单位在采购台式计算机、便携式计算机时，应当将CPU、操作系统符合安全可靠测评要求纳入采购需求，各省市自治区也接连发布相关通知，党政信创往区县乡不断下沉，并从基础软硬件替换扩展到电子公文系统以及电子政务系统国产化。
- **党政信创存量替换空间巨大。**据自主可控新事调研测算，党政信创存量规模近3000万台，截至2024年12月，党政信创PC历史总出货量近700万台，占总量比重仅约20%，2025年或将成为政企信创终端替换大年，约有1000万台的替换空间，合计市场规模约在500亿元左右。

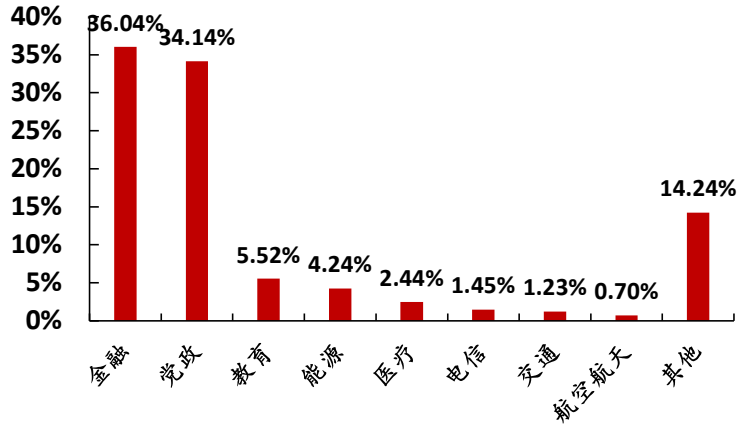
# 一、“2+8+N”逐步深入，全面自主可控可期

◆ **行业信创新政频出，“2+8+N”逐步推进。**在党政信创持续引领下，以金融、电信、电力领域为代表的行业信创正加速推进，石油、交通、航空航天在重点环节推广，医疗、教育信创趋热，信创应用从党政领域向行业领域转化。2024年9月20日，工信部发布《工业重点行业领域设备更新和技术改造指南》，针对工业软件领域、工业网络设备给出具体的更新目标。2024年12月11日，教育部和国家版权局联合印发《关于做好教育系统软件正版化工作的通知》，要求2027年底前，教育系统软件正版率显著提升，全面使用正版操作系统软件、办公软件和杀毒软件。我们认为，随着行业信创政策不断出台，行业信创有望接力党政信创，打开信创应用新格局。

图4 “2+8+N”行业信创建设节奏


资料来源：亿欧智库，上海证券研究所

图5 2024年信创各行业中标金额占比




资料来源：亿欧智库，上海证券研究所

# 一、信创订单密集落地，新一轮招标或已启动

◆ 2023年下半年开始，关基行业的信创招投标陆续启动，中信银行、工行、邮储、浦发等银行信创招投标持续落地，运营商服务器设备招标也相继落地。据亿欧智库，2024年信创整体招投标超过48亿元，“2+8”行业占比86%，是信创推动的主力军。近期，中国移动、吉林省政务服务中心等陆续披露信创相关项目，我们认为，或是新一轮信创招标开启的信号。

**表1 近期信创招标密集**

| 时间 | 招标主体 | 招标内容 |
|---|---|---|
| 2024.11 | 中国移动 | X86 服务器 1724 台、国产 C86 服务器 290 台、国产 ARM 服务器 1176 台、交换机 260 台等，总投资2.29亿元 |
| 2024.11 | 四川农商行 | 海光、鲲鹏、国产芯片服务器合计1044台，总投资1.18亿元 |
| 2024.11 | 山东高速 | 信息化建设项目，总投资4720万元 |
| 2024.12 | 中共河北省委 | 省政务云服务兼容华为云技术架构等，总投资2.6亿元 |
| 2024.12 | 吉林省政务服务和数字化建设服务中心 | 政务信息化统一建设项目，总投资2.39亿元 |
| 2025.01 | 湖南省政务服务和大数据中心 | 省级政务云服务，总投资5亿元 |

*资料来源：信创焦点，上海证券研究所*

# 目录

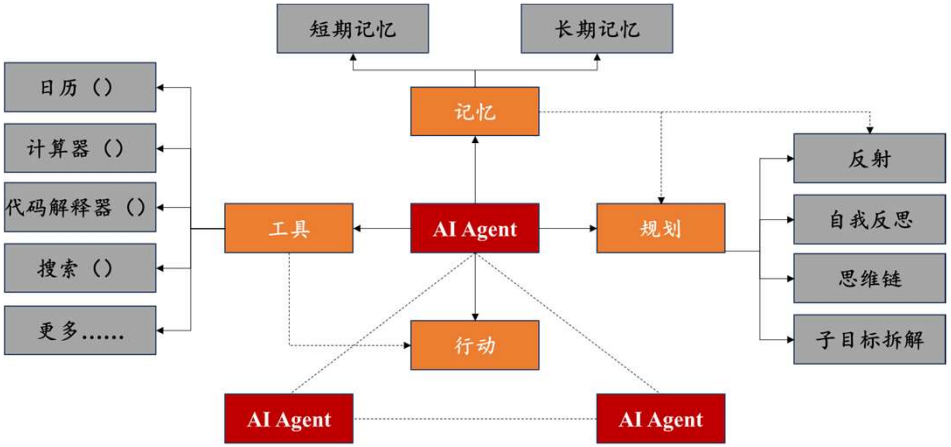
Content

# 二、AI Agent自主性、交互性凸显，打破大模型应用边界

◆ **AI Agent（人工智能代理），指能够感知环境、进行自主理解、决策和执行动作的智能体。**OpenAI将AI Agent定义为“以大语言模型为大脑驱动，具备自主理解、感知、规划、记忆和使用工具的能力，可自动化执行完成复杂任务的系统。”

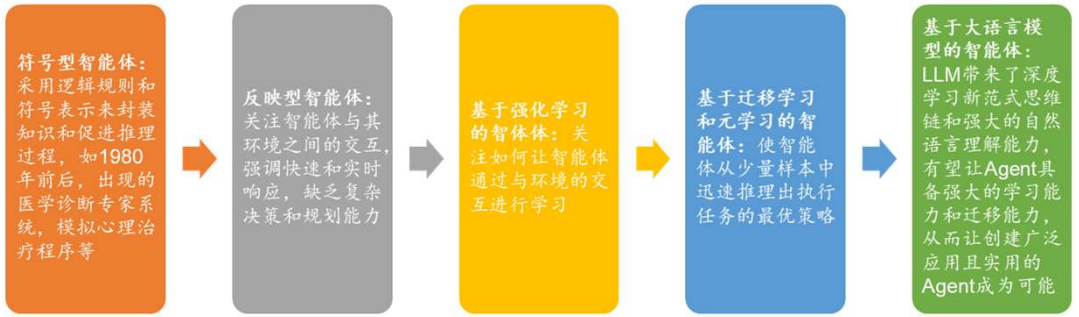
◆ **AI Agent打破大模型应用边界。**AI Agent已跨入基于大型语言模型的智能体阶段，具备自主性、适应性、交互性、智能性四大特点。 LLM是AI Agent实现的基础和前提，LLM带来了深度学习新范式，思维链和强大的自然语言理解能力有望让Agent具备强大的学习能力和迁移能力，从而让创建广泛应用且实用的Agent成为可能。而AI Agent的引入，则能够赋予大模型多轮对话管理、主动询问与澄清、策略性决策的能力，我们认为，AI Agent有望增强大模型的深入思考，突破语言模型的边界。

**图6 基于LLM驱动的Agent基本框架**




*资料来源：《LLM Powered Autonomous Agents》，腾讯研究院，来觅数据，上海证券研究所*

**图7 AI Agent发展历程**




*资料来源：头豹研究院，上海证券研究所*

# 二、AI Agent应用广泛，商用爆发时点渐行渐近

- **AI Agent可应用于B端和C端，其中B端强调专业性，**Agent多应用于金融、医疗、法律、财务、生产物流、资管、人力等专业性强的领域；**C端强调自由度，**Agent多应用于游戏、娱乐、影视、营销、广告、教育等需要更多灵感及创意的领域。
- **全球巨头加码AI Agent，商业化节点渐行渐近。**谷歌发布321个全球顶级企业的AI应用实战案例，涵盖零售巨头沃尔玛、医疗巨头 Mayo Clinic，金融巨头花旗等公司的Agent落地案例。根据乌鸦君统计，在Agent六大落地核心场景中，雇员代理的应用最为普遍，在医疗健康（17个）、金融服务（16个）、科技（15个）领域都有广泛应用。从落地行业来看，科技行业应用Agent最为广泛，零售和消费品、医疗健康、金融服务行业也落地较多。

图8 AI Agent 场景特性总览


资料来源：甲子光年，上海证券研究所

表2 Google 321个AI代理场景一览（单位：个）

| 场景 | 客户代理 | 雇员代理 | 创意代理 | 数据代理 | 代码代理 | 安全代理 |
|---|---|---|---|---|---|---|
| 零售和消费品 | 14 | 11 | 5 | 9 | 1 | 2 |
| 汽车与物流 | 6 | 3 | - | 7 | - | - |
| 医疗健康 | 4 | 17 | - | 11 | 1 | 2 |
| 金融服务 | 3 | 16 | - | 10 | 3 | 5 |
| 公共部门和非营利组织 | 12 | 9 | 1 | 7 | - | - |
| 制造、工业和电子 | 5 | 10 | - | 6 | - | - |
| 媒体、营销和游戏 | 5 | 6 | 12 | 7 | - | - |
| 酒店与旅游 | 10 | 1 | 1 | - | - | - |
| 科技 | 13 | 15 | 5 | 17 | 14 | 11 |
| 电信 | 1 | 4 | - | - | - | - |
| 商业与专业服务 | 2 | 9 | - | 5 | 1 | - |

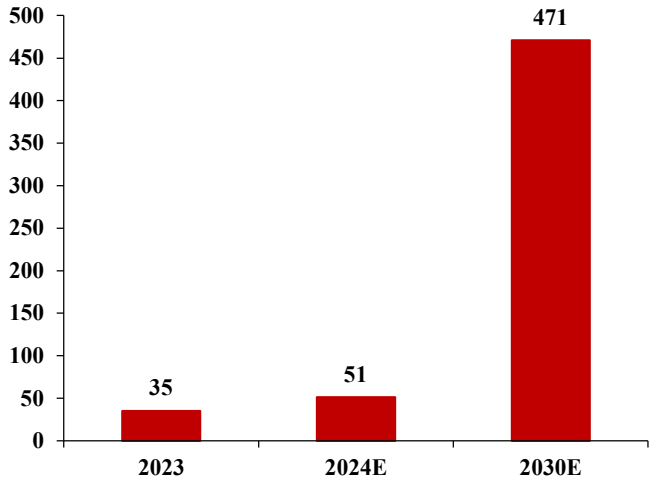
资料来源：36氪，上海证券研究所

# 二、AI Agent市场空间广阔

◆ **AI Agent处于快速发展阶段，市场发展潜力大。**根据Markets and Markets预测，全球AI Agent市场将从2024年的51亿美元增长到2030年的471亿美元，年复合增长率达44.8%。根据Gartner预测，到2028年，至少15%的日常工作决策将通过AI Agent自主完成（2024年为0%）。黄仁勋在2025年美国消费电子展上表示，AI智能体可能是下一个机器人行业，很可能是一个价值数万亿美元的机会。

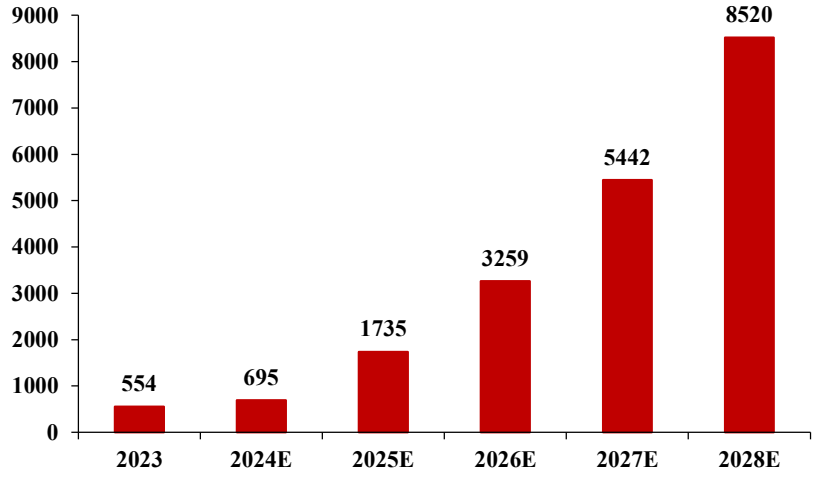
◆ **中国AI Agent市场持续增长。**2023年，AI Agent被业内正式引入，开始兴起。在To B端，AI Agent将逐渐把SaaS应用全面进行改写重构；在To C端，AI Agent作为生成式AI的商业化应用。根据头豹研究院，2023年中国AI Agent市场规模为554亿元，预计2028年中国AI Agent市场规模将达到8520亿元，2023-2028年均复合增长率达72.7%。

图9 2023-2030E全球AI Agent市场规模（亿美元）




资料来源：Markets and Markets，上海证券研究所

图10 中国AI Agent行业市场规模，2023-2028E（亿元）




资料来源：头豹研究院，上海证券研究所

# 二、B端：科技巨头持续加码，AI Agent商业化加速

◆ **微软：公布世界最大AI Agent生态系统。**2024年10月，微软发布10个Agent，覆盖销售、服务、财务、供应链等方面工作。2024年11月，在微软Ignite大会上，微软公布了世界最大AI Agent生态系统，并宣布企业用户可以通过Azure AI目录访问超过1800个AI模型，用于支持各类AI Agent的部署和运行，此外，微软还发布了5款预构建AI Agent。2025年1月，微软发布全新企业级AI助手Microsoft 365 Copilot Chat，可直接调用企业自有数据，执行端到端的超复杂自动化业务流程。

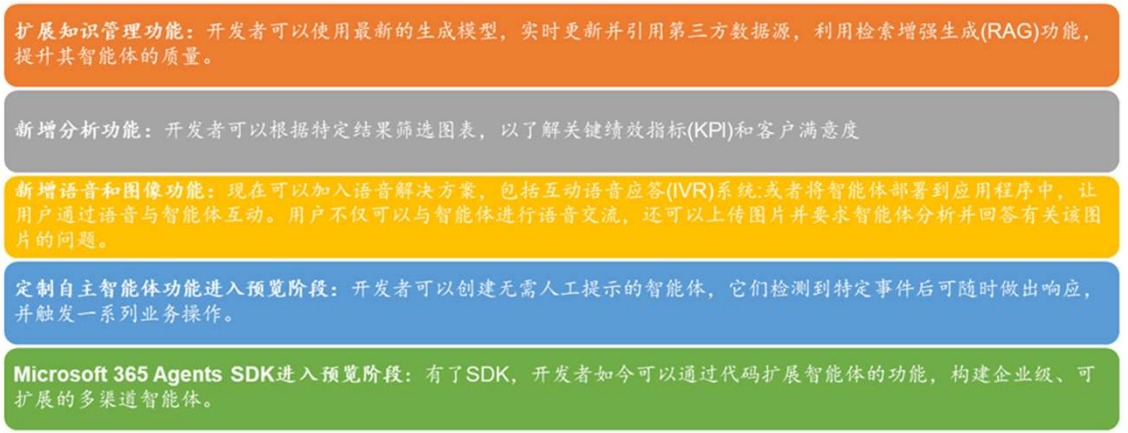
◆ **Copilot Studio平台加速智能体创建进程。**自推出以来，已经有超过10万家公司使用Copilot Studio创建了自己的AI智能体。比如，麦肯锡通过自动化的流程分配智能体，将项目受理流程从20天缩短至仅2天；Pets at Home在不到两周内部署了防欺诈智能体，每年节省数百万美元。

**表3 微软AI Agent为客户降本增效**

| 公司 | 降本增效主要体现 |
|---|---|
| Pets at Home | 利润保护团队的案例整理Agent，预计可节省七位数的年度成本 |
| 麦肯锡 | 客户入职Agent，测试显示可使流程时间缩短90%，减少30%处理工作 |
| 汤森路透 | 构建了一个专业级Agent来加快法律尽职调查工作流程，初步测试显示某些任务可以在一半时间内完成，还能帮助拓展新业务机会 |

*资料来源：量子位，上海证券研究所*

**图11 Copilot Studio平台更新**




*资料来源：新智元，上海证券研究所*

# 二、B端：科技巨头持续加码，AI Agent商业化加速

上海证券 SHANGHAI SECURITIES

- **谷歌：上线AI Agent Space一站式商用生态。**2024年11月，谷歌云宣布将提供从AI Agent的开发、部署到应用一站式商用生态。其中，谷歌发布了全球为数不多的商用AI Agent 市场（Space），类似苹果的Store。面向企业用户，用户可以在AI Agent市场中快速找到想要的AI Agent，极大简化了客户的选择和部署流程。同时，还提供了免费试用的机会。开发者则能通过用户的购买来赚取佣金。对于AI Agent的商业发展具有里程碑意义。
- **Gemini 2.0亮相，专为AI Agent打造。**2024年12月，谷歌发布新一代大模型Gemini 2.0，官方将其定位为面向智能体时代的AI模型。根据谷歌发布的基准测试结果，在多模态的图片、视频、编码、数学等能力上，仅Gemini 2.0 Flash实验版表现就已几乎全面超越Gemini 1.5 Pro 002，且速度是1.5 Pro的两倍。谷歌表示，2025年初会将Gemini 2.0扩展到更多旗下产品中，比如Project Astra。
- **此外，谷歌发布了一系列智能体。**发布Project Astra升级版，能够流畅地在多种语言和混合语言之间进行对话，并且能够理解不同口音和生僻单词，借助 Gemini 2.0，Project Astra 可以使用 Google Search、Google Lens 和 Google Maps，从而在日常生活中发挥助手的作用；发布完成复杂任务的智能体 Project Mariner，可从浏览器开始探索人机交互，能够理解和推理浏览器页面中的信息，包括像素和文本、代码、图像和表单等网页元素，然后通过Chrome 扩展程序使用这些信息为用户完成任务；发布编码智能体Jules、以及游戏和其他领域的智能体。

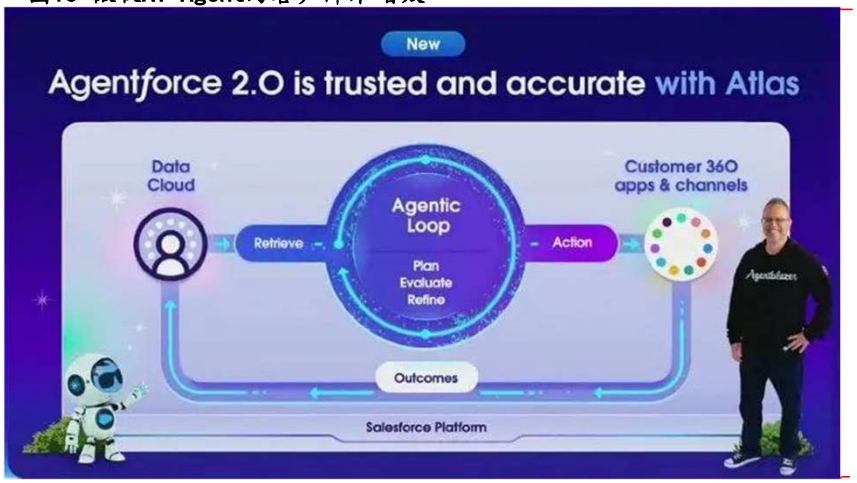
# 二、B端：科技巨头持续加码，AI Agent商业化加速

- **Salesforce：发布Agentforce。**在2024年9月的Dreamforce大会上，Salesforce全新发布Agentforce平台，企业可以创建销售代理、服务代理、营销代理等，完成多种日常任务。2024年12月，Salesforce发布了Agentforce 2.0。2.0核心的改进在于推理引擎的增强，该引擎旨在连接企业数据、业务流程和逻辑，从而提供更智能、更具上下文感知能力的AI交互体验。据董事长Benioff透露，客户在help.salesforce.com平台上每周32000次交互中，人工互动比例已从前的10000次降至5000次，83%的问题由AI代理解决。
- **AgentForce商业化进展积极，推动业绩强劲增长。**Salesforce为AgentForce引入基于使用量的定价模式，每次对话2美元。根据FY2025Q3业绩会，通过Agentforce和AI解决方案取得的客户成功推动了强劲的业绩。在第三季度，通过AI获得的100万美元以上订单数量同比增长两倍多，签署了2000多份AI交易，其中包括200多份Agentforce订单。而当考虑AgentForce面临的机遇时，这200笔交易只是冰山一角。

图12 Salesforce发布Agentforce


资料来源：Salesforce咨询，上海证券研究所

图13 微软AI Agent为客户降本增效




资料来源：Salesforce咨询，上海证券研究所

# 二、B端：科技巨头持续加码，AI Agent商业化加速

◆ **国内：大厂纷纷入局，抢跑智能体市场。**B端，百度文心智能体平台、腾讯元器、讯飞星火智能体创作中心、通义智能体、字节扣子等面向企业用户提供智能体创建平台，并开始在其AI智能助手界面中添加AI Agent入口。除这些大厂外，包括智谱AI、面壁智能等大模型创业公司，容联云、思迈特等SaaS公司，钉钉、飞书等协同办公赛道企业等，都在加码智能体开发和应用落地。

**表4 国内智能体开发平台**

| 公司/组织 | 项目名称 | 兼容大模型产品 | 主要功能点 | 智能体分发渠道 | 应用场景 |
|---|---|---|---|---|---|
| 阿里云魔搭社区 | modelscope-agent开发框架 | 开源大语言模型 | 开发者都可基于开源 LLM 搭建属于自己的智能体应用 | / | C/B端 |
| 阿里巴巴通义实验室 | 多智能体编程框架与开发平台AgentScope | OpenAI、DashScope、Gemini、Ollama等多种不同平台的模型API | 专门为多智能体应用开发者打造，便捷的“拖拽式”多智能体应用编排范式 | / | C/B端 |
| 华为方舟实验室 | 盘古智能体框架(Pangu-Agent) | 盘古大模型 | 用于将结构化推理整合到AI Agents的政策中并进行学习 | / | / |
| 字节跳动 | Coze 扣子 | 豆包、通义千问、智谱、MiniMax、Moonshot、Baichuan等 | 支持发布到多个渠道 | 平台支持用户将其一键发布到飞书、微信公众号、豆包等渠道 | C/B端 |
| 百度智能云 | Agent Builder(原灵境矩阵) | 文心一言 | 提供零代码、低代码两种低成本智能体开发模式 | 提供百度生态矩阵分发路径，打通百度搜索、小度智能硬件平台、文心一言、地图、车机等多场景、多设备，实现“开发+分发+运营+变现”一体化赋能。 | C/B端 |
| 腾讯云 | 腾讯元器 | 腾讯混元 | 一站式AI智能体创作与分发平台,主要包含两个板块，一块是开发智能体的工具，一块是商店，有智能体和插件商店 | 支持发布到元器、元宝、QQ、微信客服等平台，同时支持以 API的形式供三方软件进行调用 | C/B端 |
| 科大讯飞 | 星火企业智能平台 | 星火大模型 | 支持企业针自身具体场景打造专属智能体 | / | B端 |
| 昆仑万维 | SkyAgents | 天工大模型 | 允许用户通过自然语言输入和可视化拖拽来快速构建服务于具体业务场景的Al Agents | / | C/B端 |
| 智谱AI | 智谱清言智能体中心 | GLM-4 | 可让任何人都能够自由运用GLM-4模型并挖掘它的潜力，没有任何编程基础也能便捷地开发属于自己的大模型 | / | / |
| 面壁智能 | AI智能体应用框架XAgent | MiniCPM | XAgent被定义为超强AI智能体应用框架，能自行拆解复杂任务，面壁宣称XAgent的能力已经全面超越AutoGPT | / | B端为主 |
| | 智能体通用平台AgentVerse | | AgentVerse类似一个大模型宇宙，智能体通用平台，让每个Agent如同角色扮演一般加入其中并彼此互动 | | |
| | 多智能体协作开发框架ChatDev | | ChatDev可以看作是一个用Agent技术自动化开发软件应用 | | |

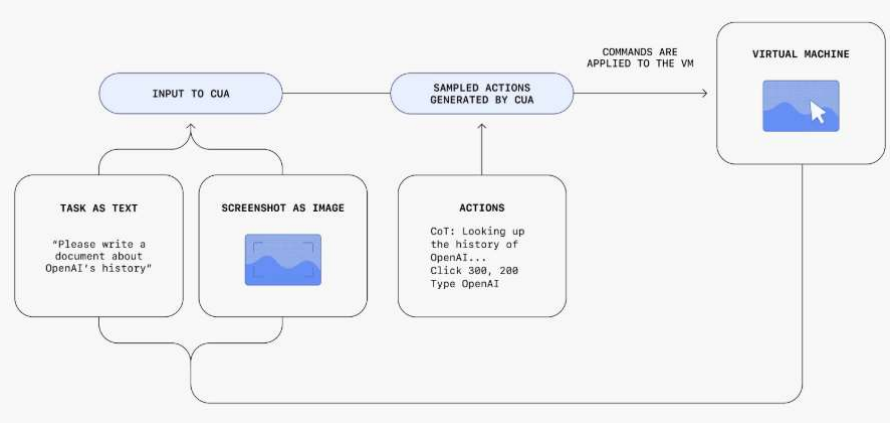
*资料来源：36氪，上海证券研究所*

# 二、C端：国内外Agent惊艳涌现

◆ **OpenAI：发布首款AI代理Operator。**1月24日，OpenAI直播发布首款AI代理工具Operator，能够代理用户执行基于网页的操作，替用户完成预订机票、预订晚餐、编写网站代码等几乎所有联网任务。Operator由一个名为CUA（计算机使用代理）的新模型驱动，结合了GPT-4o的视觉能力，以及通过强化学习实现的高级推理。Operator能够“看见”网页（截图），并使用鼠标和键盘允许的所有操作与网页进行互动。在操作中如果碰到困难，模型会调用推理能力进行自我纠正，若依然无法解决问题则会把控制权交还给人类。

◆ **发布可高效输出专业报告的Deep Research。**2月3日，OpenAI发布Deep Research（深度研究），是一个用o3模型造出的联网版推理Agent。Deep Research走专业路线，能搜索、解释和分析网络上的大量文本、图像、PDF，在极短时间内旁征博引，然后生成非常专业的综合分析报告，还附有搜索过程和索引。

**图14 Operator工作示意图**




*资料来源：财联社，上海证券研究所*

**图15 在高难度AI评估中，Deep Research准确率达26.6%**

| Model | Accuracy (%) |
|---|---|
| GPT-4o | 3.3 |
| Grok-2 | 3.8 |
| Claude 3.5 Sonnet | 4.3 |
| Gemini Thinking | 6.2 |
| **OpenAI o1** | **9.1** |
| DeepSeek-R1* | 9.4 |
| OpenAI o3-mini (medium)* | 10.5 |
| OpenAI o3-mini (high)* | 13.0 |
| **OpenAI deep research**** | **26.6** |

* Model is not multi-modal, evaluated on text-only subset.
**with browsing + python tools

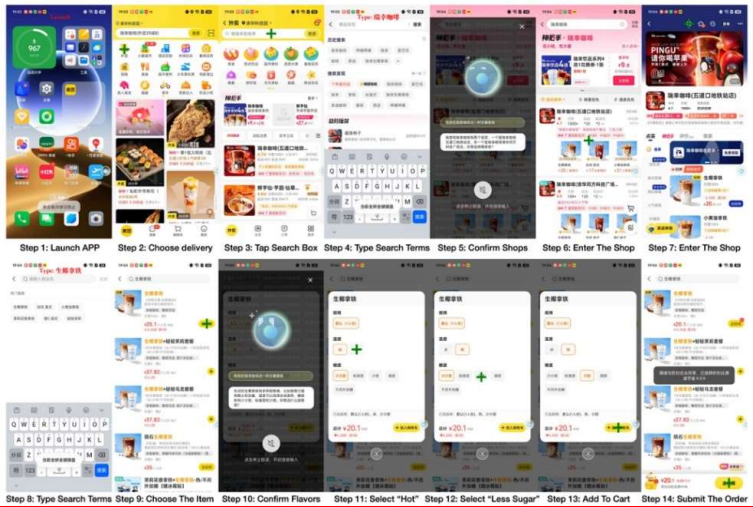
*资料来源：智东西，上海证券研究所*

# 二、C端：国内外Agent惊艳涌现

◆ **智谱发布三款AI Agent，覆盖手机、电脑、网页端。**2024年11月29日，智谱发布面向手机的phone use——AutoGLM，面向电脑的compute use——GLM PC，以及面向网页的GLM-Web能力。2024年10月25日，智谱就发布了AI手机端的AutoGLM。新升级的AutoGLM实现了一系列进步：在手机上具备了更多的能力，可以挑战更高难度的操作、支持更长的流程，甚至毫无打断地执行超过50步的操作；可以实现跨APP操作；具备短口令能力。GLM-PC则是能够操作电脑的生产力助手，具有发送信息、网页总结、文档处理、预定和参加会议、远程和定时操作等能力。

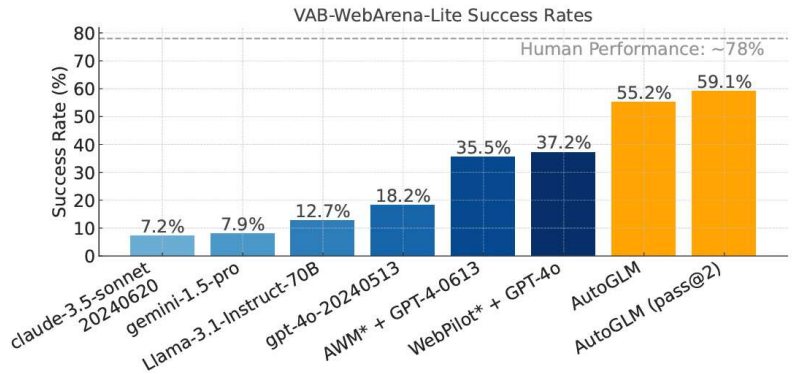
◆ **GLM-PC全新升级。**1月23日，智谱发布基于智谱多模态大模型 CogAgent的GLM-PC，是全球首个面向公众、回车即用的电脑智能体，能像人类一样「观察」和「操作」计算机，协助用户高效完成各类电脑任务。**截至目前，智谱已经有了手机智能体AutoGLM和电脑智能体GLM-PC两大系统，分别覆盖了移动设备和桌面端，实现了工具使用能力的深度突破。**

图16 AutoGLM订餐演示




资料来源：《AutoGLM: Autonomous Foundation Agents for GUIs》，上海证券研究所

图17 VAB-WebArena-Lite 上不同Agent的成功率

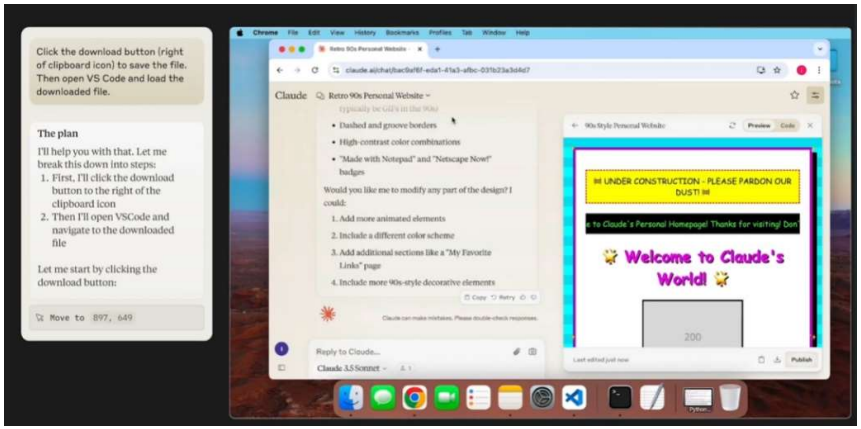



资料来源：《AutoGLM: Autonomous Foundation Agents for GUIs》，上海证券研究所

# 二、C端：国内外Agent惊艳涌现

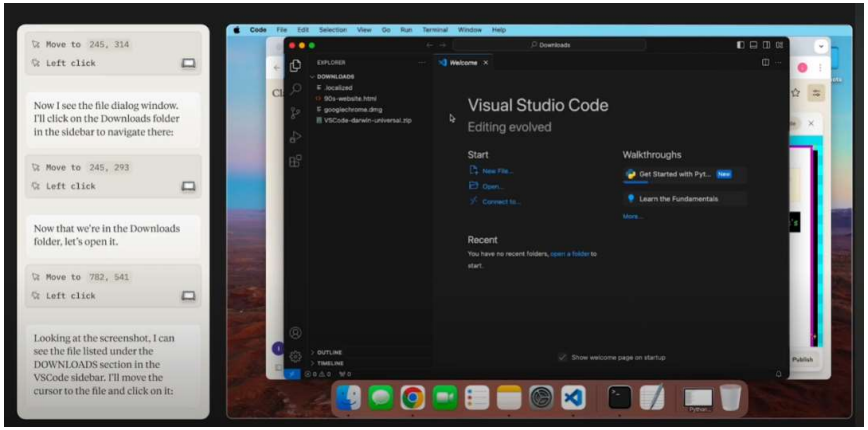
◆ **Anthropic：Claude 3.5模型更新，新增Computer use功能。**10月22日，Anthropic发布了Claude 3.5模型家族的更新，同时宣布升级版Claude 3.5 Sonnet获得Computer use功能（计算机使用能力），具体来说，Claude能够通过观看屏幕截图，实现移动光标、点击按钮、使用虚拟键盘输入文本等操作，真正模拟人类与计算机交互的方式。在多个演示视频中可以看到，Claude能够丝滑地操作电脑执行打开软件、网页搜索、文本输入、编写代码、下载文件、debug、查找网页表格并填入信息等任务，甚至还能打开外卖平台订餐。

图18 Claude通过Computer Use功能，使用Artifact功能编写代码




资料来源：DeepTech深科技，上海证券研究所

图19 Claude 能找到并打开电脑上的其他软件

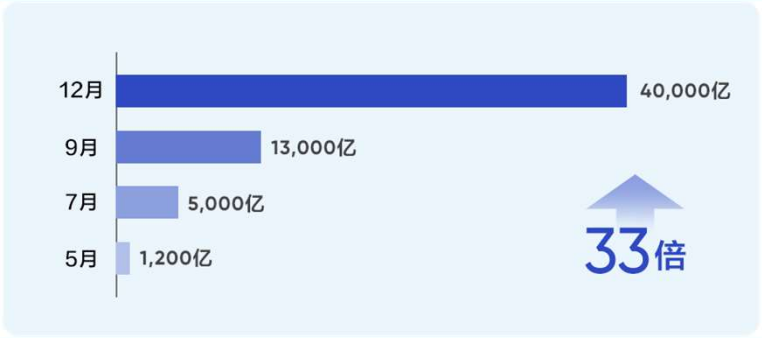



资料来源：DeepTech深科技，上海证券研究所

# 二、AI应用落地，推理算力需求提升

◆ **豆包应用全球火热，产品快速迭代，有望推动推理算力需求持续提升。**根据火山引擎，豆包大模型2024年12月日均tokens使用量超过4万亿，较5月发布时期增长超过33倍。量子位智库数据显示，截至11月底，豆包APP在2024年的累计用户规模已超过1.6亿，11月平均每天有80万新用户下载豆包，单日活跃用户近900万，位居AI应用全球第二、国内第一。根据非凡产研公众号，截止2025年1月底，豆包日活达1556万。2025年1月20日，豆包实时语音大模型正式上线，1月22日，字节跳动发布豆包大模型1.5 Pro，模型综合能力增强。

图20 豆包大模型调用量持续攀升




资料来源：火山引擎，上海证券研究所

图21 豆包大模型各应用场景调用量增长


资料来源：火山引擎，上海证券研究所

# 二、AI应用落地，推理算力需求提升

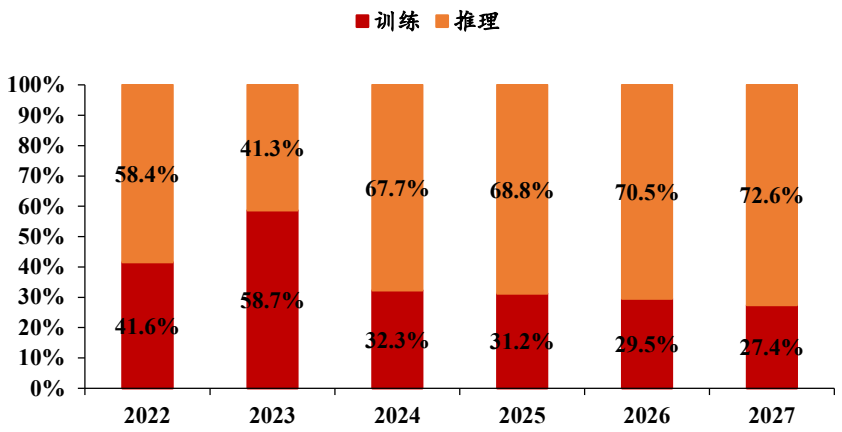
上海证券 SHANGHAI SECURITIES

- **DeepSeek火热出圈，日活超3000万。**1 月 20 日，DeepSeek 发布全新的开源推理大模型 DeepSeek-R1，在数学、编程和推理等多个任务上达到了与 OpenAI o1 相当的表现水平。根据央视网消息，近期，DeepSeek访问使用量急速上升，已经成为目前最快突破3000万日活跃用户量的应用程序。
- **DeepSeek在技术层面实现多项创新。**（1）开源强化学习引领推理计算范式转换：DeepSeek使用纯粹 RL（强化学习），无需 SFT（监督微调），不依赖冷启动数据，成功地实现了靠纯 RL（强化学习）来激励大模型的推理能力。（2）MLA 和 MoE 等引领大模型架构创新：DeepSeek 在 Transformer 架构的基础上创新了多头潜在注意力MLA（用于高效推理）和混合专家模型MoE（用于高效训练）。（3）“贴身定制”的软硬协同工程优化：从计算、存储、通信等多个层面实施了软硬协同的工程优化策略，比如混合精度训练、跨节点通信优化、双流水线机制、DualPipe 算法等。
- **大模型成本降低，促进AI应用生态繁荣。**DeepSeek自身通过“模型架构创新”和“软硬件协同工程优化”将大模型训练成本大幅度降低，大约是 Meta 的 1/10，OpenAI 的 1/20；同时，通过最开放的 MIT 开源协议，和将推理大模型蒸馏给开源小模型等一系列工程方法，为业界带来低成本的端侧模型商品。我们认为，DeepSeek的成本创新，有望降低技术门槛，带来AI平权，加速AI应用的爆发。
- **DeepSeek算力节省有望重塑AI 产业版图。**通过 MoE 和 MLA 等对经典 Transformer 架构进行的改进和迭代，DeepSeek 使用少于同行 10-20 倍的算力，完成了同等规模的预训练。随着强化学习成为后训练阶段的标配，推理计算将占比越来越大（相对预训练计算），适应预训练模式的 GPU 大卡集群计算将不再是未来 AI 算力需求的主流。我们认为，英伟达在推理计算方面的优势相对较弱，ASIC 以及国产芯片有望逐步抢占英伟达GPU的份额，迎来发展机遇。

# 二、AI应用落地，推理算力需求提升

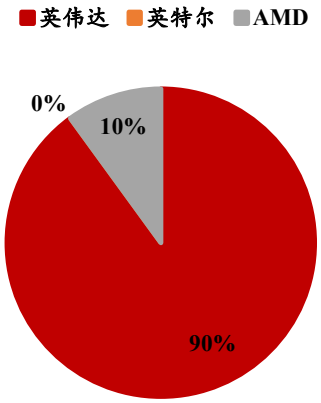
- **全球推理算力需求井喷。**巴克莱报告预计，到2026年，消费者AI日活跃用户（DAUs）将突破10亿，而企业AI代理的采用率可能占全球70亿软件任务的5%。AI推理计算需求将快速提升，预计占通用人工智能总计算需求的70%以上，推理计算的需求甚至可以超过训练计算需求，达到后者的4.5倍。行业需增加当前预测4倍的芯片资本支出，总额或接近3000亿美元。
- **国内推理算力需求持续提升。**IDC数据显示，在中国，2023上半年训练工作负载的服务器占比达到49.4%，预计全年的占比将达到58.7%。随着训练模型的完善与成熟，模型和应用产品逐步进入投产模式，处理推理工作负载的人工智能服务器占比将随之攀升。IDC预计，到2027年，用于推理的工作负载将达到72.6%。
- **推理算力需求将带动AI芯片市场呈百花齐放局面。**根据JPR，2024Q3英伟达在全球GPU市场份额高达90%。我们认为，随着推理算力的兴起， ASIC 以及国产芯片将迎来新发展机遇。

图22 中国人工智能服务器工作负载预测，2022-2027




资料来源：IDC，上海证券研究所

图23 2024Q3 AI芯片企业竞争格局




资料来源：Jon Peddie Research，快科技，上海证券研究所

# 三、投资建议

建议关注：

1. AI算力：海光信息、寒武纪、中科曙光、神州数码、软通动力、华丰科技、泰嘉股份、申菱环境、英维克、润泽科技、安博通等；
2. AI应用：金山办公、科大讯飞、万兴科技、新致软件、梅安森、鼎捷数智、汉得信息、能科科技、佳发教育、竞业达、泛微网络、软通动力、中软国际、润和软件等。

# 四、风险提示



1. AI应用落地不及预期；
2. AI需求不及预期；
3. 行业竞争加剧。

# 行业评级与免责声明

**分析师声明**

作者具有中国证券业协会授予的证券投资咨询资格或相当的专业胜任能力，以勤勉尽责的职业态度，独立、客观地出具本报告，并保证报告采用的信息均来自合规渠道，力求清晰、准确地反映作者的研究观点，结论不受任何第三方的授意或影响。此外，作者薪酬的任何部分不与本报告中的具体推荐意见或观点直接或间接相关。

**公司业务资格说明**

本公司具备证券投资咨询业务资格。

**投资评级体系与评级定义**

| | | |
|---|---|---|
| 股票投资评级： | 分析师给出下列评级中的其中一项代表其根据公司基本面及（或）估值预期以报告日起6个月内公司股价相对于同期市场基准指数表现的看法。 | |
| | 买入 | 股价表现将强于基准指数20%以上 |
| | 增持 | 股价表现将强于基准指数5-20% |
| | 中性 | 股价表现将介于基准指数±5%之间 |
| | 减持 | 股价表现将弱于基准指数5%以上 |
| | 无评级 | 由于我们无法获取必要的资料，或者公司面临无法预见结果的重大不确定性事件，或者其他原因，致使我们无法给出明确的投资评级 |
| 行业投资评级： | 分析师给出下列评级中的其中一项代表其根据行业历史基本面及（或）估值对所研究行业以报告日起12个月内的基本面和行业指数相对于同期市场基准指数表现的看法。 | |
| | 增持 | 行业基本面看好，相对表现优于同期基准指数 |
| | 中性 | 行业基本面稳定，相对表现与同期基准指数持平 |
| | 减持 | 行业基本面看淡，相对表现弱于同期基准指数 |
| 相关证券市场基准指数说明：A股市场以沪深300指数为基准；港股市场以恒生指数为基准；美股市场以标普500或纳斯达克综合指数为基准。 | | |

**投资评级说明：**

不同证券研究机构采用不同的评级术语及评级标准，投资者应区分不同机构在相同评级名称下的定义差异。本评级体系采用的是相对评级体系。投资者买卖证券的决定取决于个人的实际情况。投资者应阅读整篇报告，以获取比较完整的观点与信息，投资者不应以分析师的投资评级取代个人的分析与判断。

# 行业评级与免责声明

**免责声明**

本报告仅供上海证券有限责任公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。

本报告版权归本公司所有，本公司对本报告保留一切权利。未经书面授权，任何机构和个人均不得对本报告进行任何形式的发布、复制、引用或转载。如经过本公司同意引用、刊发的，须注明出处为上海证券有限责任公司研究所，且不得对本报告进行有悖原意的引用、删节和修改。

在法律许可的情况下，本公司或其关联机构可能会持有报告中涉及的公司所发行的证券或期权并进行交易，也可能为这些公司提供或争取提供多种金融服务。

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

本报告中的内容和意见仅供参考，并不构成客户私人咨询建议。在任何情况下，本公司、本公司员工或关联机构不承诺投资者一定获利，不与投资者分享投资收益，也不对任何人因使用本报告中的任何内容所引致的任何损失负责，投资者据此做出的任何投资决策与本公司、本公司员工或关联机构无关。

市场有风险，投资需谨慎。投资者不应将本报告作为投资决策的唯一参考因素，也不应当认为本报告可以取代自己的判断。