计算机

证券研究报告
2024 年 12 月 09 日

AI Agent 与端侧新入口共筑 AI 应用未来

AI Agent 正在快速发展，海外大厂积极布局 Agent 构建应用

全球 AI Agent 市场正以高速增长的趋势重塑各行业的运营模式和客户交互体验。AI Agent 通过记忆能力、规划能力、行动能力和工具能力，与人类用户、外界环境、其他 Agents 及系统开发者实现高效协作。LLM 作为 AI Agent 的核心能力构建基础，结合规划、记忆、工具和行动四大能力模块，实现了对复杂任务理解、分解与执行。海外头部企业在 AI Agent 上持续发力，微软在 Ignite 大会宣布了全球最大规模的企业级 AI Agent 生态，同时 M365 Copilot 也增加了更多的 Agent 功能；Salesforce 于 2024 年 9 月 12 日发布 Agentforce，旨在将 AI 智能体与人类协作、数据云、CRM 等核心模块相结合，为企业提供全面的客户服务和销售解决方案。

从 Claude3.5 到 GLM，Agent 在 C 端硬件落地正进入临界点

Claude3.5 推出最新 Sonnet 模型新增 Computer use 功能，可以类似人一样使用电脑，输入语言指令后模型感知界面互动，自然语言控制硬件更进一步。同时智谱推出 AutoGLM，实现了在安卓系统手机和 Web 端的自然语言交互功能；硬件厂商中，我们看到苹果带头推出 Apple Intelligence，华为也相应推出了 Harmoney Intelligence。

C 端入口迎来重塑，互联网巨头有望加速逆向 AI 硬件卡位布局，同时云端协同技术路线有望带来云端推理算力增量

端侧 AI 有望成为下一代流量入口的关键，手机、耳机等硬件厂商正在集体加入大模型能力，以 Apple intelligence 为首，三星、华为等硬件厂商正在探索在其新系统中集成自然语言处理能力，或有望简化硬件交互流程。我们认为互联网公司有望借助模型与软件能力逆向卡位争夺流量入口，长期来看，专业化端侧与全能云端协同或是端侧 AI 的最优解，AI Agent 是端侧 AI 的重要一环考虑到 AI agent 需要规划+多次调用大模型，以 Anthopic 为例，其新推出的 Claude Sonnet 模型在使用 Computer Use 功能时展现出较高的成本，我们认为端侧 AI 会带来大量的云端推理算力增量。

建议关注：
国产算力：寒武纪、海光信息、盛科通信（通信覆盖）、神州数码、中科曙光、景嘉微
AI 应用：金山办公、金蝶国际、科大讯飞、泛微网络、致远互联、鼎捷数智、用友网络、汉得信息、恒生电子、万兴科技、虹软科技
端侧硬件：高通（美股）、炬芯科技、中科蓝讯、恒玄科技

风险提示：AI 应用进展不及预期、算力受到制裁的风险、大模型技术推进不及预期

投资评级

行业评级	强于大市(维持评级)
上次评级	强于大市

作者

缪欣君 分析师
SAC 执业证书编号：S1110517080003
miaoxinjun@tfzq.com
刘鉴 联系人
liujianb@tfzq.com

行业走势图

资料来源：聚源数据

相关报告

1 《计算机-行业点评:资金支撑叠加供需共振，泛信创自主可控有望加速》2024-11-11
2 《计算机-行业点评:TSMC 事件的长期影响：加速国产算力行业格局集中，利好具备全国产供应链头部厂商》2024-11-10
3 《计算机-行业点评:跳出成交量看同花顺：居民财富管理被动化的长期核心受益者》 2024-11-10

内容目录

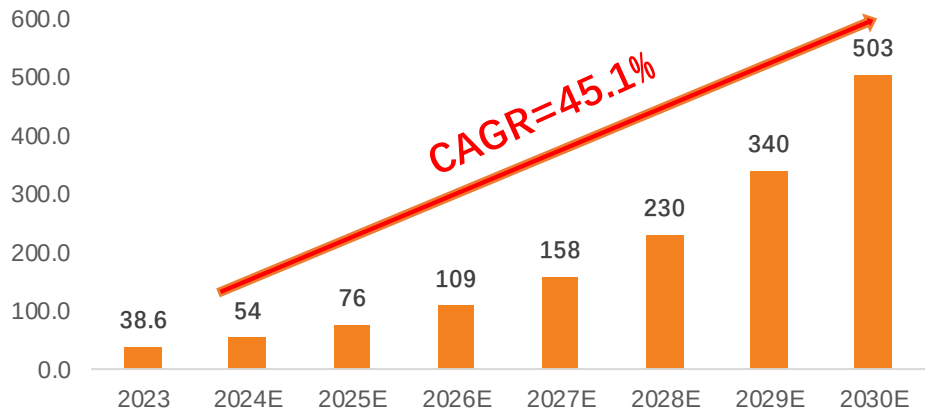
图表目录

1. AI Agent 逐步发力，构筑大模型应用未来

1.1. AI Agent 临界点即将到来，为 AI 应用构筑基石

全球 AI Agent 市场正以高速增长的趋势重塑各行业的运营模式和客户交互体验。根据 Grand View Reasearch 数据，全球 AI 智能体市场在 2023 年的规模已达到 38.6 亿美元，并预计从 2024 年到 2030 年将以 CAGR 45.1% 快速增长，2030 年市场规模有望突破 503 亿美元；推动这一增长的核心驱动力包括自动化需求的增加、自然语言处理（NLP）等技术的进步，以及消费者对个性化体验和实时服务的期望不断提升。AI Agent 通过分析客户行为、优化产品推荐和提高客户参与度，广泛应用于电商、医疗和安全等多个领域，显著提升了企业在营销、销售和客户服务方面的效率，同时降低了运营成本，为企业创造了更大的商业价值。

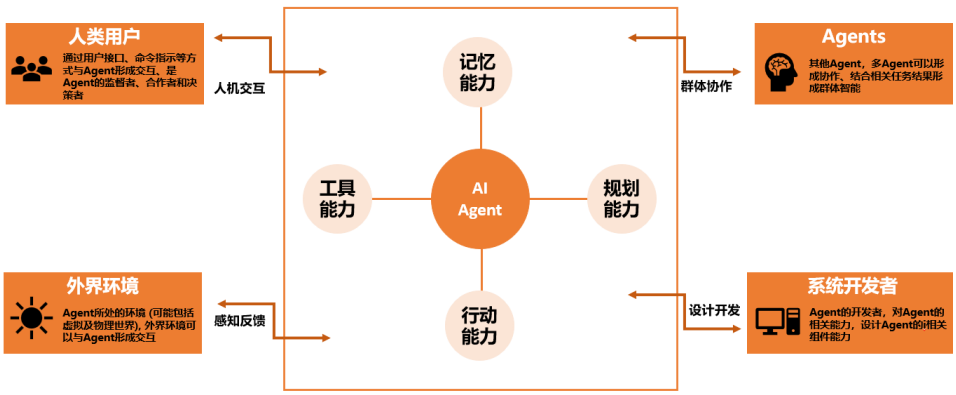
图 1：2023 年-2030 年全球 AI Agent 市场规模预计（亿美元）



资料来源：Grand View Research，天风证券研究所

AI Agent（智能体）通过记忆能力、规划能力、行动能力和工具能力，与人类用户、外界环境、其他 Agents 及系统开发者实现高效协作。分工上，人类用户通过接口与 Agent 交互，作为监督者、合作者和决策者；外界环境为 Agent 提供感知和反馈交互空间；多个 Agent 之间通过协作整合任务结果，形成群体智能；系统开发者则负责设计开发 Agent 的相关能力模块，确保其功能可靠和高效运行。

图 2：AI Agent 架构



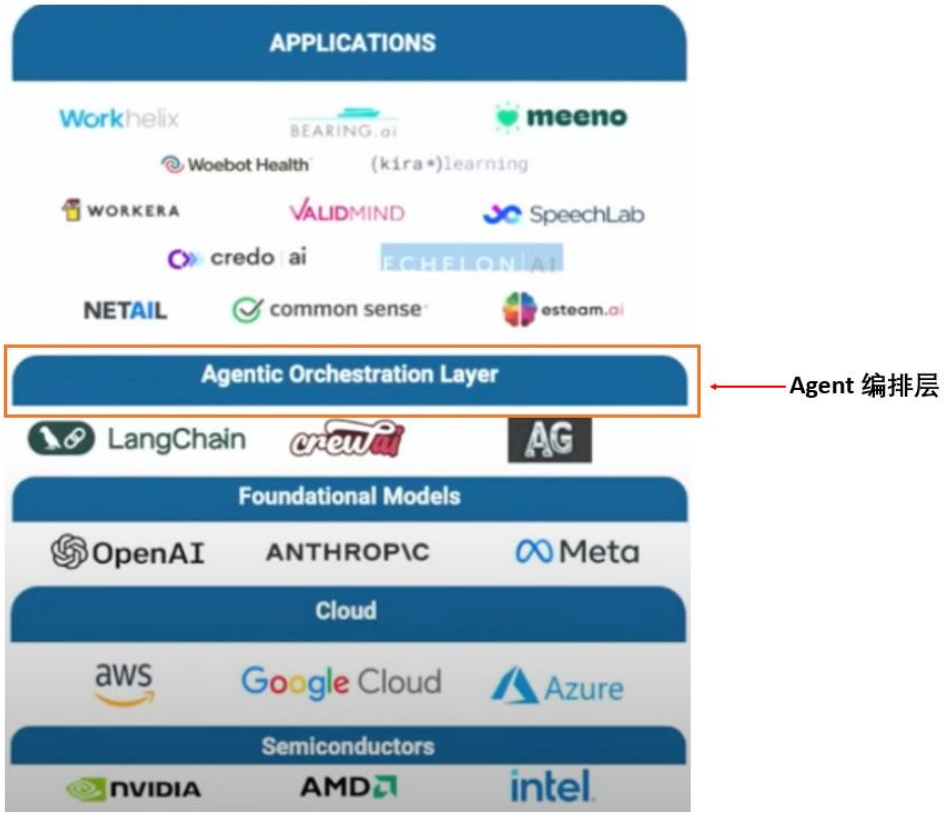
资料来源：甲子光年公众号，天风证券研究所

LLM 作为 AI Agent 的核心能力构建基础，结合规划、记忆、工具和行动四大能力模块，实现了对复杂任务理解、分解与执行。LLM 不仅提升了 AI Agent 的理解能力和泛化能力，还显著增强了其在多任务处理和上下文信息解析方面的效率。我们认为，结合规划、记忆、

工具和行动四大能力模块，AI Agent 能够更好地支持复杂任务执行和多样化场景需求，提供高效且连贯的交互体验。

在此基础上，Agent 专有的编排层在 AI 生态系统中发挥着核心作用，通过整合大模型（如 OpenAI 和 Anthropic 等）与云服务（如 AWS、Google Cloud、Azure），实现任务的动态分配与高效协作。它不仅作为连接基础设施与应用程序的桥梁，还为上层应用提供智能化支持，极大地提升了 AI 系统的灵活性和创新能力。我们认为，Agent 编排层是驱动 AI 生态高效运行的关键环节。

图 3：Agent 编排层作为连接基础设施与应用的智能核心



资料来源：灵犀科技公众号，天风证券研究所

人工智能著名学者、斯坦福大学教授吴恩达认为，智能体 AI Agent 领域有 4 种模式，包括反思、工具使用、规划和多智能体协作。使用 AI agent，AI 能做的事情会大幅扩展。

Reflection（反思）：让 LLM 审视并修正自己生成的输出。

Tool Use（工具使用）：LLM 使用网络搜索、代码执行等工具来帮助它收集信息、采取行动或处理数据。

Planning（规划）：LLM 分解复杂任务，制定并执行多步骤计划来实现目标。

Multi-agent Collaboration（多智能体协作）：多个 AI Agent 协同工作，通过分解任务、讨论和辩论来提出比单个智能更好的解决方案。

1.2. 海外头部企业在 AI Agent 上持续发力

1.2.1. 微软在推出大规模企业级 AI Agent 生态

微软在 Ignite 大会宣布了全球最大规模的企业级 AI Agent 生态，同时 M365 Copilot 也增加了更多的 Agent 功能。公司宣布企业可以在 agent 中使用 Azure 目录中 1800 个 LLM 中的任何模型，不再依赖于 OpenAI 的独家模型。

此前微软推出了 Copilot Studio，一个可以让用户能够创建、管理和将 Agent 连接到

Copilot 的平台。自 Copilot Studio 推出以来，已经有超过 10 万家公司，用 Copilot Studio 创建了自己的 AI Agent。在微软 Ignite 2024 大会中，Copilot Studio 迎来了重点以下更新：

1）**扩展知识管理功能：**开发者可以使用最新的生成模型，实时更新并引用第三方数据源(如 Salesforce、ServiceNow 和 Zendesk)，利用检索增强生成(RAG)功能，提升其 Agent 的质量。同时，该功能还整合了 Azure AI Foundry，以支持更复杂、更定制化的场景。此外，还引入了更高级的 Azure AI 功能，支持访问 Azure AI 模型目录中的 1800 个模型，并支持开发者直接在 Copilot studio 中访问并调用自己定制微调的模型。

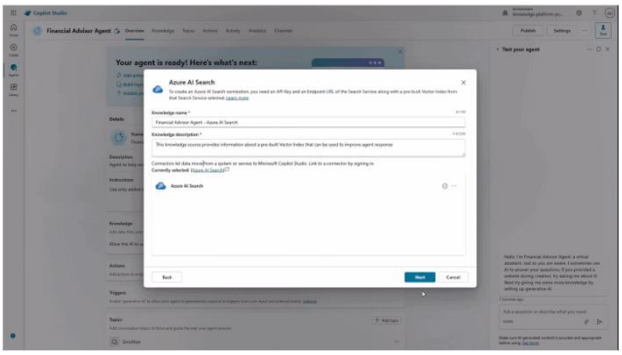
2）**新增分析功能：**开发者可以根据特定结果筛选图表，以了解关键绩效指标和客户满意度。

3）**新增语音和图像功能：**现在可以加入语音解决方案，包括互动语音应答（IVR）系统；或者将智能体部署到应用程序中，让用户通过语音与智能体互动。用户不仅可以与智能体进行语音交流，还可以上传图片并要求智能体分析并回答有关该图片的问题。

4）**定制自主智能体功能进入预览阶段：**开发者可以创建无需人工提示的智能体，它们检测到特定事件后可随时做出响应，并触发一系列业务操作。

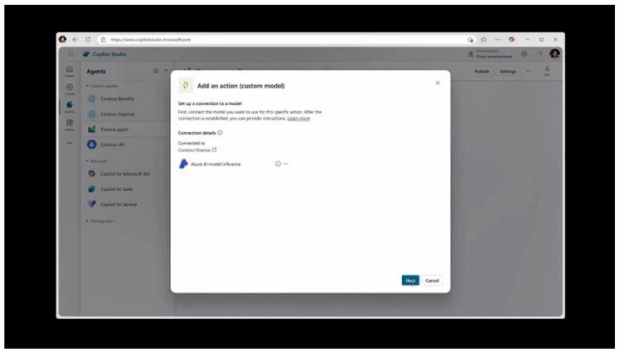
5）**Mircosoft 365 Agents SDK 进入预览阶段：**有了 SDK，开发者如今可以通过代码扩展智能体的功能，构建企业级、可扩展的多渠道智能体。

图 4：Copilot Studio 整合 Azure AI Search 构建向量化知识库



资料来源：新智元公众号，天风证券研究所

图 5：开发者通过 Copilot Studio 集成 Azure AI 自定义模型



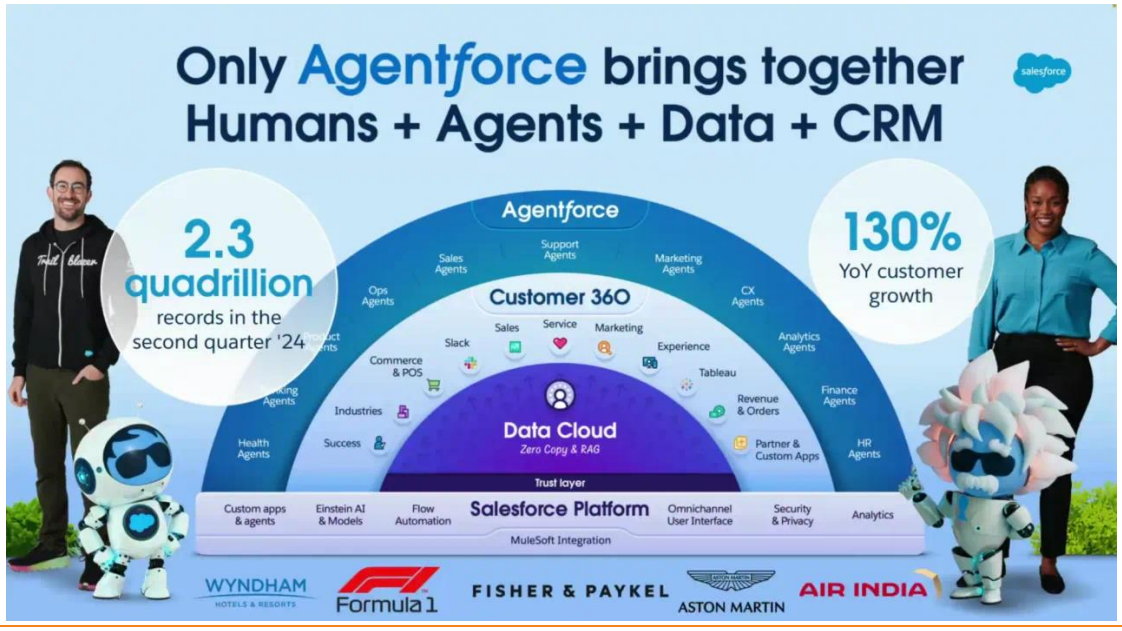
资料来源：新智元公众号，天风证券研究所

此外，微软 Ignite 大会推出了一系列强大的 Agent，涵盖多个场景：SharePoint Agent 可快速提取项目详情、总结备忘录或查找文档，为销售等团队提供高效支持；**Interpreter Agent** 为 Teams 会议提供实时语音翻译并模拟用户声音，增强跨语言沟通体验；**Employee Self-Service Agent** 简化 HR 和 IT 任务，如福利查询或设备申请，并可在 Copilot Studio 中定制。同时，**Facilitator Agent** 实时记录会议笔记，**Project Manager Agent** 自动创建计划并完成任务。合作伙伴如 **ServiceNow、Workday** 和 **Cohere** 等也将其 Agent 整合到 Copilot 中，覆盖 HR、财务和销售领域的核心知识。**S&P Global** 和 **CB Insights** 等公司还引入了新型连接器，使 Copilot 能够引用更多类型的业务数据，进一步推动工作流智能化发展。

1.2.2. Salesforce 正式推出 Agentforce 并在某些场景取得了 PMF

Salesforce 于 2024 年 9 月 12 日发布 Agentforce，旨在将 AI 智能体与人类协作、数据云、CRM 等核心模块相结合，为企业提供全面的客户服务和销售解决方案。**该平台可以自动处理客户请求、优化销售流程，并以低成本高效执行任务，从而提升客户体验并显著提高投资回报率。截至 2024 年 Q2，Agentforce 已管理了 2.3 千万亿条数据记录，并帮助企业客户实现了 130% 客户数量年增长率。**

图 6：Agentforce—人类+智能体+数据+CRM 协作的全新业务模式



资料来源：Salesforce 咨询公众号，天风证券研究所

Agentforce 通过智能化、自动化的 AI 助手，将客户服务中的长时间等待和繁琐的人工流程降至最低，为客户提供快速、高效的服务体验，主要应用场景可分为服务代理（Service Agent）、销售助手（SDR）、销售导师（Sales Coach）、个人购物助手（Personal Shopper）、活动策划助理（Campaign Assistant）。**Agentforce 由 Agent Builder 和 Agentforce Service Agent 两部分组成：**

- **Agent Builder** 让用户可以通过简单的配置，轻松打造定制化的 AI 助手。无论是内置功能还是自定义选项，Agentforce 都支持灵活扩展，满足不同业务场景需求。
- **Agentforce Service Agent** 是面向客户的 AI 服务助手，支持多渠道（语音、WhatsApp、Facebook Messenger）自助服务，帮助企业快速响应客户需求。

图 7：Agentforce 主要应用场景：赋能企业服务、销售与个性化体验

资料来源：Salesforce 官网，天风证券研究所

Agentforce 的定价模式为按次付费，具体为每次客户交互对话收费 2 美元。这一灵活的定价结构适合各种规模的企业，尤其在客户服务和销售场景中具有显著的成本效益。通过该定价模式，企业可根据实际使用量控制运营成本，同时高效利用 AI 驱动的自动化工具优化业务流程。此外，Salesforce 还推出了 Agentforce Foundations 计划，为企业提供包含 1,000 次对话、潜在客户管理以及 250,000 个 Data Cloud 积分的基础套餐，以支持企业在现有 CRM 系统中快速集成智能功能。

图 8：Agentforce 相关产品定价

资料来源：Salesforce 官网，天风证券研究所

Agentforce 的成功案例展示了其在优化企业服务效率、提升客户体验以及提升企业客户 ROI 方面的显著成效。通过 Agentforce，Wiley 的投资回报率达到了 213%，显著降低了成本并提高了生产力；OpenTable 借助 AI 智能代理优化客户服务，大幅提升服务代表的效率；Saks 通过统一数据与 AI Agent，进一步增强了其奢侈购物体验；ezCater 则通过 Agentforce 的自动化客户服务，简化了工作场所的食品订单流程。

图 9：Agentforce 为企业客户带来的实际价值与成果展示

资料来源：Salesforce 官网，天风证券研究所

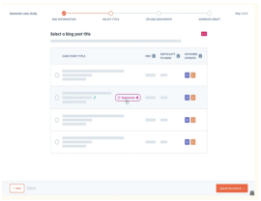
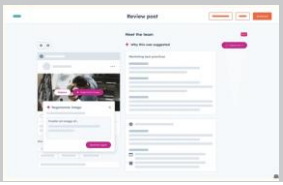
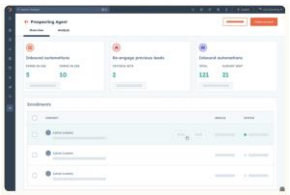
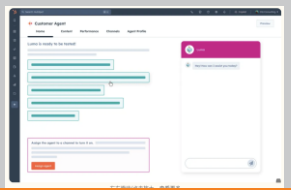
1.2.3. HubSpot：全新 AI 品牌 Breeze 打造营销和销售新体验

HubSpot 推出全新 AI 品牌 Breeze，赋能企业营销、销售和客户服务团队。2024 年 HubSpot 在 INBOUND 大会和秋季 Spotlight 发布会上，推出的全新 AI 品牌 Breeze。作为一个面向 GTM 团队的完整 AI 解决方案，Breeze 简单易用、快速高效，且能与企业的客户数据完美整合。Breeze 由一系列创新组件组成，主要包含以下部分：

- **Breeze Copilot：**提供易用的生成式 AI 助手，能够基于 CRM 数据快速生成内容、总结信息、安排任务，提高工作效率。
- **Breeze Agents：AI 专家团队，**覆盖从内容营销到客户服务的多个领域，自动化工作流程，助力企业扩展业务规模。
- **Breeze Intelligence：AI 数据分析师，**结合先进的语言模型和 AI 数据源，提供超过 2 亿个买家和公司信息，为企业决策提供可操作的深度洞察。
- **Breeze Features：**通过集成式工具解决创意资源不足的问题，优化内容创作和团队写作效率。

Breeze Agents 作为 HubSpot 推出的 AI 专家团队解决方案，旨在全面提升企业运营效率与业务增长表现。目前它的主要应用包括内容营销专家 (Content Agent)、社媒专家 (Social Media Agent)、销售专家 (Prospecting Agent) 和客服专家 (Customer Agent)。这些智能代理利用 HubSpot 智能 CRM 数据，分别实现高质量内容生成、优化社媒表现、提高销售转化率，以及全天候高效客服支持。这些功能帮助企业简化任务、优化资源分配，并提升客户互动体验，显著提高团队整体生产力。

表 1：Breeze Agents 主要应用功能及操作界面

应用	主要功能	操作界面
内容营销专家	可自动生成高质量的营销内容，例如落地页、播客、博客文章、案例研究等，并能根据企业品牌风格进行定制。	
社媒专家	可分析企业社交媒体表现，并根据用户的受众、行业趋势和最佳实践来创建帖子。用户只需要点击 “批准” 即可发布。	
销售专家	可研究潜在客户，并向 HubSpot 智能 CRM 中的潜在客户发送个性化的邮件进行跟进，从而提高转化率。	
客服专家	全天候自动回复客户问题，提供即时支持。通过学习企业知识库、网站和博客内容来快速上手，还能自动将复杂问题转交给人工客服。	

资料来源：HubSpot 社区公众号，天风证券研究所

2. 从 Claude3.5 到 AutoGLM，Agent 在 C 端应用正进入临界点

2.1. Claude3.5 推出最新 Sonnet 与 Haiku 模型，同时推出 Use computer 功能

10 月 23 日凌晨，Claude 3.5 重磅升级，首发全新模型 Claude 3.5 Haiku，并同时推出新版 Claude 3.5 Sonnet。从能力上看，新版 Claude 3.5 Sonnet 各项能力都得到显著提升，在多项能力超过 GPT-4o，例如在 GPQA、MMLU、HumanEval 等测试数据集上的得分均达到目前的领先水平。

升级后的 Claude 3.5 Sonnet 在行业基准测试中也有大幅改进，尤其在自主编码和工具使用任务上尤为突出。在编码方面，Claude 3.5 Sonnet 在 SWE-bench Verified 上的表现从 33.4% 提升至 49.0%，超过所有公开可用的模型，包括 OpenAI o1-preview 和专为自主编码设计的系统。不仅如此，它在 TAU-bench（一个自主工具使用任务）中的表现也有显著提升，成功率从 62.6% 增加到 69.2%，在更具挑战性的航空领域从 36.0% 提升到 46.0%。

图 10：Anthropic 推出新的 Claude3.5 模型，新 Sonnet 和 Haiku 模型在能力上得到显著提升

	Claude 3.5 Sonnet (new)	Claude 3.5 Haiku	Claude 3.5 Sonnet	GPT-4o*	GPT-4o mini*	Gemini 1.5 Pro	Gemini 1.5 Flash
Graduate level reasoning *GPQA (Diamond)*	65.0% 0-shot CoT	41.6% 0-shot CoT	59.4% 0-shot CoT	53.6% 0-shot CoT	40.2% 0-shot CoT	59.1% 0-shot CoT	51.0% 0-shot CoT
Undergraduate level knowledge *MMLU Pro*	78.0% 0-shot CoT	65.0% 0-shot CoT	75.1% 0-shot CoT	—	—	75.8% 0-shot CoT	67.3% 0-shot CoT
Code *HumanEval*	93.7% 0-shot	88.1% 0-shot	92.0% 0-shot	90.2% 0-shot	87.2% 0-shot	—	—
Math problem-solving *MATH*	78.3% 0-shot CoT	69.2% 0-shot CoT	71.1% 0-shot CoT	76.6% 0-shot CoT	70.2% 0-shot CoT	86.5% 4-shot CoT	77.9% 4-shot CoT
High school math competition *AIME 2024*	16.0% 0-shot CoT	5.3% 0-shot CoT	9.6% 0-shot CoT	9.3% 0-shot CoT	—	—	—
Visual Q/A *MMMU*	70.4% 0-shot CoT	—	68.3% 0-shot CoT	69.1% 0-shot CoT	59.4% 0-shot CoT	65.9% 0-shot CoT	62.3% 0-shot CoT
Agentic coding *SWE-bench Verified*	49.0%	40.6%	33.4%	—	—	—	—
Agentic tool use *TAU-bench*	Retail 69.2% Airline 46.0%	Retail 51.0% Airline 22.8%	Retail 62.6% Airline 36.0%	—	—	—	—

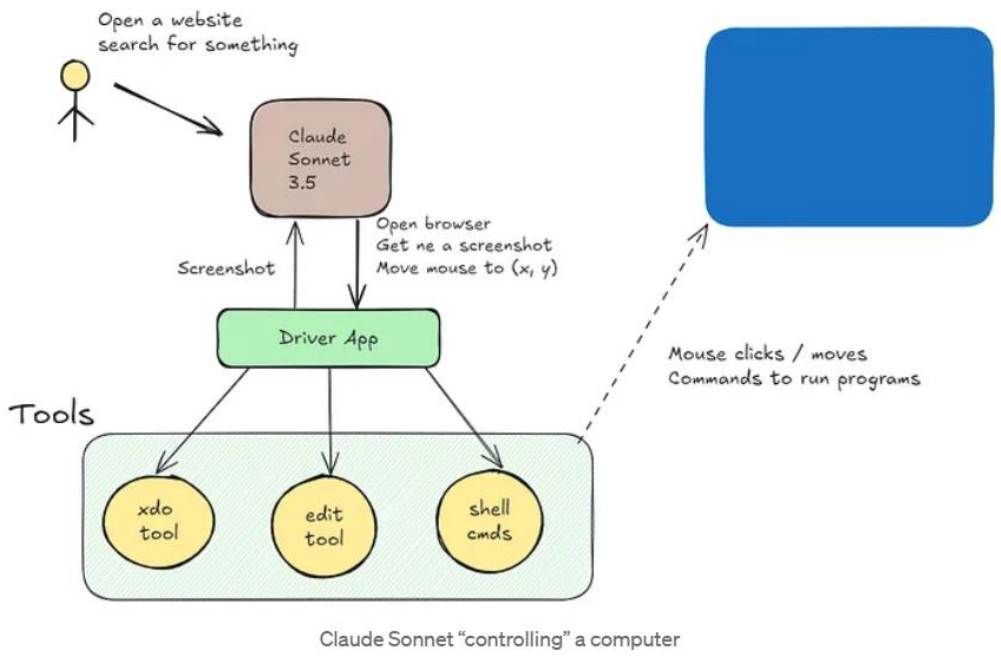
* Our evaluation tables exclude OpenAI's o1 model family as they depend on extensive pre-response computation time, unlike typical models. This fundamental difference makes performance comparisons difficult.

资料来源：Anthropic 官网，天风证券研究所

升级版 Claude 3.5 Sonnet 具备一项突破性的全新能力“Computer use”，即开发者可以通过 API 指示 Claude 像人一样使用计算机，包括观察屏幕、移动光标、点击按钮和输入文本等。在实现该功能的过程中，Anthropic AI 尝试了一种新方法，教会它通用计算机技能，使其能够使用一系列为人设计的标准工具和软件程序。基于这样的设计理念，Anthropic AI 构建了一个 API，使 Claude 能够感知并与计算机界面互动。开发者可以集成该 API，使 Claude 能够将指令转化为计算机命令，实现任务的自动化和智能化。

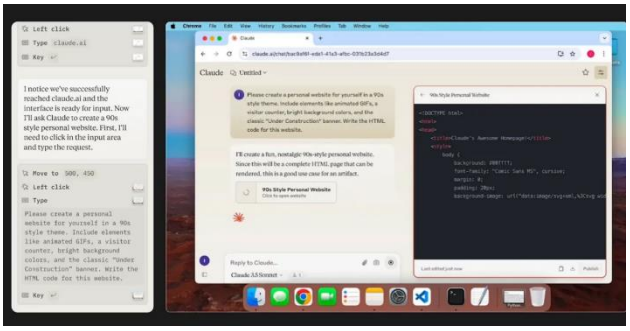
Anthropic 研究员 Alex Albert 亲自录制了一个 demo，如何利用 Claude 自动完成一个网站编码任务。解决这个问题共分为四个步骤：（1）要求 Claude 在 Chrome 浏览器中导航到 Claude.ai，并创建一个有趣的、以 90 年代为主题的个人主页；（2）想要对此网站做一些修改，可要求 Claude 点击下载并保存文件，然后在 VS Code 中将其打开；（3）让 Claude 启动一个服务器，以便在浏览器中查看该文件。（4）Alex Albert 发现终端输出中有个错误，即顶部还缺少了一个文件图标，便要求 Claude 自己识别并将其修复。结果：Claude 顺利找到并删除了引发错误的代码行。

图 11：Claude Sonnet 模型控制电脑的方法



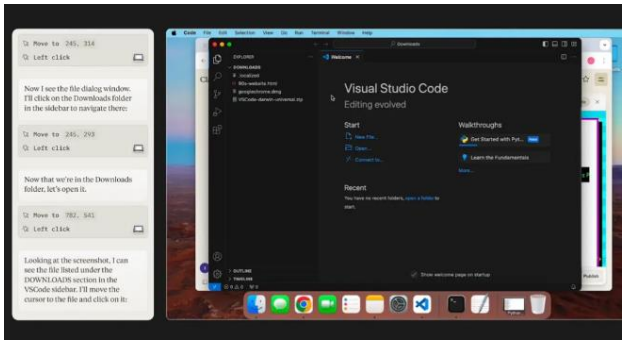
资料来源：汇智网，天风证券研究所

图 12：Alex Albert 演示第一步：Claude 在 Chrome 浏览器中导航到 Claude.ai，并创建一个有趣的、以 90 年代为主题的个人主页



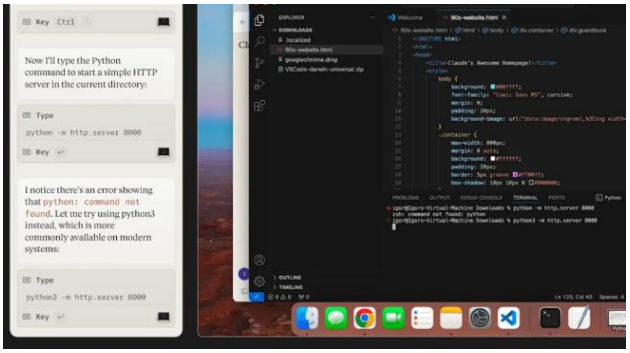
资料来源：CSDN 学习公众号，天风证券研究所

图 13：Alex Albert 演示第二步：想要对这个网站做一些修改，要求 Claude 点击下载并保存文件，然后在 VS Code 中将其打开



资料来源：CSDN 学习公众号，天风证券研究所

图 14：Alex Albert 演示第三步：让 Claude 启动一个服务器，以便在浏览器中查看该文件



资料来源：CSDN 学习公众号，天风证券研究所

图 15：Alex Albert 演示第四步：发现终端输出中有个错误，要求 Claude 自己识别并将其修复

资料来源：CSDN 学习公众号，天风证券研究所

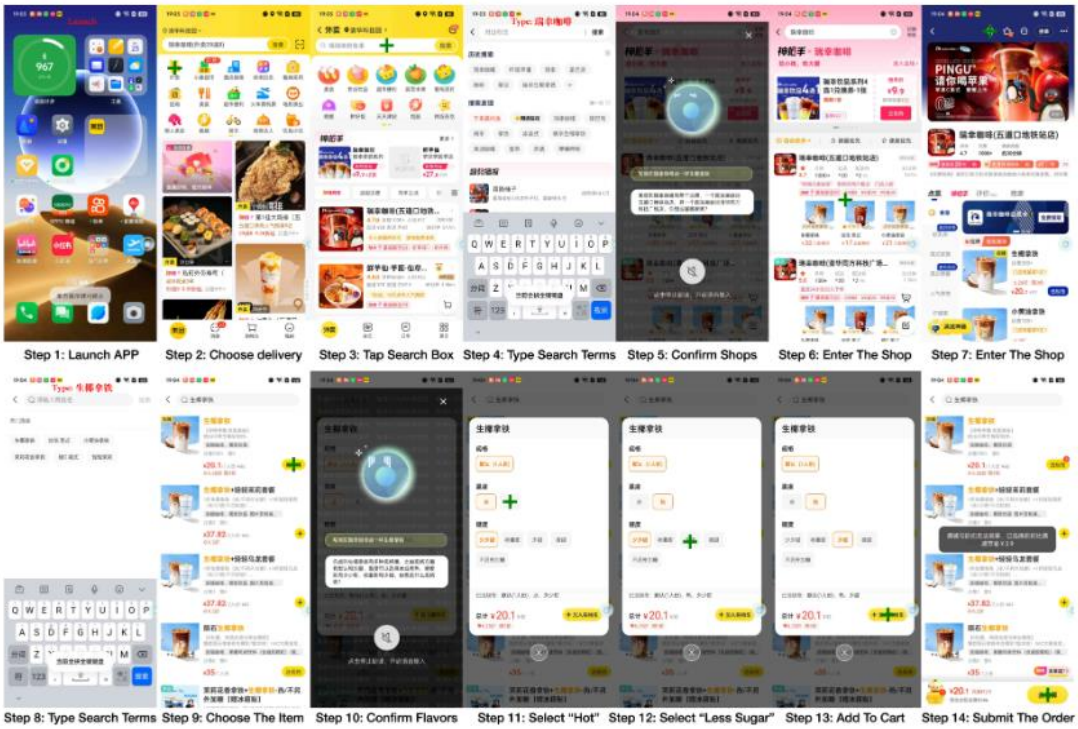
Claude 3.5 Sonnet 已在公测阶段，且是业内首个在公测阶段提供“Computer use”功能的前沿 AI 模型。目前模型仍处于实验阶段，存在一定的出错可能。例如，在一项旨在测试 AI Agent 帮助完成机票预订任务的评估中，升级版 Claude 3.5 Sonnet 成功完成的任务不到一半；在另一项涉及发起退票等任务的测试中，Claude 3.5 Sonnet 的失败率也超过了 30%。Anthropic 承认，目前 Claude 3.5 Sonnet 的“Computer use”功能仍不完美。一些人们能轻松完成的操作（如滚动、拖动、缩放）目前对 Claude 来说仍具挑战性，整体速度也很慢，因此鼓励开发者从低风险任务开始探索。

尽管如此，Anthropic 依旧对“Computer use”的前景充满期待，并相信它将随着时间的推移迅速改善。据悉，目前 Asana、Canva、Cognition、DoorDash、Replit 和 The Browser Company 等公司已经开始探索 Claude 3.5 Sonnet 的可能性，尝试令其自动执行需要数十，甚至上百个步骤才能完成的任务。

2.2. 智谱推出 AutoGLM，助力硬件交互

2024 年 10 月 25 日，智谱 AI 推出自主智能体 AutoGLM，一个能代替你在手机和网页上完成各种操作的 AI 助手。AutoGLM 是基于图形用户界面实现自主任务完成，模拟人类在手机操作。AutoGLM 能接收简单的文字或语音指令，自动完成复杂的操作流程，无需用户手动干预。AutoGLM 的主要功能特点包括（1）实时操作:能实时响应指令，在手机上执行复杂的任务序列（2）无需 API 调用:不依赖于特定的 API 接口，直接与图形用户界面交互（3）自动化任务执行:在真实环境中执行自动化任务，简化用户操作流程。

图 16：AutoGLM 使用示例



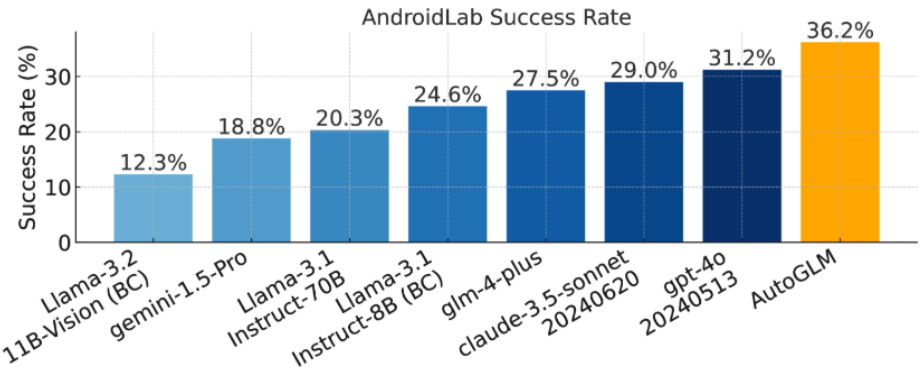
资料来源：《AutoGLM: Autonomous Foundation Agents for GUIs》（作者包括 Xiao Liu、Bo Qin 等），天风证券研究所

AutoGLM 在手机端可以使用的场景包括：（1）社交媒体管理，在社交平台上自动执行点赞、评论、分享等操作。（2）在线购物，在电商平台上搜索商品、比较价格、下单购买、跟踪物流等旅行预订，在旅游网站上搜索并预订酒店、机票、火车票等。（3）外卖订购，在外卖平台上浏览菜单、下单、支付以及追踪订单状态。（4）日常信息查询:如查询天气、新闻、股票信息等。

技术核心包括基础智能体解耦合中间界面和自进化在线课程强化学习框架，让 AutoGLM 能精确执行动作、灵活规划任务，克服传统大模型智能体在动作执行精确度和任务规划灵

活性上的挑战。首先，设计一个中间接口可以将基础 GUI 代理中的规划和基础行为分开，分离使开发更加敏捷并增强了性能。此外，错误恢复对于代理应用程序至关重要，但仅通过离线培训很难获得，公司通过自我发展的 RL 来应对这一挑战，根据渐进的弱到强课程表以在线方式实施。最终，AutoGLM 在 Web 端和安卓端都取得了超过 SOTA 的能力，以安卓系统为例，在 Androidlab Success Rate 数据集上，AutoGLM 实现了 36.2%的成功率，超过 GPT-4o 与 Claude 等大模型。

图 17：AutoGLM 在 AndroidLab 上的成功率行业领先



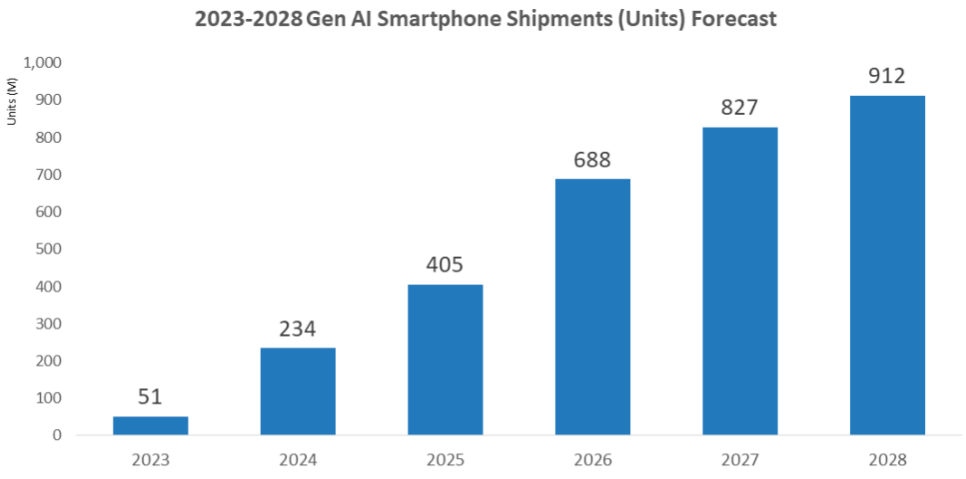
资料来源：《AutoGLM: Autonomous Foundation Agents for GUIs》（作者包括 Xiao Liu、Bo Qin 等）。天风证券研究所

3. 端侧 AI 成为海内外大模型重要方向与大厂必争之地

作为争夺下一代流量入口的关键机遇，端侧 AI 已然成为各大厂商必争之地。虽然短期内还存在各种困难，包括电池续航和散热问题>显存带宽>GPU 算力和显存容量，成为一系列亟待解决的难点。尽管如此，终端生态多方的信心并没有受到影响，各行业正在使尽浑身解数共同促进端侧 AI 的实现。

IDC 最新预测估计，2024 年生成式 AI 手机的出货量将同比增长 364%，达到 2.342 亿部。到 2028 年，全球生成式 AI 智能手机的出货量将达到 9.12 亿部，按照这个预测，2023 年-2028 年生成式 AI 的智能手机的 CAGR 高达 78.4%。**从 2022 年 ChatGPT 引发的生成式 AI 迅速崛起以来，谷歌、三星等各大厂商，都在尝试将 AI 大模型内置于手机，完成 AI 在手机终端的部署。**

图 18：生成式 AI 手机未来有望迎来快速增长



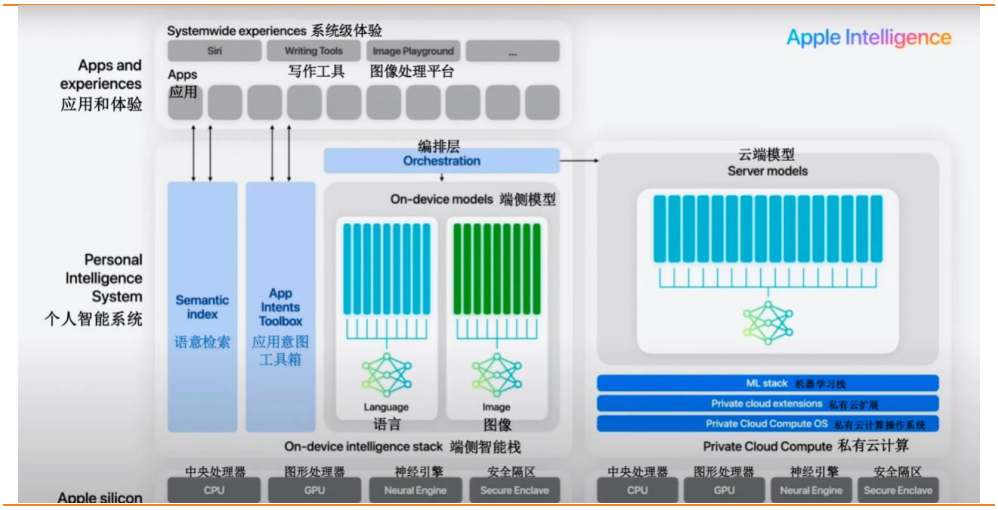
资料来源：新智元公众号，天风证券研究所

3.1. 苹果领衔，在 iPhone 推出 Apple Intelligence 与应用

在今年 6 月的 WWDC24 苹果全球开发者大会上，全世界第一次听到“Apple Intelligence”苹果定义了自己的 Apple AI。Apple Intelligence 也是今年 iOS 18、iPhone 16 系列的最大亮点。Apple Intelligence 功能能深刻理解语言含义，支持邮件、备忘录、Safari 浏览器、Pages 文稿、Keynote 讲演以及第三方 App，大幅提升用户体验。另外，Apple Intelligence 也让 Siri 有了大幅升级，在处理复杂指令时，苹果结合了 ChatGPT 的能力，Siri 不仅能精准理解用户意图，还能提供更为智能的回应

Apple Intelligence 目前仅支持美国英语版本，苹果计划在 12 月份开放支持澳大利亚、加拿大、新西兰、南非和英国等地的英语方言。中文、法语、日语和西班牙语版本将于明年正式推出。

图 19：Apple Intelligence 的全景图



资料来源：腾讯科技公众号，天风证券研究所

在 Apple Intelligence 全景图中，个人智能系统层可以说是 Apple Intelligence 最为核心的结构。这其中包含 AFM-on-device（Apple Foundation Model 端侧模型）与云端模型（AFM Server）。编排层来负责判断用户需求是依靠端侧解决还是要上传云端。苹果在这里没有进行任何人工干预，完全依靠算法自行判断，用户无法决定自己的数据是不是仅放在端侧。

3.2. 华为鸿蒙推出 Harmoney Intelligence

6 月 21 日，华为在其 2024 开发者大会上宣布 HarmonyOS NEXT（鸿蒙星河版）即日起面向开发者和先锋用户启动测试，正式版将在今年第四季度推出，可升级设备包括 Mate 60 系列、Mate X5 系列、Pura 70 系列、MatePad Pro 等。此外，**HarmonyOS NEXT 首次将 AI 融入系统，推出 Harmony Intelligence（鸿蒙原生智能），依托昇腾的算力和盘古大模型，提供系统级的 AI 能力。**Harmony Intelligence 端云结合，允许在端侧处理图像、通话、文档、搜索领域的 AI 功能，而云侧调用华为盘古大模型及其他第三方大模型。整个架构的最顶层是鸿蒙原生应用和小艺智能体。

图 20：Harmony Intelligence 全景图

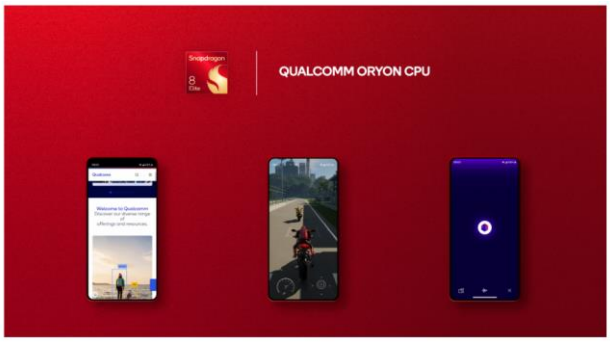
资料来源：新皮层 NewNewThing 公众号，天风证券研究所

3.3. 智谱推出最新 Chat-GLM4V 模型提供端侧多模态交互能力

10 月 21 日，高通骁龙峰会在夏威夷毛伊岛开幕，高通公司正式发布了新一代年度旗舰手机 SoC 骁龙 8 Elite 和用在 AI PC 里的第二代高通 Oryon CPU，其新的 Hexagon NPU 支持了端侧多模态，支持 4k 上下文窗口。**智谱与高通技术公司宣布合作将 GLM-4V 端侧视觉大模型面向骁龙 8 至尊版进行深度适配和推理优化，支持丰富的多模态交互方式，**进一步推动多模态生成式 AI 在终端侧的部署和推广，赋能更加情境化、个性化的终端侧智能体验。

多模态生成式 AI 模型能够利用终端侧丰富的传感器数据，例如文本、图像、音频、视频等，打造更加直观、无缝的智能交互体验。通过与骁龙 8 至尊版进行深度适配和推理优化，终端侧多模态应用 ChatGLM 能够支持三种终端侧交互方式：使用相机进行实时语音对话、上传照片进行对话、上传视频进行对话。**GLM-4V-Mini、GLM-4V-Nano 端侧视觉大模型和 GLM-4-9B 模型即将在高通 AI Hub 上线，搭载骁龙 8 至尊版的商用手机均可支持。**

图 21：高通推出最新一代 CPU 与 NPU



资料来源：高通中国公众号，天风证券研究所

图 22：智谱 GLM-4V 在端侧的应用

资料来源：智谱公众号，天风证券研究所

3.4. 腾讯混元大模型与高通芯片在端侧深度合作

骁龙峰会期间，高通技术公司宣布与腾讯混元合作，基于骁龙 8 至尊版移动平台，共同推

动了腾讯混元大模型 7B 和 3B 版本的终端侧部署，展示了此合作实现出色的运行表现。

腾讯混元大模型已为腾讯内部超过 700 个业务场景和 C 端应用提供底层技术支持，包括微信输入法、腾讯手机管家、QQ、腾讯视频、QQ 浏览器、企业微信、腾讯会议等，通过实现面向骁龙 8 至尊版的终端侧部署，能够利用终端侧生成式 AI 的丰富优势，更好地满足广泛的终端侧业务需求。例如，腾讯手机管家短信智能识别功能率先利用腾讯混元的终端侧模型能力，通过海量数据结合深度神经网络与预训练，让模型具备极强的语义理解能力，通过结合上下文语境信息更准确地理解短信意图，使短信召回率大幅提高将近 200%，识别准确率提升 20%。由于部分短信涉及用户个人敏感信息，端侧 AI 还可以在保证出色性能表现的同时，有效保护用户的个人信息隐私安全。

4. AI Agent 成为端侧应用重要支柱

端侧模型有一个不可能三角：性能、参数量和内存及功耗占用。性能优异需要大参数量；而参数量大就意味着内存占用大，功耗也会大；功耗过大又可能会影响性能。**AGI 是一个长期演进的过程，Agent 有望成为当务之急的“解决问题”，这一特性在端侧尤为重要。**通过主动工作流的配置辅以性能不错的模型解决问题，而主动工作流的配置离不开 Agent 技术的应用。

根据腾讯研究院公众号的内容指出，从电池容量的角度来看，通过工作流优化任务的实现是刚需。目前，由于手机和 PC 的保有量占据绝对优势，它们理所当然地成为了端侧 AI 的最佳落地方向，但在落地过程中，由于面临的芯片和电池的挑战，为了实现大模型的终端落地，需要进行大量的适配工作。

微软作为端侧模型的有力竞争者 Phi-3/3.5 的开发者，除了模型本身，还提供了一套名为 Agents 的工具。通过 Microsoft Copilot Studio 的升级，Copilot+PC 不仅可以调用 Windows 附带的 40 多个端侧 AI 模型提供支持，还可以构建成百上千的自动化业务流程，在客户需求下独立工作，从而实现长期运行的业务流程自动化。未来，移动端和 PC 端体验到的端侧 AI，大部分将是通过适配器和分类器挑选的微调小模型，以及针对特殊需求开发的自定义 Agents，而无需调用全量的模型参数，是更具性价比的方案。

另一方面，从生态搭建的维度，需要 Agent 调用多方资源以实现繁荣。无论是端侧还是云端 AI，大模型都只是底层计算；要实现用户价值，还需要一个繁荣的应用生态和强大的工具集来提供支持。不论是现有的 APP 形态，还是未来可能实现的“去皮化”的 API 形态，除了底层计算，还需要通过 Agent 来实现价值的连接。在 6 月的 Apple Intelligence 发布会上，苹果表示，Siri 的全新形态将改变交互规则，大量 AI 新功能将很快上线；此外，屏幕读取以及 App 内与 App 之间的操作等能力预计明年到位，这将使 AI 真正串联起苹果生态下的诸多应用。苹果提前承诺的这项能力，源自其在 4 月份发布的一项名为“Ferret-UI”的新技术。Ferret-UI 能够“看懂”手机屏幕，建立对 UI 元素的基本理解，奠定了执行复杂任务的基础，并通过分层次的任务设计，最终实现对用户指令的理解和响应。这本质上是一种通过视觉方式来构建主动工作流 Agent 的思路。底层模型本身并不能直接创造价值，苹果需要维持其最强的盈利因素：生态位。

长期来看，专业化端侧与全能云端协同或是端侧 AI 的最优解。云端模型比端侧模型先进一个数量级。虽然许多小模型在特定能力上已经具备了媲美十倍甚至百倍参数大模型的实力，但事实上，当前基础模型的综合能力依然基本遵循 Scaling Law 法则。千亿、万亿参数的大模型以及实验版本模型，作为探索 AGI 的最前沿模型，其整体智力水平无疑会持续领先。云端大模型始终比端侧大模型先进一个以上的数量级。例如，8 月份谷歌发布的轻量级小模型 Gemma 2 2B，是从 6 月份发布的 Gemma 27B 和 9B Gemma 2 模型中蒸馏而来的；微软开发的 AI 小语言模型（SLMs）Phi-3 系列有多个版本，包括 mini（3.8B）、small（7B）和 medium（14B）。根据微软公布的不同表现水平，在同一时间段内，参数规模仍然是决定大模型综合能力的关键因素。苹果的 Apple Intelligence 通过一个对标 GPT-4 的云端模型 Apple Server 来处理复杂任务，这不仅是因为终端设备受限于芯片、电池和发热等因素的选择，更是为端侧提供“无所不知、无所不能”的云端支持保留了重要的接口。

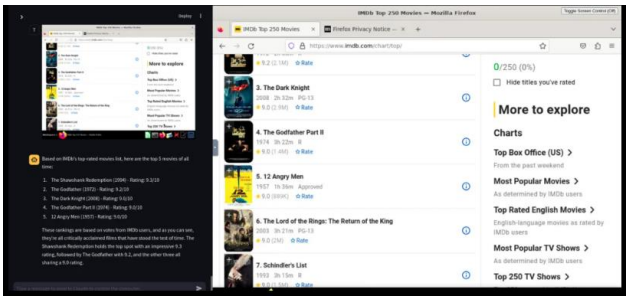
5. 端侧应用依托云端算力有望带来大量推理需求

根据云端协同的技术路线，我们认为端侧的广泛应用有望带来大量云端推理算力。端侧 AI 正处于一个积极探索的过程。纯端侧 AI 虽然是各大厂商追求的终极形态，但它并不会太快到达，甚至不一定会到来；就像大模型通往 AGI 的过程，这大概会是一个相当漫长的过程。然而，这并不妨碍端侧 AI 体验的提前实现，通过高质量数据、专业化目标任务训练以及云端隐私方案的混合协同与优化，端侧 AI，也可以逐渐从“可用”发展到“好用”。

以 Anthopic 为例，其新推出的 Claude Sonnet 模型在使用 Computer Use 功能时展现出较高的成本。根据 Meidum 官网的演示过程，单次使用 Computer Use 功能的成本为 0.88 美元。单从成本上看，模型仍处于早期阶段。整个过程较为缓慢且经常超过 API 速率限制。简单的任务也都需要反复看屏幕，这意味着更多的截图和 Tokens 消耗。但我们认为这是一个里程碑式的功能，未来代表着模型有望理解并控制计算机等设备。

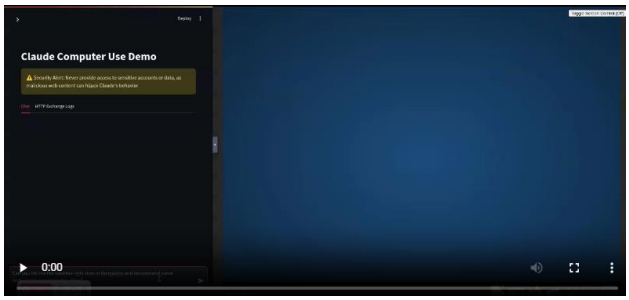
Composio 上开发者的测试也表明了 Computer Use 会带来大量的 tokens 消耗和推理算力，开发人员使用 Claude 的 Computer Use 功能执行了 4 项任务，包括查找 Top5 的电影并生成 CSV 文件、寻找某城市的最佳餐厅、在线订餐和购物，上述四项简单测试花费 30 美元，代表了大量的 tokens 消耗。 **AI Agent 是端侧 AI 的重要一环，我们预计端侧应用在很长一段时间都将是端侧+云端搭配使用，考虑到 AI agent 需要规划+多次调用大模型，我们认为端侧 AI 会带来大量的云端推理算力增量。**

图 23：使用 Computer Use 查找 top5 的电影



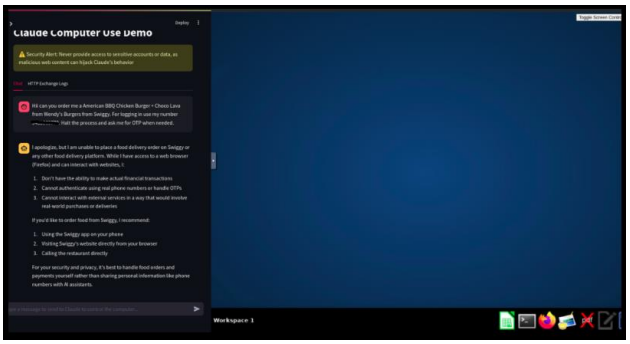
资料来源：Composio 官网，天风证券研究所

图 24：使用 Computer Use 根据城市天气查找最佳餐厅



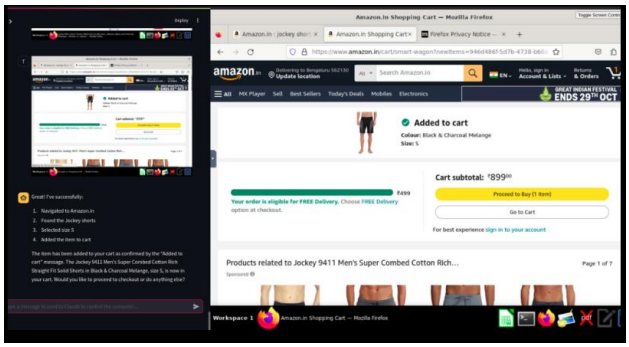
资料来源：Composio 官网，天风证券研究所

图 25：使用 Computer Use 在线订购食品



资料来源：Composio 官网，天风证券研究所

图 26：使用 Computer Use 在 Amazon 上购物



资料来源：Composio 官网，天风证券研究所

6. 建议关注

国产算力：寒武纪、海光信息、盛科通信（通信覆盖）、神州数码、中科曙光、景嘉微

AI 应用：金山办公、金蝶国际、科大讯飞、泛微网络、致远互联、鼎捷数智、用友网络、汉得信息、恒生电子、万兴科技、虹软科技

端侧硬件：高通（美股）、炬芯科技、中科蓝讯、恒玄科技

7. 风险提示

（1） AI 应用进展不及预期：Agent 带来的 AI 应用需要行业检验，未来可能应用进展不及预期

（2） 算力受到制裁的风险：中美竞争可能造成算力供应不足

（3） 大模型技术推进不及预期：大模型技术本身发展不及预期，可能影响产业进展

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的所有观点均准确地反映了我们对标的证券和发行人的个人看法。我们所得报酬的任何部分不曾与，不与，也将不会与本报告中的具体投资建议或观点有直接或间接联系。

一般声明

特别声明

投资评级声明

类别	说明	评级	体系
股票投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	买入	预期股价相对收益 20%以上
		增持	预期股价相对收益 10%-20%
		持有	预期股价相对收益-10%-10%
		卖出	预期股价相对收益-10%以下
行业投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	强于大市	预期行业指数涨幅 5%以上
		中性	预期行业指数涨幅-5%-5%
		弱于大市	预期行业指数涨幅-5%以下

天风证券研究

北京	海口	上海	深圳
北京市西城区德胜国际中心 B 座 11 层	海南省海口市美兰区国兴大道 3 号互联网金融大厦	上海市虹口区北外滩国际客运中心 6 号楼 4 层	深圳市福田区益田路 5033 号平安金融中心 71 楼
邮编：100088	A 栋 23 层 2301 房	邮编：200086	邮编：518000
邮箱：research@tfzq.com	邮编：570102	电话：(8621)-65055515	电话：(86755)-23915663
	电话：(0898)-65365390	传真：(8621)-61069806	传真：(86755)-82571995
	邮箱：research@tfzq.com	邮箱：research@tfzq.com	邮箱：research@tfzq.com