计算机行业2024年12月暨2025年度策略

AI应用方兴日盛，推理算力蓄势待发

优于大市

核心观点

AI应用：模型迭代夯实应用底座，AI应用多领域开花。24年OpenAI发布GPT-O系列模型，思维链技术提升推理能力，谷歌Gemini大模型持续迭代，12月发布Gemini 2.0 Flash，支持多模态输出，模型能力持续提升，夯实AI应用底座。同时，大模型API调用价格下降，利好AI应用厂商降本，加速AI应用在各场景渗透。目前，AI应用在医疗、金融、教育、办公、CRM等场景均有落地，看好25年AI应用蓬勃发展，有望迎来爆发元年。

AI算力：云厂商资本开支持续增长，AI应用拉动推理需求爆发。23年以来，全球人工智能快速发展，云厂商大力进行AI基础设施建设，24Q3微软、谷歌、Meta、亚马逊资本开支分别同比+78.6%、+62.2%、+36.1%、+88.3%，驱动新一轮资本开支上升周期。同时，B+C端应用逐步落地，拉动推理算力需求增长，根据《2023-2024年中国人工智能计算力发展评估报告（IDC&浪潮）》发布数据，预计24年中国推理算力占比为67.7%，同比+26.4个pct，预计未来推理算力占比将持续提升。此外，字节跳动大模型API调用量快速提升，预计其算力需求将快速增长，字节跳动相关产业链有望充分受益。

AI Agent：市场爆发前夕，有望颠覆传统工作范式。AI Agent市场处于早期阶段，根据Roots Analysis预测数据，预计24年全球AI Agent市场规模为52.9亿美金，预计2035年达到2168亿美金，对应24-35年CAGR为40.15%。未来随着AI Agent自动化程度提升，有望转向基于面向目标架构的工作范式；同时，根据Y Combinator（美国著名创业孵化器）披露数据，24年冬季入营项目中，AI Agent项目占比接近80%，处于规模化应用前夕。

量子计算：谷歌发布Willow量子芯片，市场规模快速增长。量子计算指基于量子特性（叠加、纠缠、量子干扰等）进行存储数据和执行计算，提供了指数加速，亦适用于并行计算；24年12月谷歌发布全新量子芯片Willow，首次实现低于阈值的量子纠错，产业落地初现曙光。根据IDC披露数据，23年全球量子计算及其相关市场规模为16.05亿元，预计28年增长至89.48亿美金，对应23-28年CAGR为41.01%，市场规模快速增长。

投资建议：看好AI应用及国产算力。全球大模型持续迭代，模型能力持续提升，赋能AI应用发展；大模型API调用价格持续下降，利好应用侧厂商降本，推动AI应用渗透率提升，看好25年AI应用持续落地，建议关注金山办公等。同时，AI应用的蓬勃发展，拉动推理侧算力需求提升，持续看好国产算力发展；此外，字节跳动大模型API调用量快速提升，潜在算力需求巨大，相关产业链有望受益，建议关注海光信息、浪潮信息等。

风险提示：大模型研发进展不及预期、云厂商资本开支投入不及预期、国产算力迭代及供应不及预期。

重点公司盈利预测及投资评级

公司代码	公司名称	投资评级	昨收盘（元）	总市值（百万元）	EPS 2024E	EPS 2025E	PE 2024E	PE 2025E
688111	金山办公	优于大市	292.35	135,219	3.26	4.05	89.68	72.19
688041	海光信息	优于大市	136.70	317,737	0.78	1.00	175.26	136.70
000977	浪潮信息	优于大市	50.85	74,858	1.24	1.45	41.01	35.07

资料来源：Wind、国信证券经济研究所预测

行业研究·行业月报

计算机

优于大市·维持

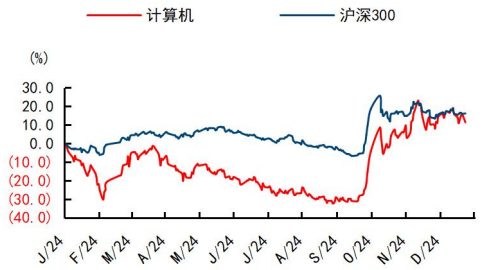
证券分析师：熊莉
021-61761067
xiongli1@guosen.com.cn
S0980519030002

证券分析师：库宏垚
021-60875168
kuhongyao@guosen.com.cn
S0980520010001

证券分析师：艾宪
0755-22941051
aixian@guosen.com.cn
S0980524090001

联系人：云梦泽
021-60933155
yunmengze@guosen.com.cn

市场走势



资料来源：Wind、国信证券经济研究所整理

相关研究报告

《人工智能专题：Openai发布会梳理》——2024-12-23
《人工智能专题：字节原动力大会总结及API调用收入预期》——2024-12-20
《FORCE2024原动力大会即将召开，关注字节AI产业链机会》——2024-12-14
《海外AI应用近况点评-AI应用逐步落地，推理算力需求提速》——2024-12-10
《美股AI软件公司情况追踪》——2024-11-17

内容目录

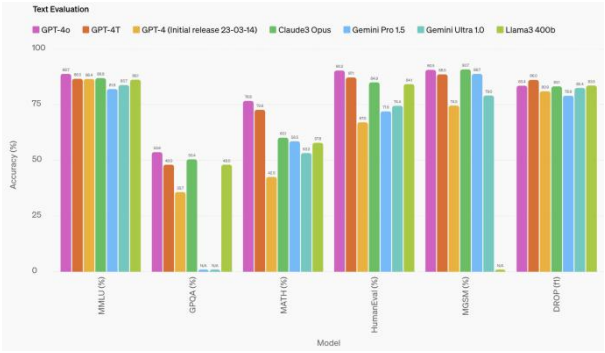
图表目录

应用：模型迭代夯实应用底座，AI 应用多领域开花

基座：大模型能力持续提升，API 调用成本下降

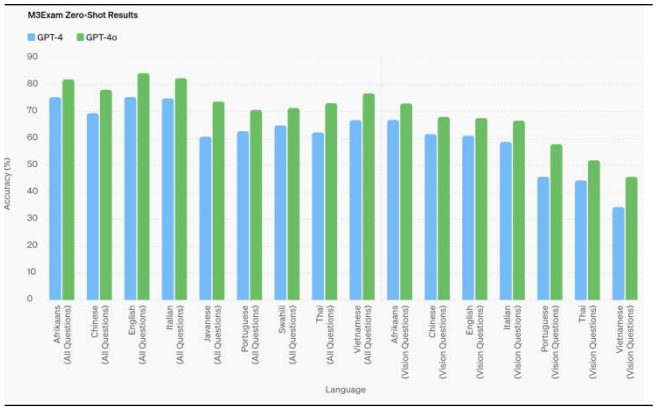
OpenAI：24 年发布 GPT-O 系列模型，思维链技术提升推理能力。1）**GPT-4O**：24 年 5 月 OpenAI 发布 GPT-4O，提供同时理解文本、音频、图像等内容的多模态能力，在 232 毫秒内能对音频输入做出反应，平均为 320 毫秒，实时性能力基本达到人类级别；同时，在基准测试维度，GPT-4O 在文本、推理和编码领域达到 GPT-4 Turbo 水平，在多语言、音频、视觉能力创下新高，此外，在 M3Exam（多语言和视觉评估基准）中，GPT-4O 全面超越 GPT-4。**2）OpenAI-O1：**24 年 9 月 OpenAI 发布 O1 模型，引入思维链技术，推理能力大幅提升，在竞赛数学、编程领域远超 GPT-4O，在物理、生物、化学等科学问题上，达到人类博士水平，此外在 MATH、MMLU 等多基准超越 GPT-4O。**3）OpenAI-O3：**24 年 12 月发布 O3 模型，在竞赛数据、科学（PHD-Level）领域进一步超越 O1 模型，同时在 ARCAGI 基准中，低算力条件下达到 75.7%、高算力条件下达到 87.5%，首次超越人类 85%的水平阈值。

图1：GPT-4o 在多测试基准超过 GPT-4 Turbo



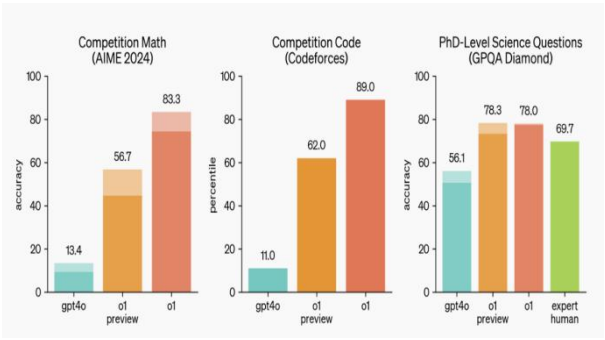
资料来源：OpenAI 官网，国信证券经济研究所整理

图2：GPT-O1 推理占比大幅提升



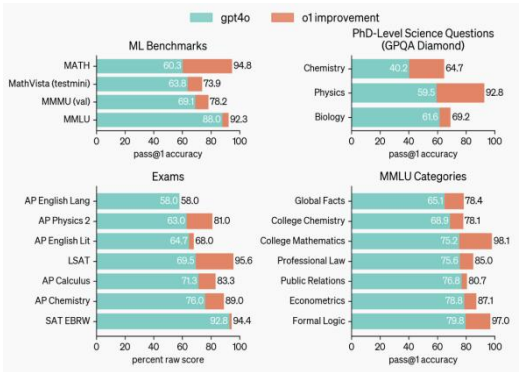
资料来源：OpenAI 官网，国信证券经济研究所整理

图3：O1 模型在竞赛数学、编程领域远超 GPT-4O



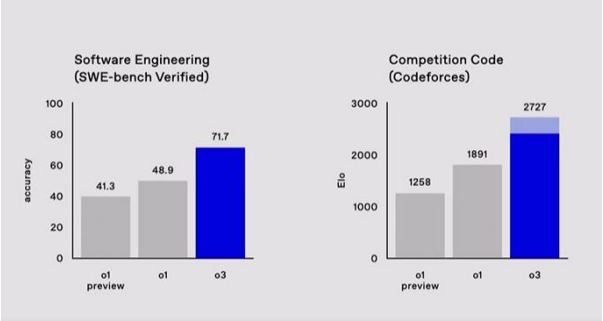
资料来源：OpenAI 官网，国信证券经济研究所整理

图4：O1 模型在 MATH、MMLU 等多基准超越 GPT-4O



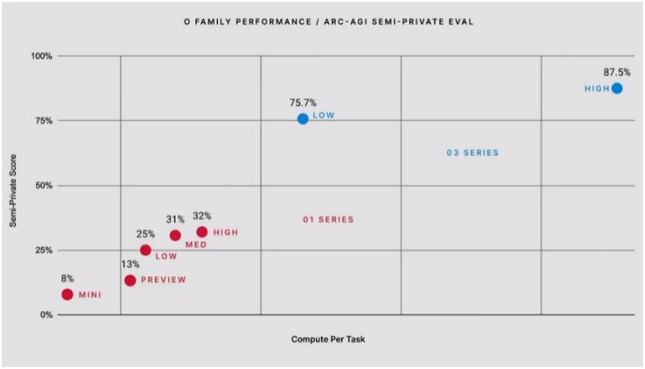
资料来源：OpenAI 官网，国信证券经济研究所整理

图5：O3 模型在竞赛数学、科学（PHD-Level）超越 O1 模型



资料来源：OpenAI 官网，国信证券经济研究所整理

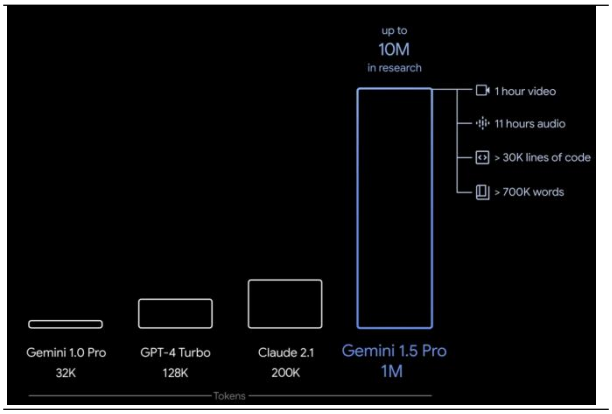
图6：高算力模式下 O3 在 ARCAGI 基准中达到 87.5%



资料来源：OpenAI 官网，国信证券经济研究所整理

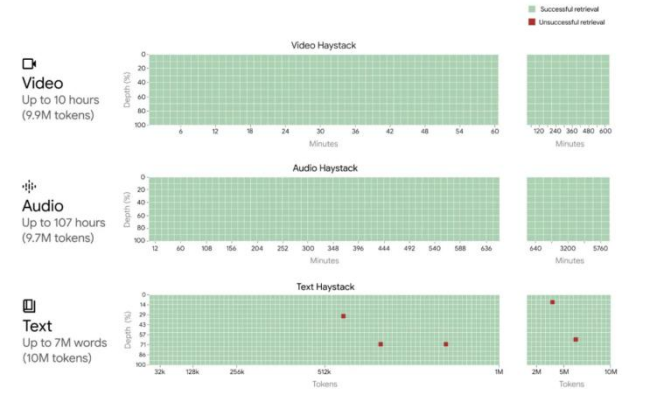
谷歌：Gemini 大模型持续迭代，超长上下文技术优势。24 年 2 月谷歌发布 Gemini 1.5，最高可支持 10,000k Token 超长上下文，在多模态海底捞针测试中，文本领域（1000 万 Token）检索准确率高达 99.2%、音频领域（11 小时）100%成功检索隐藏的音频片段、视频领域（3 小时）100%成功检索各种隐藏的视觉元素。24 年 12 月谷歌发布 Gemini 2.0 Flash（Gemini 2.0 系列模型中较小的一款），在关键基准测试中超越 Gemini 1.5pro，且速度提升一倍；同时，该模型除了支持图像、视频、音频等多模态输入外，还支持多模态输出，包括原生生成的图文混合内容等。

图7：Gemini 拥有超长上下文



资料来源：谷歌官网，国信证券经济研究所整理

图8：Gemini 1.5 在多模态大海捞针测试中表现出色



资料来源：谷歌官网，国信证券经济研究所整理

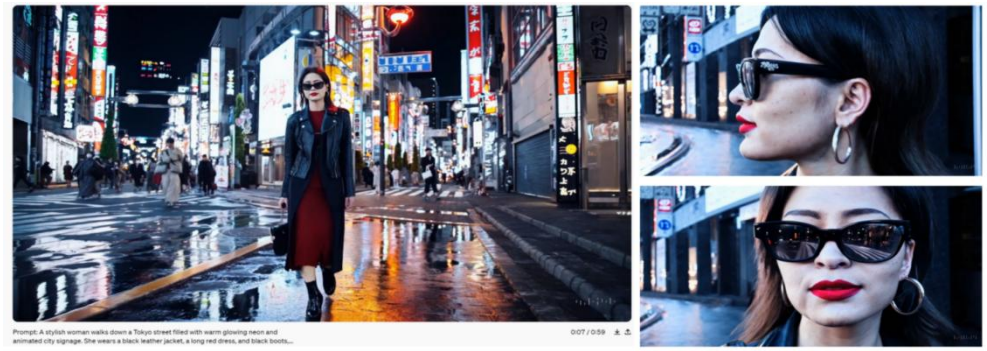
图9：Gemini 2.0 Flash 多测试基准超越 Gemini 1.5 pro

CAPABILITY	BENCHMARK	DESCRIPTION	Gemini 1.5 Flash 002	Gemini 1.5 Pro 002	Gemini 2.0 Flash Experimental
General	MMLU-Pro	Enhanced version of popular MMLU dataset with questions across multiple subjects with higher difficulty tasks	67.3%	75.8%	**76.4%**
Code	Natural2Code	Code generation across Python, Java, C++, JS, Go . Held out dataset HumanEval-like, not leaked on the web	79.8%	85.4%	**92.9%**
	Bird-SQL (Dev)	Benchmark evaluating converting natural language questions into executable SQL	45.6%	54.4%	**56.9%**
	LiveCodeBench (Code Generation)	Code generation in Python. Code Generation subset covering more recent examples: 06/01/2024 - 10/05/2024	30.0%	34.3%	**35.1%**
Factuality	FACTS Grounding	Ability to provide factuality correct responses given documents and diverse user requests. Held out internal dataset	82.9%	80.0%	**83.6%**
Math	MATH	Challenging math problems (incl. algebra, geometry, pre-calculus, and others)	77.9%	86.5%	**89.7%**
	HiddenMath	Competition-level math problems. Held out dataset AIME/AMC-like, crafted by experts and not leaked on the web	47.2%	52.0%	**63.0%**
Reasoning	GPQA (diamond)	Challenging dataset of questions written by domain experts in biology, physics, and chemistry	51.0%	59.1%	**62.1%**
Long context	MRCR (1M)	Novel, diagnostic long-context understanding evaluation	71.9%	**82.6%**	69.2%
Image	MMMU	Multi-discipline college-level multimodal understanding and reasoning problems	62.3%	65.9%	**70.7%**
	Vibe-Eval (Reka)	Visual understanding in chat models with challenging everyday examples. Evaluated with a Gemini Flash model as a rater	48.9%	53.9%	**56.3%**
Audio	CoVoST2 (21 lang)	Automatic speech translation (BLEU score)	37.4	**40.1**	39.2
Video	EgoSchema (test)	Video analysis across multiple domains	66.8%	71.2%	**71.5%**

资料来源：谷歌官网，国信证券经济研究所整理

工具类模型：24 年 2 月 OpenAI 发布文生视频 Sora，12 月正式商用。Sora 基于 Transformer 和 Diffusion 技术，可以完成超长视频生成、保持多角度视频一致性，以及通过理解物理世界提高生成视频的逼真度。24 年 5 月，谷歌发布文生图模型 Imagen 3，生成图片质量持续提升。

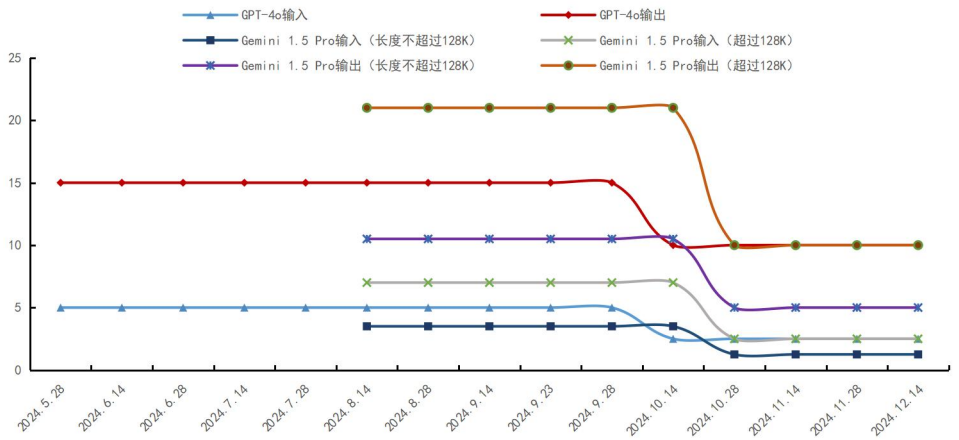
图10：Sora 可以保持多角度视频一致性



资料来源：OpenAI 官网，国信证券经济研究所整理

大模型 API 调用价格下降：根据 OpenAI 和谷歌官网 API 调用价格数据，24 年 10 月双方主力模型 API 调用价格均出现大幅下降，其中 GPT-4o 输入 API 调用价格为 2.5 美元/百万 Tokens（下降 50%），输出 API 调用价格为 10 美元/百万 Tokens（下降 33%）；谷歌 Gemini 1.5 Pro 输入 API 调用价格为 2.5 美元/百万 Tokens（下降 64%，超过 128k），Gemini 1.5 Pro 输出 API 调用价格为 10 美元/百万 Tokens（下降 52%，超过 128k），大模型 API 调用价格下降利好 AI 应用厂商成本下降，进而传导至终端 AI 应用消费者费用的下降。

图11：大模型 API 调用价格下降（单位：美元/百万 Tokens）



资料来源：OpenAI 官网、谷歌官网，国信证券经济研究所整理

场景：各场景 AI 应用蓬勃发展，有望迎来爆发元年

AI+医疗：生成式 AI 可以为医疗行业赋能，提升了医疗健康行业的智能化、高效化和便利化程度。

- **智能化**：1）智能诊断：利用大数据、人工智能等技术，智慧医疗可以根据患者就诊数据、疾病季节性变化等生成高精度诊断模型，快速分析患者症状、病历以及大模型数据，实现准确诊断疾病。2）智能治疗：通过对患者的历史就诊数据、基因数据、生活习惯等进行综合分析，智慧医疗能够为患者提供个性化的治疗方案与康复计划，持续监测康复进展，提高了患者的护理质量。3）临床支持：通过分析大量医疗数据，AI 可以为医疗专业人员提供循证医学建议，提高诊断准确性和治疗选择的合理性。

- **高效化**：1）优化问诊流程：智慧医疗能够通过智能化排班与大数据分析，合理安排医生与护士工作时间，同时智能分配患者至诊室候诊，减少等待时间过长的情况，确保患者及时就诊率。2）提高工作效率：通过智能化的病历与药品管理系统，医生可以快速检索和查看患者的病历信息，减少重复劳动，药剂师可以快速准确地配发药品，全面提高从诊断到配药的全流程效率。3）医学文献分析：大语言模型能够高效阅读与总结大量医学文献，帮助研究人员和临床医生了解最新研究进展与医学实践，减少工作量的同时确保医疗实践始终处于创新前沿。

- **便利化**：1）随时随地获取医疗服务：得益于人工智能，APP、小程序、网站等多渠道可为患者提供在线咨询、远程诊断等服务，帮助患者随时随地获取医疗服务，无需到医院排队等待，提高就医效率。2）远程医疗支持：大语言

模型可以作为健康聊天机器人的底层智能支持，不仅能够提供持续的、个性化的健康相关支持，还能够提供医疗建议、检测健康状况甚至提供心理健康支持，为用户提供全方面的医疗服务。3）医学文献分析：大语言模型能够高效阅读与总结大量医学文献，帮助研究人员和临床医生了解最新研究进展与医学实践，减少工作量的同时确保医疗实践始终处于创新前沿。

AI+金融：金融企业不断进行大模型落地实践。同花顺 24 年 1 月发布业内首个金融对话大模型——问财 Hithink，并通过网信办备案，提供 70、130、300、700、1300 亿五个版本，支持 API 调用接口、网页调用、私有化部署等；To B 端，在 iFind 中融入 AI 功能，例如智能搜索、智能投研、ChatFinD、AI 写作等，To C 端，发布智能投顾 i 问财。新致软件 24 年 6 月推出全方位人工智能能力连接平台——新致新知平台，在金融领域，公司面向银行、保险、投研三个子行业，提供全场景智能化解决方案。

图12：同花顺 HithinkGPT 提供一站式 AI 解决方案

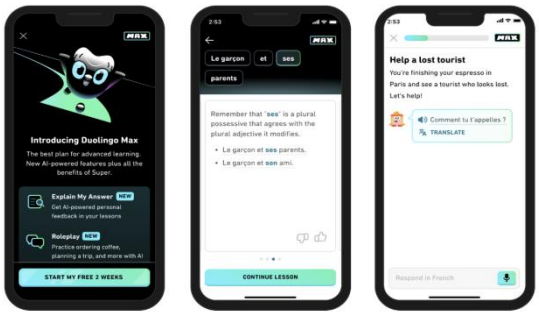
资料来源：同花顺官网，国信证券经济研究所整理

图13：新致新知拥有完整的人工智能支撑体系

资料来源：新致软件官网，国信证券经济研究所整理

AI+教育：23 年 3 月 Duolingo 推出基于 GPT-4 的 Duolingo Max，包括“Explain my answer”和“Role play”两项新功能，其中，“Explain my answer”帮助用户对每道练习题进行解析，进行错误归因和逻辑理解；“Role play”帮助用户在和 AI 对话中训练口语，实现情景式的学习体验。2024 年 9 月，Duolingo App 新增了视频通话与大冒险两项 AI 功能。佳发教育自研“灵汐文本生成大模型算法”通过网信办算法备案认证，公司 AI 能力已在体育、英语、理化生实验等多个学科的教学和考试场景中运用，并取得一定成果。

图14：Duolingo Max 包括“Explain my answer”和“Role play”两项新功能



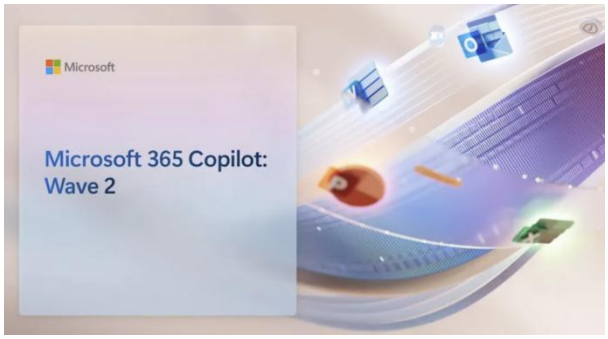
资料来源：Duolingo，国信证券经济研究所整理

图15：灵汐教育大模型实现教育行业业务、数据及应用的深度融合

资料来源：新致软件官网，国信证券经济研究所整理

AI+办公：24 年 9 月，微软发布 Microsoft 365 Copilot（Wave 2），新增 Copilot Pages 功能，将 AI 生成的短暂内容转化为持久可编辑的素材，支持团队实时协作和内容迭代，同时 Copilot in Excel、PowerPoint、Teams、Outlook、Word 等功能持续提升。根据微软披露信息，Teladoc 的客户服务代理每周通过使用 Copilot 起草对常见客户问题的回复，可节省 5 个小时的工作时间，Honeywell 的 Copilot 用户平均每周使用 Microsoft 365 Copilot 可以节省 92 分钟工作时间，提升工作效率。24 年 7 月，金山办公发布 WPS AI 2.0 版本，发布 AI 写作助手、AI 阅读助手、AI 数据助手、AI 设计助手，进一步提升办公效率。

图16：微软发布 Microsoft 365 Copilot Wave2



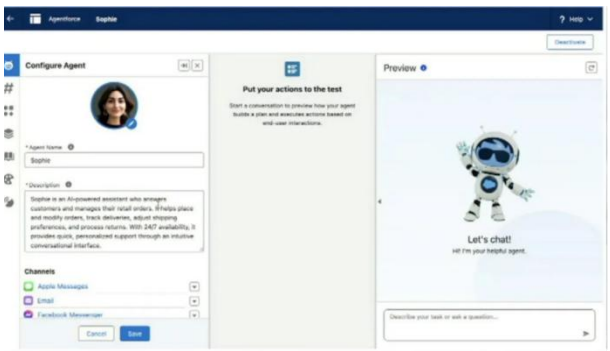
资料来源：微软官网，国信证券经济研究所整理

图17：WPS AI 发布 4 个 AI 办公助手

资料来源：金山办公官网，国信证券经济研究所整理

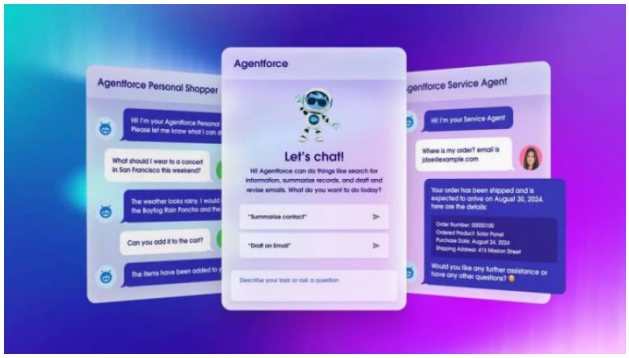
AI+CRM：以赛富时为例，24 年 10 月，公司正式推出名为 Agentforce 的 AI 代理软件服务产品，由两部分功能组成：1）Agent Builder 让用户可以通过简单的配置，轻松打造定制化的 AI 助手，并支持自定义功能等灵活扩展，满足不同业务场景需求；2）Agentforce Service Agent 是面向客户的 AI 服务助手，支持多渠道（如语音、WhatsApp、Facebook Messenger）自助服务，帮助企业快速响应客户需求。公司为 Agentforce 引入了基于使用量的定价模式，初步的产品定价约为每次 AI Agent 对话 2 美元，极大降低了客户的试用门槛，从而加速了其产品的推广。目前公司 AI 产品推广迅速，三季度公司超 100 万美元的 AI 订单数量同比增长超三倍，并签署了超 2000 个 AI 合同，其中包括超 200 个 Agentforce 订单。目前超过 80000 名系统集成商完成了 Agentforce 培训，数百家 ISV 和技术合作伙伴正在构建和销售 Agent。公司计划于当地时间 12 月 17 日发布 Agentforce 2.0，新品推出或将进一步加速公司 AI 产品放量进程。

图18：Agentforce 支持客户打造定制化 AI 助手



资料来源：赛富时官网，国信证券经济研究所整理

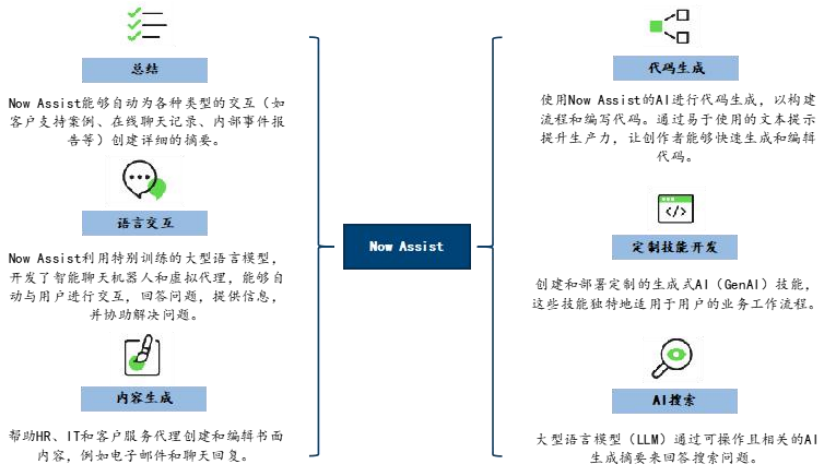
图19：Agentforce 打造全新 AI 客服



资料来源：赛富时官网，国信证券经济研究所整理

AI+自动工作流：以 ServiceNow 为例，23 年 5 月，公司宣布与英伟达达成合作伙伴关系，将共同开发企业级生成式 AI 功能；7 月，公司联手英伟达、埃森哲推出 AI 灯塔计划，旨在让下游客户快速、安全地采用生成 AI 工具；2024 年 2 月，公司与英伟达扩大合作，推出特定领域的生成式 AI 解决方案，基于 Now 平台构建，使用 NVIDIA Triton 推理服务器提供服务，并通过 NVIDIA NeMo 进行定制；10 月，公司与英伟达的合作关系进一步扩大，双方将使用 NVIDIA NIM Agent Blueprints 在 Now 平台上共同开发原生 AI 代理，为下游客户提供即开即用的解决方案。基于与英伟达的合作，公司于 2023 年 9 月推出了 Now Assist：1）Now Assist for ITSM：提供交互摘要和事件记录，使代理人能够在处理新问题时快速获取信息以解决事件，事后自动生成解决方案注释以加快结案时间； 2）Now Assist for CSM：能够快速为案例和聊天生成摘要，减少手动工作量，并使客服人员能够更快地解决客户问题，并提供个性化服务，从而提升客户满意度； 3）Now Assist for HR Service Delivery：代替 HR 团队快速为员工提供所需答案，面对薪资差异、员工信息更新等问题，HR 可以通过查看即时生成的案例主题摘要、实时聊天及之前的解决方案迅速解决各种问题；4）Now Assist for Creator：可以将自然语言文本转换为高质量的代码建议，使开发团队能够在 Now 平台上更快地创建和扩展应用程序，该工具基于 ServiceNow 工程代码进行训练，因此通过 Now Assist for Creator 生成的结果通常比其他任何代码生成技术产生的代码质量更高、更具可扩展性和安全性。2024 年 9 月，公司发布了最新的 Now 平台更新 Xanadu，该版本大幅扩展了 Now 平台的 AI 功能，新增了超 350 个开箱即用的新一代 AI 功能，覆盖数据可视化自动化、聊天和电子邮件回复生成等多个领域。

图20：Now Assist 主要功能包括总结、语音交互、内容生成、代码生成等



资料来源：Service Now 官网，国信证券经济研究所整理

算力：云厂商资本开支持续增长，AI 应用拉动推理需求爆发

需求侧：云厂商资本开支持续增长，市场增速指引乐观

复盘云计算厂商历史，全球性资本开支增长仅有两次：第一次，2010 年美国制定“云优先”的发展战略，全球云计算蓬勃发展，各厂商资本开支均快速增长；第二次，23 年以来，全球人工智能快速发展，云厂商大力进行 AI 基础设施建设，驱动新一轮资本开支上升周期。

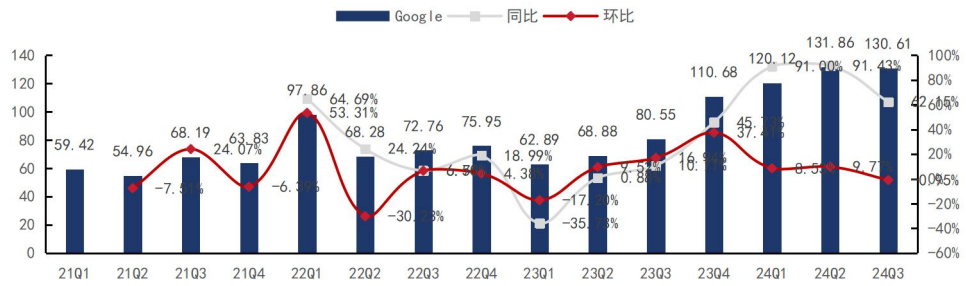
微软：24 年资本开支维持高同比增速。24Q3 资本支出（包括融资租赁）200 亿美元，同比+78.6%、环比+5.3%，其中用于购买财产、厂房和设备的现金支付为 149 亿美元，同比+50.7%。根据公司财报电话会议披露，资本开支总体投向 AI 和云，其中约一半用于长期资产，另一半用于采购服务器（包括 CPU 及 GPU）。公司预计，随着公司扩大人工智能服务规模，资本支出将继续增加。

图21：微软 FY25Q1（=24Q3）资本开支 200 亿美金，同比+78.6%、环比+5.3%

资料来源：微软财报，国信证券经济研究所整理

谷歌：24 年资本开支维持高位。谷歌 FY24Q3 资本开支为 130.61 亿美元，同比+62.15%、环比-0.95%，环比基本持平。资本开支主要用于服务器、网络设备的购买以及数据中心的建设，融资租赁成本在财务上不显著；同时，预计 Q4 资本开支将与 Q3 持平。

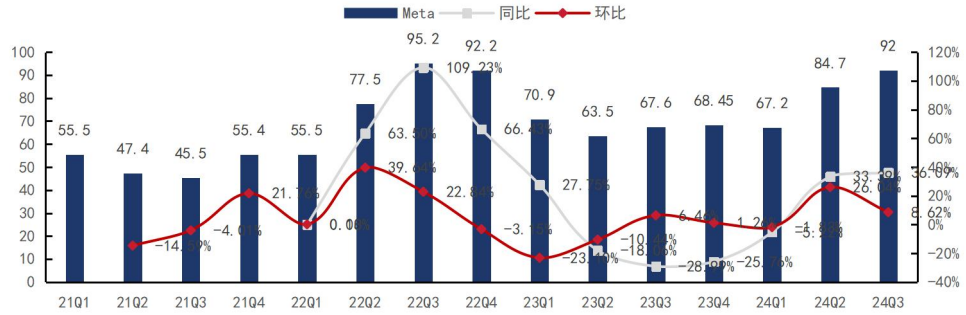
图22：谷歌 FY24Q3 资本开支 131 亿美金，同比+62.15%、环比-0.95%



资料来源：谷歌财报，国信证券经济研究所整理

Meta：上调资本开支指引下限。24Q3 资本支出（包括融资租赁支出）为 92 亿美元，同比+36.1%、环比+8.6%，主要用于服务器、数据中心和网络基础设施的投资，部分支出受三季度服务器交付进度的影响，预计将在第四季度支付。Meta 预计 2024 年资本支出将介于 380 亿美元至 400 亿美元之间，较此前预计的 370 亿美元至 400 亿美元有所上调。同时，Meta 预计 2025 年的资本支出会显著增长。

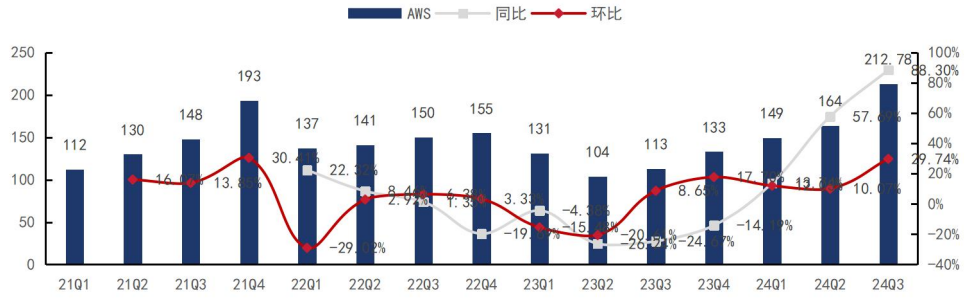
图23：Meta FY24Q3 资本开支 92 亿美金，同比+36.1%、环比+8.6%



资料来源：Meta 财报，国信证券经济研究所整理

亚马逊：资本开支连续五季度环比增长。24Q3 资本开支达 212.78 亿美元，同比+88.33%、环比+29.7%，连续五个季度环比增长，大部分支出主要投资于 AI 服务需求，同时包括支持北美和国际业务的技术基础设施。全年资本开支预计 750 亿，下个季度预计 231 亿美金。

图24：亚马逊 FY24Q3 资本开支 213 亿美金，同比+88.33%、环比+29.7%



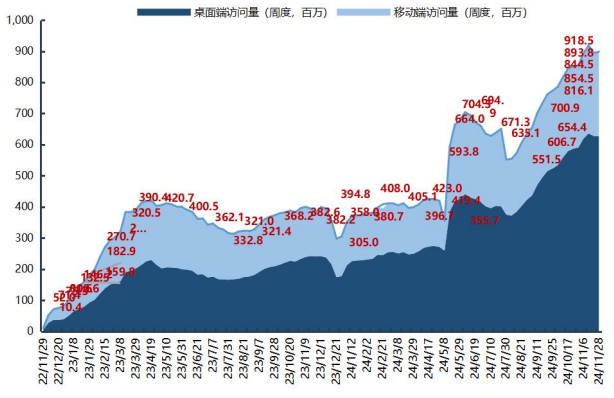
资料来源：亚马逊财报，国信证券经济研究所整理

超威半导体（AMD）全球 AI 芯片指引乐观。24 年 12 月，超微半导体（AMD）CEO 对未来 AI 芯片市场指引乐观，预计 28 年 AI 芯片市场规模将突破 5000 亿美元大关，年均增长率高达 60%。

推理侧：B+C 端应用逐步落地，思维链等新技术拉动推理算力需求增长

C 端：AI 应用数据持续增长。从全球来看，ChatGPT 周度访问量持续提升，根据 Similarweb 数据，最近一周（11 月 28 日-12 月 4 日）访问量合计为 8.99 亿次，环比+0.5%；根据 AI 产品榜数据，11 月 ChatGPT 的 MAU 数据为 287.25M，环比+11.27%。从国内来看，豆包、文小言、Kimi、智谱清言等 AI 应用快速发展，根据 AI 产品榜数据,11 月豆包、文小言、Kimi、智谱清言 MAU 分别为 59.98M、12.99M、12.82M、6.37M，分别环比+16.92%、3.33%、27.40%、22.18%。

图25：ChatGPT 周度访问量数据持续上升



资料来源：Similarweb，国信证券经济研究所整理

图26：11 月全球 TOP10 AI 产品有 7 款 MAU 环比增长

全球排名	AI产品榜	产品名	应用(APP)简短描述 aicpb.com	11月上榜应用 APP MAU	11月上榜应用 MAU变化
1		ChatGPT	The official app by OpenAI	287.25M	11.27%
2		豆包	AI 智能助手 | 抖音	59.98M	16.92%
3		Nova	聊天AI与AI写作机器人	49.63M	5.67%
4		ChatOn	Powered by ChatGPT & GPT-4o	28.84M	6.66%
5		Remini	人工智能修图	27.96M	-2.16%
6	c.ai	Character AI	Chat Ask Create	26.88M	5.74%
7		FaceApp	AI 人脸编辑器	26.48M	0.20%
8		Ask AI	Chat with Ask AI	26.35M	-7.16%
9		Talkie AI	Chat With Character | MiniMax	25.19M	22.14%
10		Chatbot AI	Chatbot AI & Smart Assistant	23.1M	3.85%

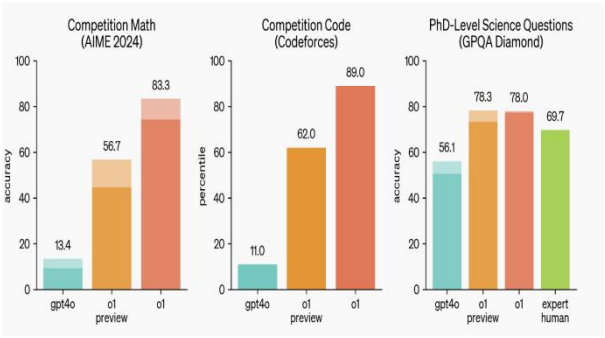
资料来源：AI 产品榜，国信证券经济研究所整理

B 端：AI 赋能公司业绩增长。在广告领域，AppLovin 在 AppDiscovery 平台使用 AXON 2.0 AI 驱动技术拉动广告商支出增长，三季度收入同比+39%、净利润同比+300%，且四季度展望乐观；在数据分析领域，Palantir 推出 AIP 平台，集成多款大模型，用于数据分析，拉动其商业客户收入增长，三季度收入同比+30%（其中美国商业业务同比+54%），且上调全年收入指引为 28.05-28.09 亿美元（前值为 27.42-27.50 亿美元）。Salesforce、DocuSign、Asana 等公司受益于 AI 驱动，三季度业绩表现出色。

AI 应用逐步落地，拉动推理算力需求增长。从单次推理来看，主要包括分词（Tokenize）、嵌入（Embedding）、位置编码（Positional Encoding）、Transformer 层、Softmax，推理主要计算量在 Transfomer 解码层，对于每个 token、每个模型参数，需要进行 2×1 Flops = 2 次浮点运算，则单词推理算力消耗为模型参数量×（提问 Tokens + 回答 Tokens）×2，随着模型参数量增长、模型向多模态发展，单次推理算力消耗持续增长。从推理次数来看，AI 应用逐步落地，模型推理次数提升，拉动推理算力需求快速增长。

OpenAI 发布 GPT-O1，通过思维链提升模型推理能力。24 年 9 月 12 日，OpenAI 发布 GPT-O1，同 GPT-4O 相比，GPT-O1 在数学、代码、科学问题（PhD 级别）评分显著提升。GPT-O1 在回复用户问题之前会生成一条较长的内部思维链，将复杂的问题拆分为更简单的步骤，且当前方法无效时，会进一步尝试其他方式，引入思维链将显著提升模型的推理能力。

图27：GPT-O1 在数学、代码、科学问题（PhD 级别）评分显著高于 GPT-4o



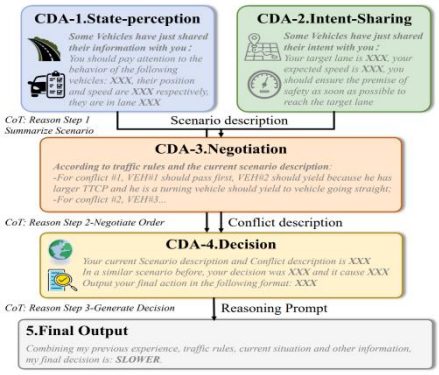
资料来源：OpenAI 官网，国信证券经济研究所整理

图28：GPT-O1 推理占比大幅提升

资料来源：JimFan（From X），国信证券经济研究所整理

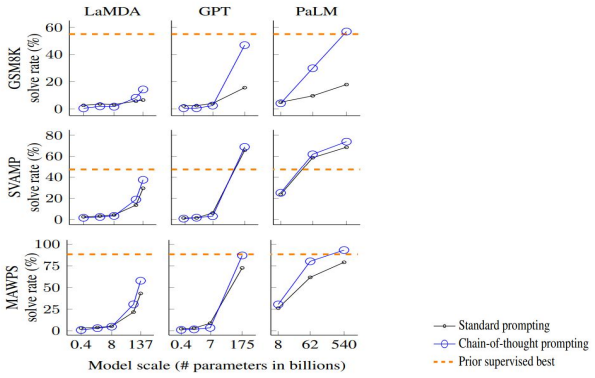
思维链（CoT）拉动推理算力增长。1）思维链（CoT）需要多步推理进而大幅提升推理算力的需求，同时推理时间的增长亦是推理算力消耗增长的反映；2）根据 Jason Wei 等人在 23 年发布的文章《Chain-of-Thought Prompting Elicits Reasoning in Large Language Models》，思维链仅对 1000 亿以上参数模型的推理有显著提升；此前，为节省推理算力消耗，大多数模型通过蒸馏等方式缩小模型参数量，而思维链反向限定模型参数量下限，进而拉动推理阶段算力需求增长。

图29：思维链多步推理提升推理阶段算力消耗



资料来源：Shiyu Fang 等著-《Towards Interactive and Learnable Cooperative Driving Automation: a Large Language Model-Driven Decision-Making Framework》-arXiv（2024）-P6，国信证券经济研究所整理

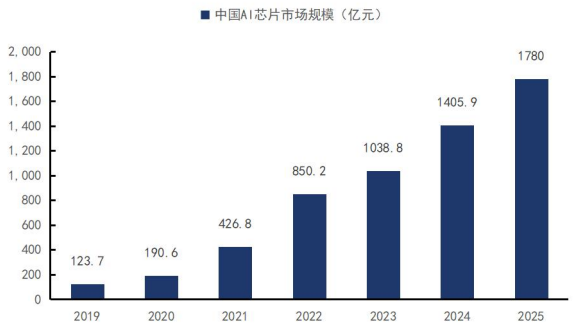
图30：思维链（CoT）在 1000 亿参数模型上才能带来显著提升



资料来源：Jason Wei 等著-《Chain-of-Thought Prompting Elicits Reasoning in Large Language Models》-arXiv（2023）-P5，国信证券经济研究所整理

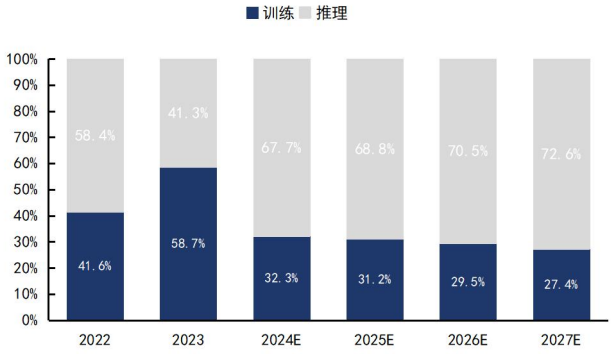
中国 AI 芯片市场规模快速增长，推理算力占比有望提升。根据亿欧智库数据，23 年中国 AI 芯片市场规模约 1038.8 亿元，预计 25 年增长至 1780 亿元，对标 23-25 年 CAGR 为 30.9%，中国 AI 芯片市场规模快速增长。随着 AI 应用逐步落地以及思维链等技术的运用，推理侧算力需求有望快速提升，根据《2023-2024 年中国人工智能计算力发展评估报告（IDC&浪潮）》发布数据，预计 24 年中国推理算力占比为 67.7%，同比+26.4 个 pct。

图31：中国 AI 芯片市场规模快速增长



资料来源：亿欧智库，国信证券经济研究所整理

图32：预计 24 年开始推理算力占比大幅提升



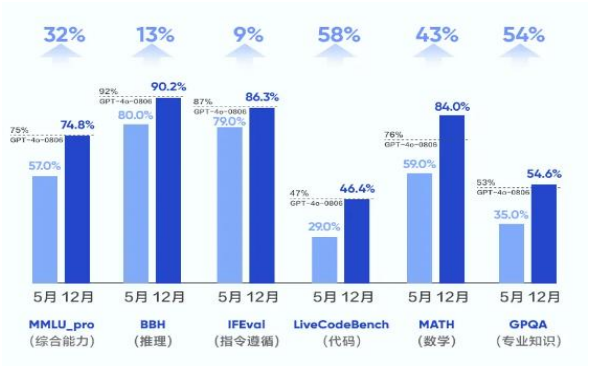
资料来源：《2023-2024 年中国人工智能计算力发展评估报告（IDC&浪潮）》-P18，国信证券经济研究所整理

- **博通**：从客户来看，公司为谷歌、Meta 等客户定制 AI Asic 芯片；从技术来看，公司可以提供高复杂度的定制加速卡（XPU）和 Asic，主要包括：1）计算：处理单元架构（客户提供）、设计流程和性能优化（博通）；2）内存：HBM PHY、集成与性能（博通）；3）网络 I/O：架构及执行（博通）；4）封装：2.5D、3D 和硅光子体系结构（博通）。
- **Marvell**：公司 19 年收购 Avera，随后宣布提供定制 Asic SOC 服务，目前客户主要有亚马逊等。
- **海光信息**：公司深耕 AI 芯片领域，采用 GPGPU 架构，其 DCU 芯片在推理领域表现出色。
- **寒武纪**：公司云、边、端三位协同，发布思元 370 加速卡，推理领域表现出色。
- **云天励飞**：公司 Deep Edge 系列推理卡已经适配了包括云天天书、通义千问、百川智能、以及 Llama2/3 等在内的近十个主流大模型。

字节跳动：大模型 API 调用量快速提升，潜在算力需求巨大

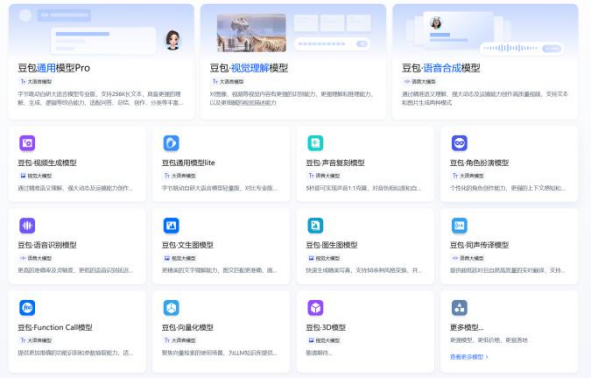
字节跳动模型能力持续提升，产品矩阵愈加丰富。1）模型能力持续提升：豆包通用模型 pro 完成新版本迭代，综合任务处理能力较 5 月份提升 32%，在推理上提升 13%，在指令遵循上提升 9%，在代码上提升 58%，在数学上提升 43%，在专业知识领域能力提升 54%。豆包通用模型 Pro 已全面对齐 GPT-4o 的能力，但使用价格远低于后者，推理输入价格为 0.0008 元/千 tokens。**2）产品矩阵愈加丰富**：除豆包通用模型 Pro 外，字节新发布视觉理解模型，模型能够综合理解用户给出的文本和图像信息，并给出准确的回答，在金融、医疗、教育、旅游等诸多行业有广阔的应用前景，且模型输入价格仅为 0.003 元/千 tokens，比行业价格便宜 85%；此外，豆包音乐模型、文生图模型持续迭代升级。

图33：字节跳动模型能力持续提升



资料来源：字节跳动官网，国信证券经济研究所整理

图34：字节跳动模型产品矩阵愈加丰富

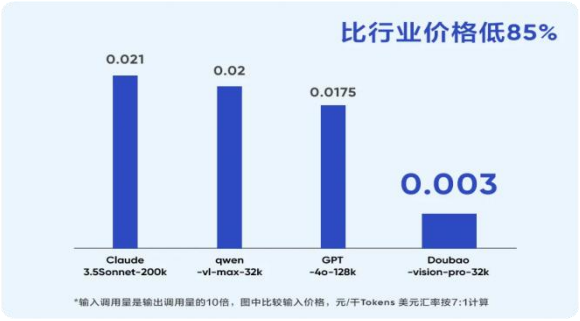


资料来源：字节跳动官网，国信证券经济研究所整理

图35：字节跳动视觉理解模型能力增强，场景拓宽

资料来源：火山引擎，国信证券经济研究所整理

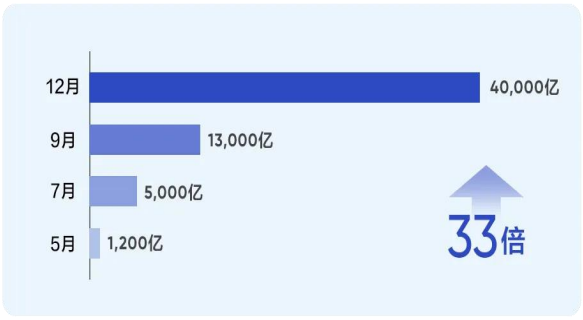
图36：字节跳动视觉理解模型价格远低于同行



资料来源：火山引擎，国信证券经济研究所整理

豆包大模型 API 调用量迅速提升，多场景快速渗透。1）API 调用量快速提升：截至 12 月中旬，豆包通用模型的日均 tokens 使用量已超过 4 万亿，较五月首次发布时日均 1200 亿增长了 33 倍；目前，豆包大模型已经与八成主流汽车品牌合作，并接入到多家手机、PC 等智能终端，覆盖终端设备约 3 亿台，来自智能终端的豆包大模型调用量在半年时间内增长 100 倍。**2）多场景快速渗透：**最近 3 个月，豆包大模型在信息处理场景的调用量增长了 39 倍，帮助企业更好的分析和处理内部及外部的数据；客服与销售场景增长 16 倍，帮助企业更好的服务客户，扩大销售；硬件终端场景增长 13 倍，AI 工具场景增长 9 倍，学习教育等场景也有大幅增长。

图37：豆包大模型 API 调用量迅速提升



资料来源：火山引擎，国信证券经济研究所整理

图38：豆包大模型在多场景迅速渗透

资料来源：火山引擎，国信证券经济研究所整理

豆包 API 调用收入快速增长，拉动推理算力需求提升。我们的计算基于以下假设：1）目前豆包 Pro 大模型能力已全方位对齐 ChatGPT-4o，我们假设豆包通用大模型的参数量已达到万亿以上，与 ChatGPT 参数量齐平，目前万亿模型普遍采用 MoE 架构，我们假设每次调用 API 时使用的模型参数为 1100 亿；2）数据中心算力调用情况在一天内可出现多次波峰及波谷，我们假设算力设备投建时，所需的算力容量为平均需求的 3 倍；3）假设数据中心中，单卡平均利用率为 45%；4）假设豆包的所有模型 API 调用平均价格为 1.5 元/百万 Tokens；5）假设未来 API 调用将有 40%为付费调用，剩余 60%为免费用户的调用；6）目前国外领先云服务商如微软、Meta 等普遍预期计算设备折旧年限为 6 年，我们假设字节跳动数据中心折旧年限同为 6 年。基于上述假设，我们预期在日均 100 万亿 Tokens 调用量时，公司年收入可达 219 亿元。

图39：豆包 API 调用算力需求及收入（基于日均 100 万亿 API 调用

假设日均API调用量为100万亿Tokens	
通用模型	数据
豆包通用模型参数量（亿）	1100
预期日均Tokens调用量（亿）	1000000
每日推理总时长（秒）	86400
平均每秒调用Tokens（亿）	11.57
假设峰值算力需求/平均需求（倍）	3
推理所需理论算力=2*模型参数量*每秒调用Tokens/100（EFLOPS）	763.89
H800单卡算力（TFLOPS）	1979
假设单卡利用率（%）	45%
实际需要卡数（万张）	**85.78**
API调用平均单价（元/百万Tokens）	1.5
API调用付费率（%）	40%
日均API调用收入（亿元）	0.6
年收入（亿元）	**219**
单张H800芯片价格（万元）	15
显卡折旧年限（年）	6
年均算力投入（亿元）	**214.44**

资料来源：火山引擎、英伟达，国信证券经济研究所整理

图40：豆包 API 调用算力需求及收入（基于日均 1000 万亿 API 调用

假设日均API调用量为1000万亿Tokens	
通用模型	数据
豆包通用模型参数量（亿）	1100
预期日均Tokens调用量（亿）	10000000
每日推理总时长（秒）	86400
平均每秒调用Tokens（亿）	115.74
假设峰值算力需求/平均需求（倍）	3
推理所需理论算力=2*模型参数量*每秒调用Tokens/100（EFLOPS）	7638.89
H800单卡算力（TFLOPS）	1979
假设单卡利用率（%）	45%
实际需要卡数（万张）	**857.77**
API调用平均单价（元/百万Tokens）	1.5
API调用付费率（%）	40%
日均API调用收入（亿元）	6
年收入（亿元）	**2190**
单张H800芯片价格（万元）	15
显卡折旧年限（年）	6
年均算力投入（亿元）	**2144.43**

资料来源：火山引擎、英伟达，国信证券经济研究所整理

AI Agent：市场爆发前夕，有望颠覆传统工作范式

AI Agent：以大语言模型为大脑驱动的系统，具备自主理解、感知、规划、记忆和使用工具能力

AI Agent 以大语言模型（LLM）为基础，对目标进行决策并执行动作的智能化应用。AI Agent 以大语言模型（LLM）为基础，根据任务目标进行任务规划和问题拆解，同时可以调用必要的组件和工具，按照既定的工作流依次执行任务。AI Agent 具备四大能力，即规划能力、工具能力、行动能力、记忆能力，并通过与人类用户交互、外界环境交互，具备灵活性、自主性等优势。

- **规划能力**：AI Agent 通过调用大语言模型（LLM）思维链能力对任务进行分解和规划，以便高效处理复杂任务；同时，通过反思和自省框架，AI Agent 任务规划能力可以持续提升。
- **工具能力**：AI Agent 可以自动调用工具，并根据规划获取的每一步任务判断是否需要调用外部工具完成，工具包括 API、软件库、硬件设备或其他服务。
- **行动能力**：AI Agent 基于规划和记忆执行具体的行动，包括同外部世界的互动等。
- **记忆能力**：包括短期记忆和长期记忆，允许 Agent 存储和检索信息，支持学习和长期知识积累，其中，短期记忆受到有限上下文窗口的限制，长期记忆涉及信息长时间保留和检索。

图41：AI Agent 架构，包括规划能力、工具能力、行动能力、记忆能力

资料来源：甲子光年，国信证券经济研究所整理

市场及所处阶段：规模化应用前夕，未来全球千亿美金市场

AI Agent 市场处于早期阶段，商业化产品落地前夕。根据人类和 AI 完成任务的比重，将 AI 产品划分为 ChatBot、Copilot、AI Agent，最早为 ChatBot 阶段，人类完成绝大部分工作，人类可以向 AI 询问意见、了解信息，但 AI 不直接处理工作；目前处于 Copilot 阶段，人类和 AI 进行协作，AI 根据人类 Prompt 完成工作初稿，人类进行修改确定，典型产品为微软 Microsoft Copilot；25 年有望进

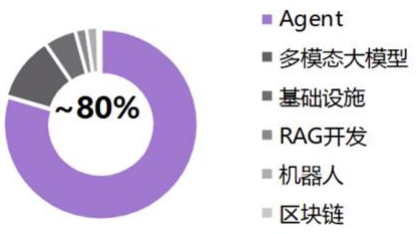
入 AI Agent 阶段，人类仅负责设定目标、提供资源和监督结果，AI 完成绝大部分工作，终极形态的 AI Agent 只需要用户起始指令和结果反馈，过程中不需要人的介入。目前，商业类 AI Agent 产品处于探索期，根据 Y Combinator 披露数据，24 年冬季入营项目中，AI Agent 项目占比接近 80%，处于规模化应用前夕。

图42：AI Agent 中人类工作占比大幅下降

名称	自动化的实现方式	含义
Chatbot	/	人类完成绝大部分工作，类似向AI询问意见，了解信息，AI提供信息和建议但不直接处理工作
Copilot	借助复杂的提示词完成自动化	人类和AI进行写作，工作量相当。AI根据人类prompt完成工作初稿，人类进行目标设定，修改调整，最后确认
Agent	通过设定目标完成自动化	AI完成绝大部分工作，人类负责设定目标、提供资源和监督结果，AI完成任务拆解，工具选择，进度控制，现目标完成后自主结束工作

资料来源：甲子光年，国信证券经济研究所整理

图43：24年Y Combinator冬季入营项目中Agent项目占比近80%



资料来源：Y Combinator、甲子光年，国信证券经济研究所整理

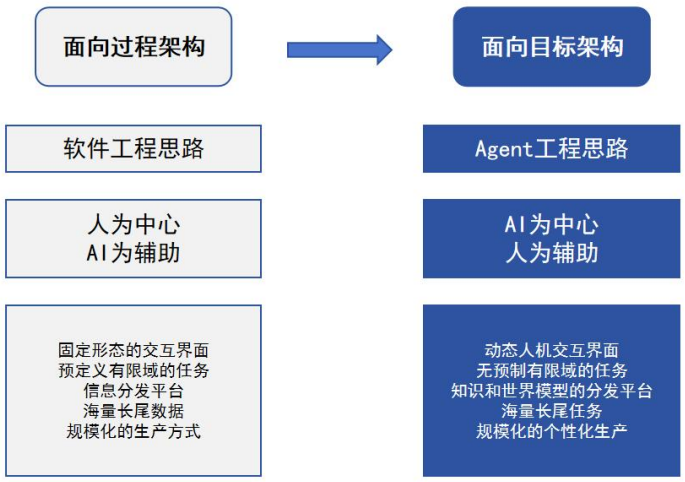
AI Agent 持续提升，有望改变传统工作范式。根据 AI Agent 自动化程度分为 L1（部分自动化）、L2（有条件自动化）、L3（高度自动化）、L4（完全自动化），感知能力、认知能力、执行能力和规划能力四个维度逐步提升。传统的工作范式是基于面向过程架构，以人为中心，AI 为辅助，有固定形态的交互界面、预定义有限域的任务、信息分发平台等；基于 AI Agent 的工作范式基于面向目标架构，以 AI 为中心，人为辅助，变更为动态人机交互界面、无预限制有限域任务、知识和世界模型的分发平台等。

表1：AI Agnet 自动化程度将持续提升

等级	感知能力	认知能力	执行能力	规划能力
L1 （部分自动化）	“所见即所得”的感知，处理单一模态下的相对简单的数据类型，应用于简单场景	利用大量人类监督信号获得一定程度的理解语言、利用语言人机交互能力	少量的常见标准工具的调用，简单的工具调用逻辑	静态地执行特定的、预定义的任务。涉及少量的、简单串并联的流程节点
L2 （有条件自动化）	多模态感知能力，能处理更广泛的数据类型，应用于更多样、更长尾、更复杂的场景	全面的认知能力，包含记忆能力、决策能力、高度智能对话能力、内容生成能力	可使用的工具数量、类型、实现的业务逻辑的复杂度得到极大提升	以业务规模达到端到端最大化自动化为目标，可以规划和编排大量流程节点和复杂逻辑
L3 （高度自动化）	综合利用认知能力，环境交互结果，在少量人类干预下获得超高精度的感知力	通过综合利用环境知识、人类少量的监督信号，达到高精度的认知水平	在人类少量干预下，可以实现绝大多数工具调用代码	能够主动洞察问题域和求解域的环境变化，实现业务流程的灵活适应和编排，环境适应能力强
L4 （完全自动化）	在无人工干预下智能体自主进化获得超高精度的感知能力	利用环境信号自主学习提升认知水平	能自动学习工具使用的方式，实现 100%的自动化调用工具的能力	能利用过程反思、经验沉淀、难例挖掘等高度智能化的决策机制，自主提升规划和编排能力，自主进化

资料来源：甲子光年，国信证券经济研究所整理和预测

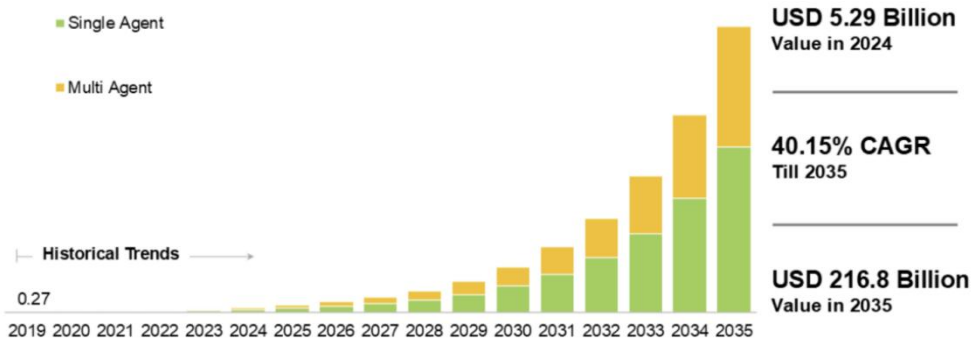
图44：Agent 有望改变传统工作范式



资料来源：甲子光年，国信证券经济研究所整理

全球 AI Agent 市场规模快速增长，国内 24 年市场有望达到百亿元。根据 Roots Analysis 预测数据，预计 24 年全球 AI Agent 市场规模为 52.9 亿美金，预计 2035 年达到 2168 亿美金，对应 24-35 年 CAGR 为 40.15%。国内 AI Agent 快速发展，根据华经产业研究院披露数据，23 年中国 AI Agent 市场规模为 59.81 亿元人民币，预计 24 年将超过百亿元，市场规模快速增长。

图45：24 年全球 AI Agent 市场规模为 52.9 亿美金，预计 35 年达 2168 亿美金



资料来源：Roots Analysis，国信证券经济研究所整理

参与者及落地场景：百花齐放，通用场景有望率先落地

根据厂商资源禀赋，AI Agent 参与者主要可以分为大模型厂商、AI Agent 应用厂商、垂类应用厂商。

- **大模型厂商：**基于自身大模型优势，提供 AI Agent 平台工具和 AI Agent 应用搭建服务，代表公司为互联网大厂（字节跳动、阿里巴巴、百度、科大讯

飞等）和大模型初创公司（智谱、Minimax 等）。

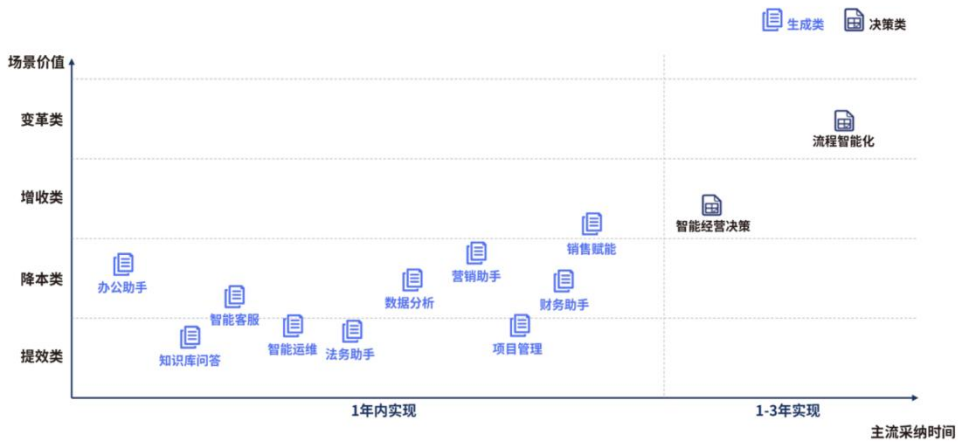
- **AI Agent 应用厂商**：AI Agent 原生应用厂商，提供 AI Agent 平台工具和 AI Agent 应用搭建，底层适配多款大模型，具体可细分为 AI Agent 工具厂商（例如 Dify）、AI Agent 方案厂商（例如澜码科技）。
- **垂直应用厂商**：通常为垂类领域中具备较深经验、客户资源积累的厂商，熟悉垂直场景下业务痛点，可以接入多家大模型产品，具体可分为无自研产品的集成方案厂商和有自研产品的应用产品厂商。

图46：中国 AI Agent 百花齐放

资料来源：InfoQ 研究中心，国信证券经济研究所整理

AI Agent 场景：通用场景有望率先落地。通用场景 AI Agent 可分为生成类应用和决策类应用，生成类应用主要包括办公助手、财务助手、营销助手等，相关场景包括办公场景（会议日程管理、会议总结、差旅申请、报销填报、员工管理、人岗匹配等）、财务场景（费用申请、业务、财税政策问答、客商政策交互服务、费用审核、客户账款催收、流程效率分析、资金洞察报告、采购成本分析报告、税利测算助手等）、营销场景（客户信息检索、智能画像总结、营销产品推荐、智能自助填单、智能话术推荐等），生成类 AI Agent 有望率先落地。决策类 AI Agent 主要包括智能经营决策、流程智能化等，有望未来 1-3 内实现。

图47：通用场景有望率先落地



资料来源：爱分析，国信证券经济研究所整理

AI Agent 场景：行业场景开始试点，能源、金融、政务领域有望率先落地。根据爱分析（ifenxi）披露数据，将行业落地分为观望学习、探索可研、试点速赢、全面推广四个阶段，目前大多数行业集中在探索可研、试点速赢节点，其中能源、金融、政务行业进度较快，主要由于行业数字化预算投入大，数字化基础设施相对完善。

图48：AI Agent 行业场景开始试点，能源、金融、政务领域有望率先落地

资料来源：爱分析（ifenxi），国信证券经济研究所整理

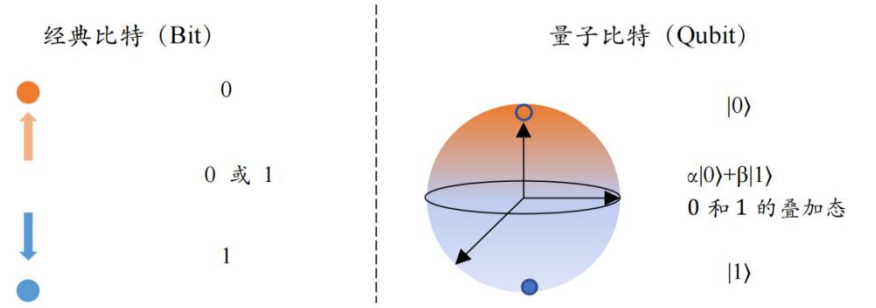
量子计算：谷歌发布 Willow 量子芯片，市场规模快速增长

量子计算：基于量子特性进行存储数据和执行计算

量子计算指基于量子特性（叠加、纠缠、量子干扰等）进行存储数据和执行计算。量子计算中的基本信息单位是量子比特（用 0、1 或 0 和 1 的叠加状态来表示），相较于经典计算机（信息基本单位是比特，用 0 或 1 二进制表示），量子计算机提供了指数加速，亦适用于并行计算。

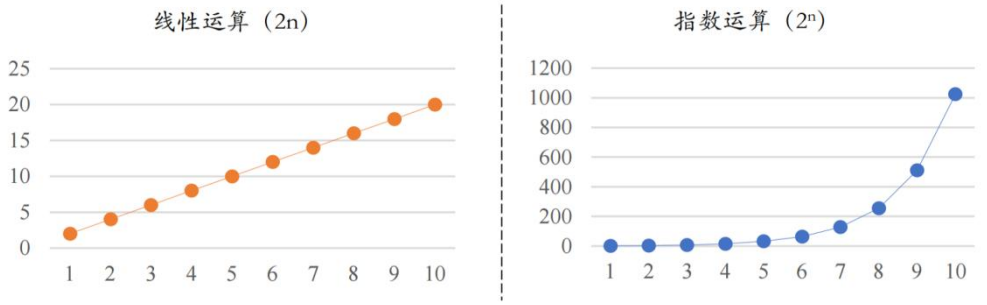
- 量子叠加：量子力学基本原理，任意两个或者多个量子态可以被加在一起，结果是另一个有效量子态。
- 量子纠缠：量子纠缠现象仅出现在量子物理学中，量子粒子可以跨越很远的距离连接并共享一个量子态，改变一个量子粒子状态会对另一个量子粒子状态产生相关影响。
- 量子干扰：量子比特固有行为，由于叠加而影响其坍塌方式的可能性。

图49：经典比特和量子比特存在差别



资料来源：光子盒研究院，国信证券经济研究所整理

图50：量子计算机提供了指数加速

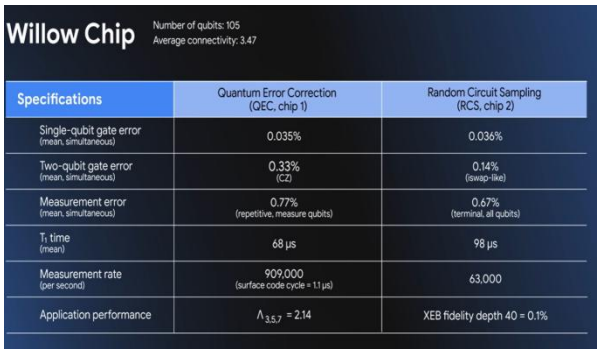


资料来源：光子盒研究院，国信证券经济研究所整理

谷歌发布 Willow 芯片：首次低于阈值的量子纠错能力

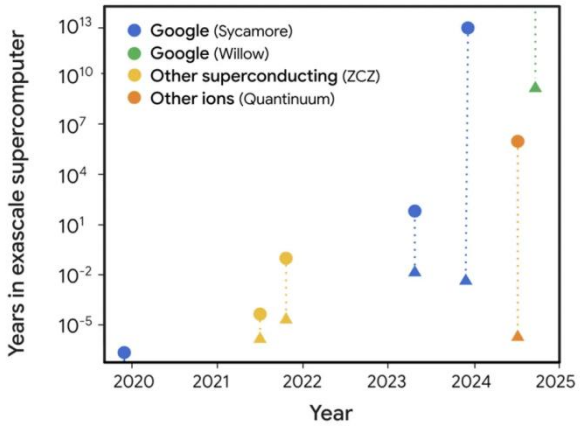
谷歌发布全新量子芯片 Willow，在量子纠错和随机电路采样两项基准测试表现出色。12 月 9 日，谷歌发布全新量子芯片 Willow，目前有 105 个量子比特，T1 时间（量子比特保持激发状态的时间）达到近 100 微秒，比上一代产品提升 5 倍，其在量子纠错和随机电路采样两项基准测试中表现出色。

图51：Willow 在量子纠错和随机电路采样测试结果出色



资料来源：谷歌官网，国信证券经济研究所整理

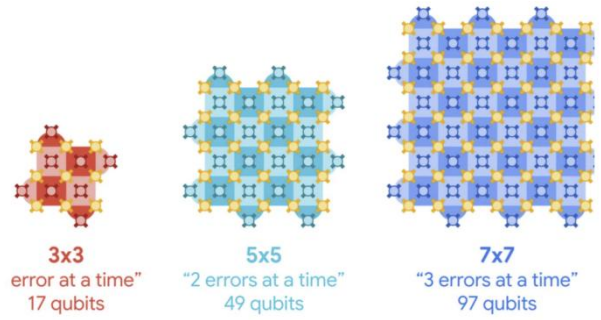
图52：Willow 超越传统计算机



资料来源：谷歌官网，国信证券经济研究所整理

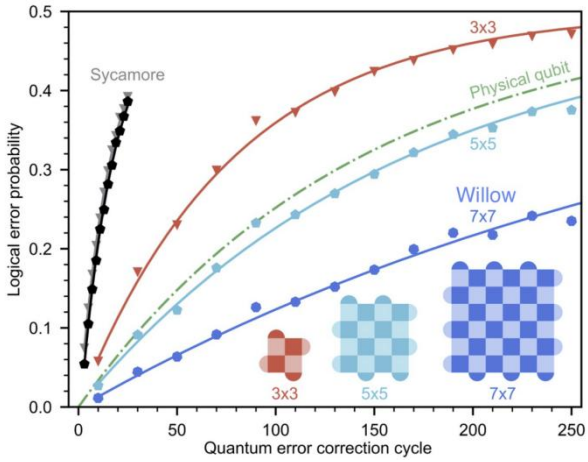
量子计算首次实现低于阈值的量子纠错。量子比特分组协同工作可以使量子计算更加可靠，通过将其分组为“表面码”（及 d×d 的量子比特网络），在通常情况下，随着晶格增大（即表面码中量子比特数量增加），额外增加的错误数目会多过纠正的错误，反而会拖慢处理器性能；而谷歌 Willow 芯片实现了在增加量子比特数量的同时，成功减少了误差，误差率指数级下降，实现了低于阈值的量子纠错（即拓展量子比特数量时，能够降低误差率）。

图53：随着晶格增大，额外增加的错误数目亦增多



资料来源：谷歌官网，国信证券经济研究所整理

图54：逻辑量子比特性能随表面码规模的扩展而提升



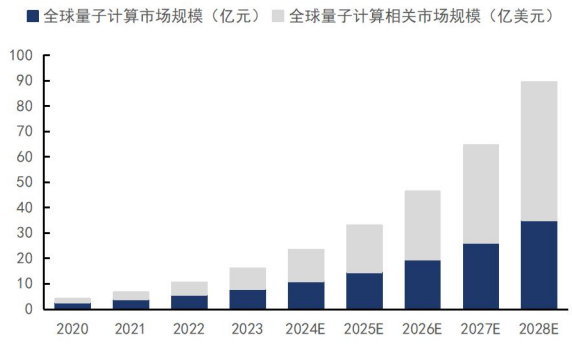
资料来源：谷歌官网，国信证券经济研究所整理

市场及技术路线：市场规模快速增长，各技术路线齐头并进

23 年全球量子计算及其相关市场规模 16.05 亿美元，市场规模快速增长。根据 IDC 披露数据，23 年全球量子计算及其相关市场规模为 16.05 亿元，其中全球量子计算市场规模 7.95 亿美金（主要包括量子计算系统、量子启发计算技术、量子计算平台等），全球量子计算相关市场规模为 8.10 亿美金（主要为基于云的量子计算相关的技术和服务、量子密码学、量子通信等）。IDC 预计 28 年全球量子计算及其相关市场规模为 89.48 亿美金，对应 23-28 年 CAGR 为 41.01%，其中全球量子计算市场规模 34.96 亿美金，全球量子计算相关市场规模为 54.52 亿美金，分别对应 23-28 年 CAGR 为 34.47%、46.44%。

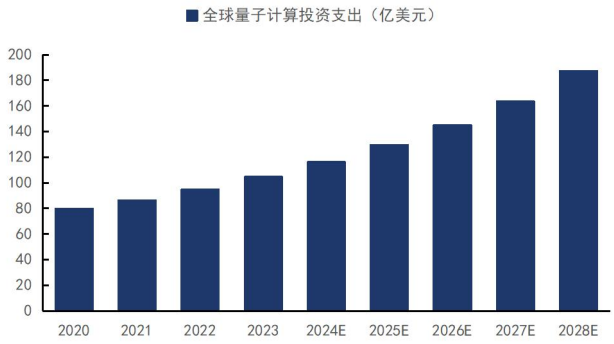
23 年全球量子计算相关投资为 104.6 亿美金，预计相关投资稳步增长。根据 IDC 披露数据，23 年全球量子计算相关投资为 104.6 亿美金，预计 28 年达到 187.3 亿美金，对应 23-28 年 CAGR 为 12.35%，量子计算相关投资稳步增长。

图55：23 年全球量子计算及相关市场规模为 16.05 亿美元



资料来源：IDC，国信证券经济研究所整理

图56：23 年全球量子计算相关投资为 104.6 亿美金



资料来源：IDC，国信证券经济研究所整理

目前，量子计算主要技术路径包括超导量子计算路线、离子阱量子计算路线、中性原子量子计算路线、光量子计算路线，各具技术优势。

- **超导量子计算路线：**使用超导体的电荷、相位、磁通量来形成量子比特，优势在于量子比特可控性强、拓展性好、可依托现有成熟集成电路工艺，但为了保障退相干时间，必须在接近绝对零度的真空环境下运行，需要强大的低温制冷系统。24 年主要进展集中在构建新型超导量子比特，以实现提高良率、增加相干时间等效果。目前，主要参与者包括 IBM、谷歌、英特尔、中科大、本源量子、国盾量子、中科院、浙江大学等。
- **离子阱量子计算路线：**量子比特存储在每个离子的稳定电子态中，量子信息可以通过离子在共享陷阱中的集体量子化运动进行传输，其具有更高的门保真度，允许量子比特与更多相邻的量子比特进行交互，且能更长时间地保持其量子态，但离子阱比特上的门操作速度较慢，这在处理系统中产生的实时误差上影响很大，24 年主要进展集中在增加量子比特数。目前，主要参与者包括 IonQ、Quantinuum、Quantum Machines、浙江大学等。
- **中性原子量子计算路线：**中性原子是指核外电子等于核内质子数的原子，量子信息被编码成非常稳定的低能原子态，具备相干时间较长、可控的相互作用、良好的扩展性和构型灵活性等优势，但其在保真度、双比特门的连通性

方面存在劣势。24 年主要进展集中在拓展量子比特数。目前，主要参与者包括 Atom Computing、QuEra、ColdQuanta 等。

- **光量子计算路线**：使用光子来编码量子比特，通过对光子的量子操控及测量来实现量子计算，具有量子比特相干时间长、操控简单、拓展性好等优势，但量子比特之间逻辑操作较为困难。24 年主要进展集中在量子纠错、量子比特控制、量子存储等。目前，主要参与者包括中科大、图灵量子、玻色量子、牛津大学等。

图57：全球量子计算机技术路线及对应量子比特数

路线	机构	型号	发布日期	量子比特数
超导	IBM	Condor	2023.12	1121
		System Two（由3块Heron组成）	2023.12	133×3
	谷歌	Sycamore 2.0	2023.07	72
	IQM	IQM Radiance	2023.11	20或54
	Rigetti	Ankaa-2	2024.01	84
	北京量子信息科学研究院	ScQ-P136（Quafu云平台）	2023.05	136（最大）590+（总和）
	量子信息与量子科技创新研究院	骁鸿	2024.04	504
	国盾量子、中电信量子	祖冲之2.1	2021.10	66计算量子比特+110耦合量子比特
	本源量子	本源悟空	2024.01	72
离子阱	Quantinuum	System Model H2-1	2023.05/2024.06	32/56
	IonQ	Forte	2022.05	36
	华翊量子	HYQ-A37	2023.07	37
中性原子	QuEra	Aquila	2022.11	256
	Atom Computing	Phoenix	2021.07	100
	Pasqal	Orion Alpha	2023	200
	中科酷原	汉原一号	2024.06	100+
光量子	中科大	九章三号	2023.10	255
	Xanadu	Borealis	2022.06	216
	玻色量子	天工量子大脑550W	2024.04	550

资料来源：光子盒，国信证券经济研究所整理

投资建议:看好 AI 应用及国产算力

全球大模型持续迭代，模型能力持续提升，赋能 AI 应用发展；大模型 API 调用价格持续下降，利好应用侧厂商降本，推动 AI 应用渗透率提升，看好 25 年 AI 应用持续落地，建议关注金山办公等。同时，AI 应用的蓬勃发展，拉动推理侧算力需求提升，持续看好国产算力发展；此外，字节跳动大模型 API 调用量快速提升，潜在算力需求巨大，相关产业链有望受益，建议关注海光信息、浪潮信息等。

风险提示

大模型研发进展不及预期。

云厂商资本开支投入不及预期。

国产算力迭代及供应不及预期。

免责声明

分析师声明

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。

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

重要声明

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 层
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