计算机行业2025年10月投资策略

视频模型迎来GPT时刻，海外大厂加码AI投资

优于大市

核心观点

阿里云栖大会召开，布局AI全栈能力。2025年9月24日，阿里云云栖大会于杭州召开。模型方面，通义大模型实现七连发，在模型智能水平、Agent工具调用和Coding能力、深度推理、多模态等方面实现多项突破。会上，阿里首次系统阐述了通往ASI的演进路线，未来三年阿里将投入超3800亿元用于建设云和AI硬件基础设施，持续升级全栈AI能力。硬件方面，阿里云在此次大会上全面展示了从底层芯片、超节点服务器、高性能网络、分布式存储、智算集群到AI平台、模型训练推理服务的全栈AI技术能力。

海外大厂加码ASIC芯片，算力需求加速。随着生成式AI和大模型进入推理时代，全球主要云计算厂商纷纷布局自研芯片，ASIC已成为云厂商在高强度推理场景下降低成本、优化性能、增强生态主导力的重要方向。据The Information报道，谷歌已经开始将其自研的AI芯片TPU部署在较小型云服务商运营的数据中心中。当前谷歌已接触包括CoreWeave和Crusoe在内的多家公司，商讨托管TPU的事宜。

算力需求飙升，全球资本开支加速。受下游算力需求推动，甲骨文FY2026财报披露其RPO飙升至4550亿美元，同比增长359%，其中仅第一季度新增即达3170亿美元，公司上修全年CapEx指引至约350亿美元。英伟达和OpenAI宣布达成合作，为OpenAI的下一代AI基础架构部署至少10吉瓦的英伟达系统。各海外大厂大幅提升资本开支，AI投入进入白热化。

Sora 2发布，视频生成模型进入GPT时刻。OpenAI发布最新的旗舰视频与音频生成模型Sora 2，在现实风格、电影风格以及动漫风格的视频生成上都表现出色，能够创造复杂的背景声效、语音和音效，并具备高度的真实感。OpenAI正式发布一款新的社交iOS应用，用户可以创作、混合彼此的生成内容，在可定制的Sora动态中发现新视频。DeepSeek宣布正式发布DeepSeek-V3.2-Exp模型，在不影响模型输出效果的前提下，可以实现长文本训练和推理效率的大幅提升。智谱AI发布GLM-4.6，在多个基准上胜过了Claude Sonnet 4/Claude Sonnet 4.5，并在寒武纪芯片上实现FP8+Int4混合量化部署。

投资建议：当前全球AI快速推进，算力投资持续加码，模型能力加速迭代，有望在未来加速商业化进程，形成产业闭环，推动算力投资持续提升，建议关注AI应用及算力相关个股，如海光信息、博通、甲骨文、新大陆等。

风险提示：AI应用落地不及预期、市场需求不及预期、行业竞争加剧、宏观经济波动、新技术研发不及预等。

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 2025E	EPS 2026E	PE 2025E	PE 2026E
688041.SH	海光信息	优于大市	252.60	587127.80	1.69	2.36	149.47	107.03
000997.SZ	新大陆	优于大市	29.56	29947.17	1.18	1.42	25.05	20.82
AVGO.O	博通	优于大市	333.39	1574389.27	5.38	7.35	61.97	45.36
ORCL.N	甲骨文	优于大市	289.01	823907.5708	6.68	8.48	43.26	34.08

资料来源：Wind、国信证券经济研究所预测

行业研究·行业投资策略

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：艾宪
0755-22941051
aixian@guosen.com.cn
S0980524090001

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

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

联系人：侯睿
hourui3@guosen.com.cn

联系人：蔺亚婴
linyaying@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

《计算机行业2025年9月投资策略暨财报总结-25H1业绩显著改善，海外大厂Capex持续上行》——2025-09-16
《人工智能行业专题：2025Q2海外大厂CapEx和ROIC总结梳理》——2025-08-15
《计算机行业2025年7月投资策略-AI ASIC市场规模快速增长，稳定币产业链蓄势待发》——2025-07-15
《计算机行业2025年6月暨中期投资策略-AI产业快速迭代，持续看好Agent和算力租赁》——2025-06-13
《稳定币香港政策落地，关注板块投资机会》——2025-06-04

内容目录

图表目录

阿里云栖大会召开，布局 AI 全栈能力

大模型七连发，重构 AI 能力底座

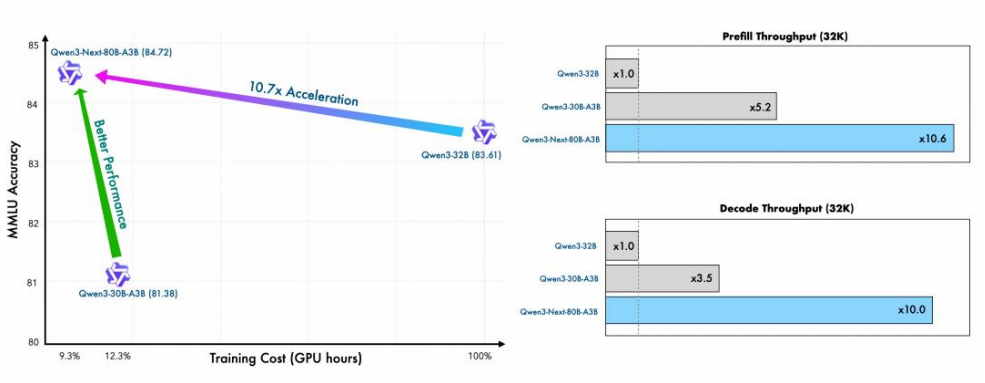
2025 年 9 月 24 日，阿里云云栖大会于杭州召开，本届大会以"云智一体·碳硅共生"为主题，全面展示了云计算与 AI 技术的最新突破。模型方面，通义大模型实现七连发，在模型智能水平、Agent 工具调用和 Coding 能力、深度推理、多模态等方面实现多项突破。其中，阿里通义旗舰模型 Qwen3-Max 全新亮相，作为通义千问家族中最大、最强的基础模型，其预训练数据量达 36T tokens，总参数超过万亿，分为指令（Instruct）和推理（Thinking）两大版本，在 Coding 编程能力和 Agent 工具调用能力方面领先。具体测评中，Qwen3-Max 在大模型用 Coding 解决真实世界问题的 SWE-Bench 评测中获得 69.6 分，位列全球第一梯队。在聚焦 Agent 工具调用能力的 Tau2 Bench 测试上，Qwen3-Max 取得 74.8 分，超过 Claude Opus4 和 DeepSeek V3.1。

图1：Qwen3-Max 能力全面领先

资料来源：公司官网，国信证券经济研究所整理

下一代基础模型架构 Qwen3-Next 及系列模型正式发布，模型总参数 80B 仅激活 3B，性能即可媲美 Qwen3 旗舰版 235B 模型，实现模型计算效率的突破。大模型目前的发展趋势是上下文长度与参数规模两方面的持续扩展，Qwen3-Next 顺应大模型的发展趋势设计，针对性地引入了多项创新：包括混合注意力机制、高稀疏度的 MoE 架构以及多 token 预测（MTP）机制等核心技术，模型训练成本较密集模型 Qwen3-32B 大降超 90%，长文本推理吞吐量提升 10 倍以上。

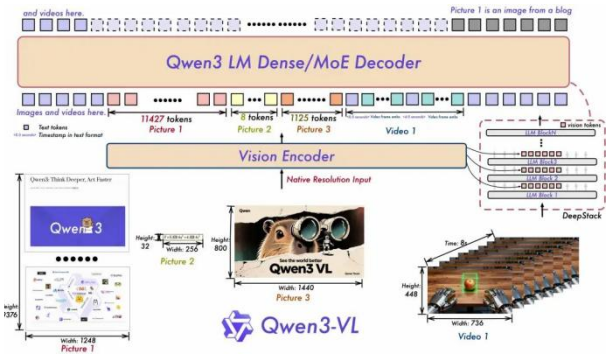
图2：Qwen3-Next 实现计算效率突破



资料来源：公司官网，国信证券经济研究所整理

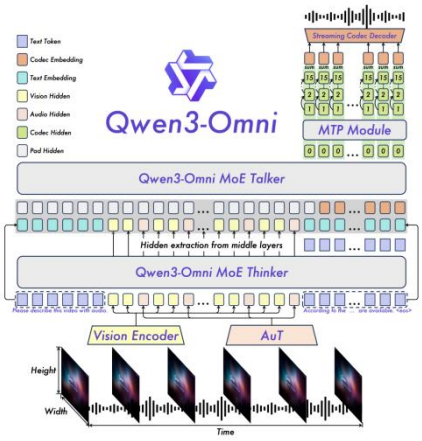
编程、多模态模型连续发布，测评结果全球顶尖。新的 Qwen3-Coder 与 Qwen Code、Claude Code 系统联合训练，应用效果显著提升，推理速度更快，代码安全性也显著提升。Qwen3-Coder 开源后调用量曾在 OpenRouter 上激增 1474%，位列全球第二。视觉理解模型 Qwen3-VL 在视觉感知和多模态推理方面实现突破，在 32 项能力测评中超过 Gemini-2.5-Pro 和 GPT-5。Qwen3-VL 拥有视觉智能体和视觉 Coding 能力，不仅能看懂图片，还能像人一样操作手机和电脑，自动完成许多日常任务。此外，Qwen3-VL 还升级了 3D Grounding（3D 检测）能力，为具身智能夯实基础；扩展支持百万 tokens 上下文，视频理解时长扩展到 2 小时以上。全模态模型 Qwen3-Omni 开源了三大版本，能够完全覆盖文本、图像、音频、视频等全模态输入，实时流式响应，实现像真人一样实时对话。

图3：Qwen3-VL 实现多模态输入统一处理



资料来源：公司官网，国信证券经济研究所整理

图4：Qwen3-Omni 全模态处理能力



资料来源：公司官网，国信证券经济研究所整理

视觉基础模型通义万相 Wan2.5-preview 系列模型推出，涵盖文生视频、图生视频、文生图和图像编辑四大模型。Wan2.5 能生成和画面匹配的人声、音效和音乐 BGM，首次实现音画同步的视频生成能力，降低电影级视频创作的门槛。视频生成

的时长达 10 秒，支持 24 帧每秒的 1080P 高清视频生成。通义万相 2.5 还全面升级了图像生成能力，可生成中英文文字和图表，支持图像编辑功能，输入一句话即可完成图像处理。同时，公司发布语音大模型通义百聆 Fun，包括语音识别大模型 Fun-ASR 和语音合成大模型 Fun-CosyVoice，可用于客服、销售、直播电商、消费电子、有声书、儿童娱乐等落地场景。

图5：通义万相多模型布局

资料来源：公司官网，国信证券经济研究所整理

此次更新，通义大模型家族完成全场景布局，形成了覆盖从 0.5B 到 480B 的全尺寸以及基础模型、编程、图像、语音、视频的全模态模型矩阵。截至目前，阿里已开源 300 余款通义大模型，全球下载量突破 6 亿次，衍生模型突破 17 万个，位居全球第一，超 100 万家客户接入了通义大模型。会上，阿里首次系统阐述了通往 ASI 的三阶段演进路线：

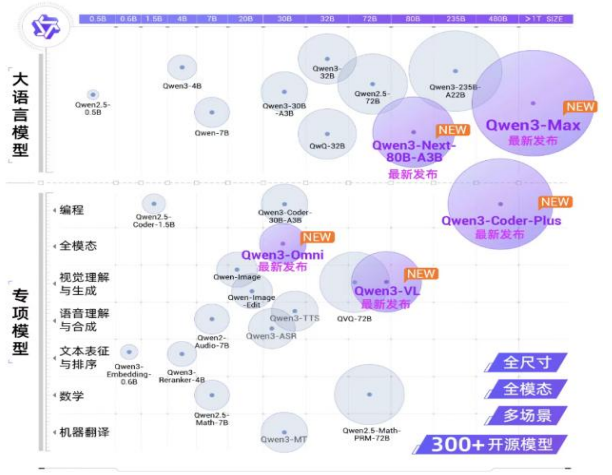
1）智能涌现：AI 通过学习海量人类知识具备泛化智能；

2）自主行动：AI 掌握工具使用和编程能力以辅助人，这是行业当前所处的阶段；

3）自我迭代：AI 通过连接物理世界并实现自学习，最终实现超越人。

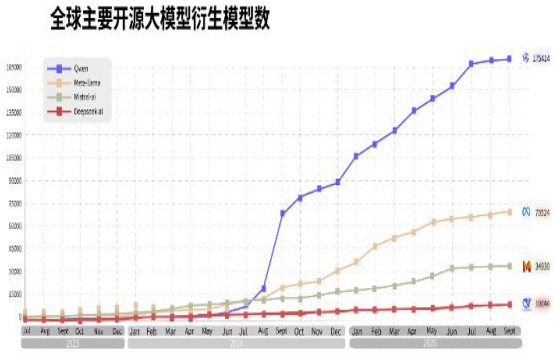
为此，未来三年阿里将投入超 3800 亿元用于建设云和 AI 硬件基础设施，持续升级全栈 AI 能力。

图6：通义千问模型家族



资料来源：公司官网，国信证券经济研究所整理

图7：通义成为最具影响力的开源模型



资料来源：公司官网，国信证券经济研究所整理

布局全栈 AI 能力，打造超级计算机

阿里云在此次大会上全面展示了从底层芯片、超节点服务器、高性能网络、分布式存储、智算集群到 AI 平台、模型训练推理服务的全栈 AI 技术能力。服务器层面，阿里云发布新一代磐久 128 超节点 AI 服务器，由阿里云自主研发设计，具备高密度、高性能的优势，可高效支持多种 AI 芯片，单柜支持 128 个 AI 计算芯片。在网络层面，新一代高性能网络 HPN8.0 采用训推一体化架构，存储网络带宽拉升至 800Gbps，GPU 互联网络带宽达到 6.4Tbps，可支持单集群 10 万卡 GPU 高效互联。在存储层面，阿里云分布式存储面向 AI 需求全面升级，高性能并行文件存储 CPFS 单客户端吞吐提升至 40GB/s，可满足 AI 训练对快速读取数据的需求；表格存储 Tablestore 为 Agent 提供高性能记忆库和知识库;对象存储 OSS 推出 Vector Bucket，为向量数据提供高性价比的海量存储，相比自建开源向量数据库，成本下降 95%。

图8：阿里云全栈 AI 能力

资料来源：公司官网，国信证券经济研究所整理

加大全球数据中心投建，与英伟达合作物理 AI 落地。阿里云已与英伟达在 Physical AI 领域的软件工具栈合作，阿里云人工智能平台 PAI 将集成 NVIDIA Isaac Sim、Isaac Lab、Cosmos、Physical AI 数据集等，形成覆盖数据预处理、仿真数据生成、模型训练评估、机器人强化学习、仿真测试在内的全链路平台支撑。这一合作将缩短具身智能、辅助驾驶等应用的开发周期，加速 Physical AI 创新落地。同时，阿里云宣布新一轮全球基础设施扩建计划，将在巴西、法国和荷兰首次设立云计算地域节点，并将扩建墨西哥、日本、韩国、马来西亚和迪拜的数据中心，以便服务全球客户日益增长的 AI 和云计算需求。

图9：阿里云全球布局

资料来源：公司官网，国信证券经济研究所整理

海外大厂加码 ASIC 芯片，算力需求加速

ASIC 芯片具备多种优势，市场空间广阔

AI 芯片指专门用于运行人工智能算法且做了优化设计的芯片，为满足不同场景下的人工智能应用需求，AI 芯片逐渐表现出专用性、多样性的特点。根据设计需求，AI 芯片主要分为中央处理器（CPU）、图形处理器（GPU）、现场可编程逻辑门阵列（FPGA）、专用集成电路（ASIC）等，相比于其他 AI 芯片，ASIC 具有性能高、体积小、功率低等特点。2016 年，Google 发布 TPU 芯片（ASIC 类），ASIC 克服了 GPU 价格昂贵、功耗高的缺点，开始逐步应用于 AI 领域，成为 AI 芯片的重要分支。

图10：不同类型 AI 芯片对比

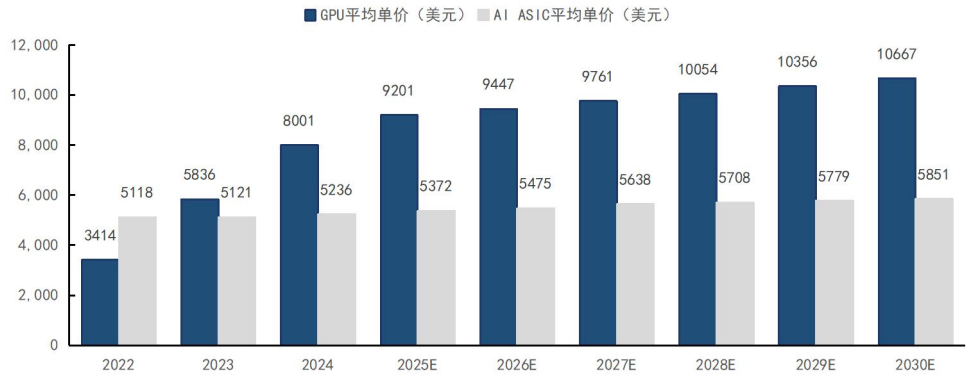
类别	CPU	GPU	FPGA	ASIC
特点	拥有大量的缓存和复杂的逻辑控制单元	一种由大量运算单元组成的大规模并行计算架构芯片	可对其集成的基本门电路和存储器进行重新定义	全定制化芯片，其无法通过修改电路进行功能拓展
功耗	高	高	中	低
优势	✓ 灵活性 ✓ 通用性强 ✓ 复杂指令和任务 ✓ 系统管理	✓ 大量并行核 ✓ AI处理出色表现	✓ 可配置的逻辑门 ✓ 灵活性 ✓ 可重新编程性	✓ 可用库设计的定制化逻辑 ✓ 更快的处理速度 ✓ 体积小
劣势	✓ 核数少 ✓ 时延严重 ✓ 效率低	✓ 功耗高 ✓ 体积大	✓ 编程复杂	✓ 固定的功能 ✓ 前期定制化成本高
代表厂商	Intel、AMD	NVIDIA、AMD	Xilinx、Altera	Google、寒武纪
	 Intel Sapphire Rapids	 NVIDIA H100	 Xilinx Versal AI Core	 Google TPU

资料来源：Ashutosh Mishra 等著-《Artificial Intelligence and Hardware Accelerators》-2023 年 Springer 出版-P35，国信证券经济研究所整理

相比较 GPU 等其他 AI 芯片，ASIC 芯片具有以下优势，使其在 AI 发展中重要性快速提升：

1）成本优势：由于 GPU 芯片的通用性、灵活性，其设计及流片成本较高，进而导致 GPU 的平均单价较高。从历史趋势来看，根据 IDC 统计数据，2022-2024 年受 AI 大模型驱动，GPU 性能需求快速提升，导致 GPU 产品的平均单价快速提升（对应 CAGR 为 53.1%）。从短期来看，2024 年 GPU 平均单价为 8001 美元，根据爱集微披露数据，英伟达 B200 单卡售价约 3-4 万美金，AI ASIC 平均单价为 5236 美元，具备价格优势。从长期来看，IDC 预测 GPU 平均单价自 2025 年后稳中有升，ASIC 平均单价基本维稳；

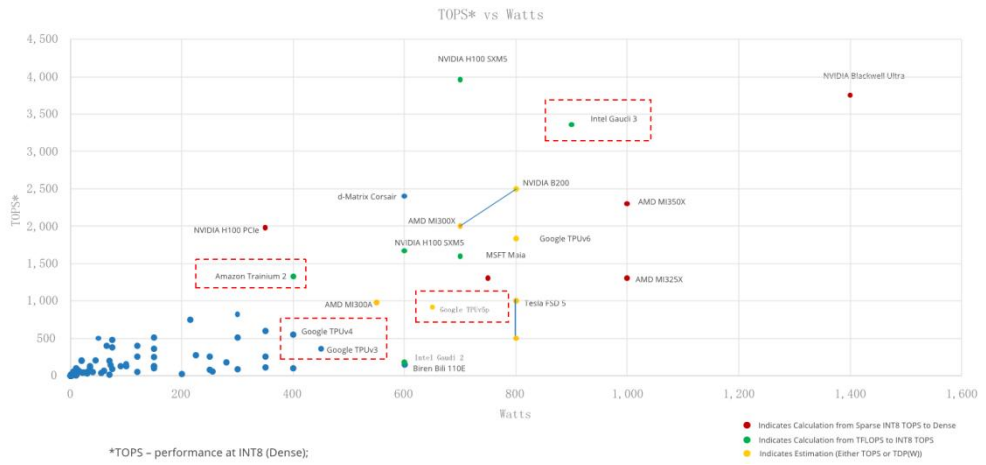
图11：GPU 和 AI ASIC 平均单价及预测



资料来源：IDC，国信证券经济研究所整理

2）能耗优势：由于 AISC 芯片偏定制化设计，专为特定任务优化，因此其在执行特定任务时功率较低。据 IDC 统计数据，在同等算力水平下，ASIC 的功率更低，能耗优势明显；

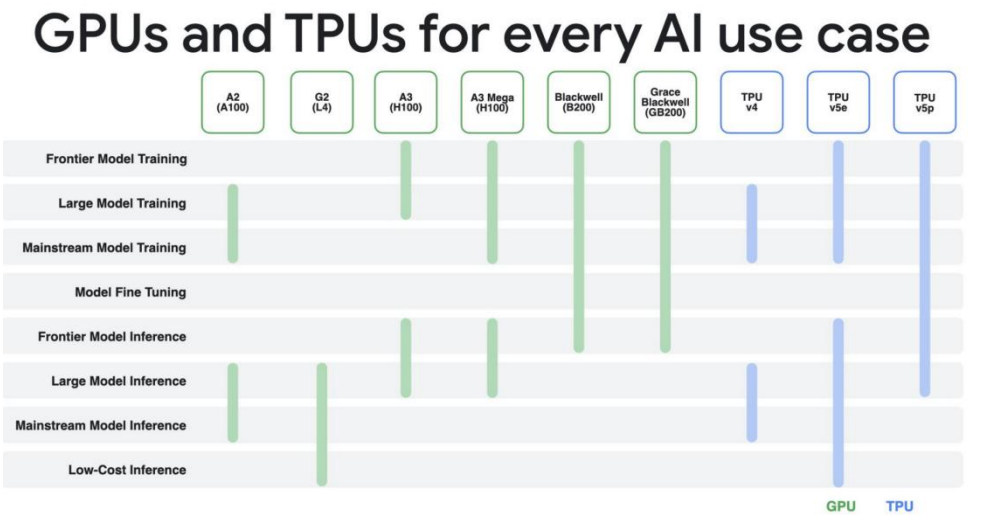
图12：AI 芯片算力和功率矩阵图



资料来源：IDC，国信证券经济研究所整理

3）推理优势：大厂对 ASIC 产品细分，处理训练、推理任务更加灵活。以谷歌 TPU 产品为例，自第五代 TPU 产品开始，产品进一步细分为“e”系列和“p”系列两个版本，TPU v5e 强调成本效益“cost-efficient”和可拓展性，TPU v5p 专注于超大基础模型训练，AI 模型训练速度、性价比表现出色。目前，TPU v5e 和 TPU v5p 已经基本覆盖全部 AI 任务，TPU v5e 在推理侧具备优势，已覆盖低成本推理、主流模型推理、大模型推理等任务。

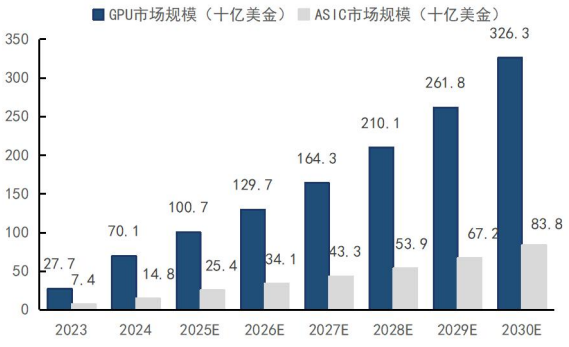
图13：AI ASIC（以 TPU 为例）基本覆盖全部 AI 任务



资料来源：Google Cloud，国信证券经济研究所整理

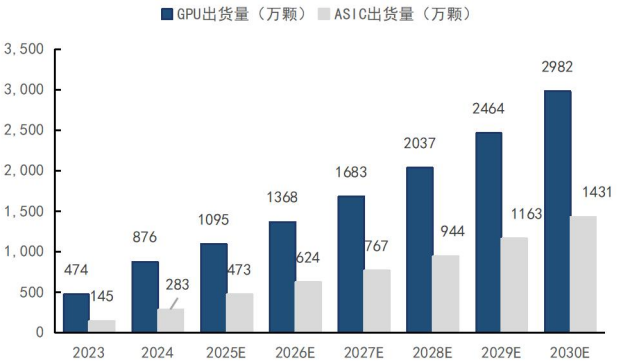
全球 AI 芯片市场超 4000 亿，ASIC 芯片快速增长。从市场规模来看，根据 IDC 披露数据，2024 年 GPU、AI ASIC 芯片市场规模分别为 701、148 亿美金，预计 2030 年分别增长至 3263、838 亿美金，对应 2024-2030 年 CAGR 分别为 29.2%、33.5%。从出货量来看，2024 年 GPU、AI ASIC 芯片出货量分别为 876、283 万颗，预计 2030 年增长至 2982、1431 万颗，对应 2024-2030 年 CAGR 分别为 22.6%、31.0%，ASIC 芯片占比稳步提升。

图14：GPU、ASIC 芯片市场规模情况



资料来源：IDC，国信证券经济研究所整理

图15：GPU、ASIC 芯片出货量情况



资料来源：IDC，国信证券经济研究所整理

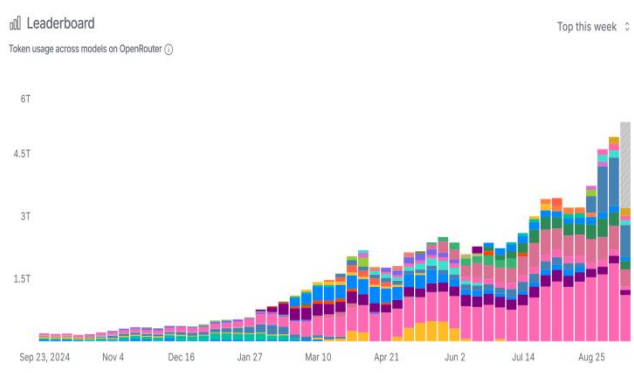
大厂布局自研芯片，谷歌授权 TPU 托管

大模型 token 处理量快速提升，推理算力需求加速。1）从龙头应用来看：据谷歌 2025 年 4 月 I/O 大会披露数据，当月 Gemini 大模型每月 token 处理量达 480 万亿，同比提升了约 50 倍；此外，根据谷歌 2025 年 7 月财报电话会议披露数据，Gemini 大模型每月处理的 token 量从 4 月至 7 月翻了一倍，达到接近 960 万亿 token/月；据 Intelligence by Intent 预测内容，预计 2025 年 8 月 Gemini 大模型每月处理的 token 量将达到 1250 万亿。2）从第三方平台来看：选取 AI 模型 API 聚合平台 OpenRouter 为例，平台集成 Grok Code Fast1、Claude Sonnet 4、Deepseek v3.1 等优秀模型，2025 年 9 月 2 日-9 月 8 日该平台周度 token 处理量为 4.9T，周度环比+6.5%，随着模型 token 使用量的快速增长，推理算力需求加速。

图16：谷歌 Gemini 大模型每月 Token 处理量

资料来源：Intelligence by Intent，国信证券经济研究所整理

图17：OpenRouter 中相关模型 Token 调用量



资料来源：OpenRouter，国信证券经济研究所整理

随着生成式 AI 和大模型进入推理时代，全球主要云计算厂商纷纷布局自研芯片，ASIC 已成为云厂商在高强度推理场景下降低成本、优化性能、增强生态主导力的重要方向：

1）谷歌：全球 ASIC 领导者，第七代 AI ASIC 芯片有望 2026 年放量。谷歌 2015 年发布第一代 AI ASIC 芯片 TPUv1，分别于 2018、2020、2022、2023、2024 年发布 TPUv2、TPUv3、TPUv4、TPUv5（包括 v5e 和 v5p）、TPUv6；2025 年谷歌发布全新一代 AI ASIC 芯片 Ironwood（TPU v7），单芯片峰值 Flops 达 4614 TFLOPS，约为 TPU v5p 的 10 倍，峰值能效是上一代 Trillium 的 2 倍，是 TPU v2 的 29.3 倍，有望 2026 年大规模放量；

2）Meta：第二代 AI ASIC 芯片有望 2025 年底量产。2023 年 Meta 发布第一代 AI 推理加速卡 MTIA v1，其从 2020 年开始设计，采用台积电 7nm 制程工艺，运行频率 800MHz，TDP 仅 25W，FP16 下 51.2 TFLOPS 算力；2024 年 Meta 发布下一代 MTIA 芯片 MTIA v2，采用台积电 5nm 工艺，运行频率 1.35GHz，TDP 为 90W，FP16 下提供 354TFLOPS 算力，有望 2025 年底大规模量产；

3）亚马逊：第三代 AI ASIC 芯片有望 2025 年底量产，更高比例资本开支用于自研芯片。2022 年亚马逊发布第一代自研 AI 芯片 Trainium 1，提供一定并行计算的能力，但由于其互联网络性能有限（仅 4 个 Neuronlink-v2 互联端口），并未大规模使用；2023 年亚马逊发布第二代 AI 芯片 Trainium 2，每颗芯片包含 8 个 NeuronCore-v3 和 4 个 HBM3 堆栈，且通过 Neuronlink-v3 网络实现芯片间高速通信，FP16 下提供 667 TFLOPS 算力，相较于 Trainium 1 提升 3.4 倍，目前已经量产。2024 年亚马逊发布第三代 AI 芯片 Trainium 3，采用 3nm 制程生产工艺，相较于上一代 AI 芯片，计算能力提升 2 倍，能源效率提升 40%，有望 2025 年底大规模量产。根据亚马逊 FY2025Q2 财报电话会披露信息，目前 Trainium 2 已经大规模用于 Anthropic 的 Claude 大模型的训练，相比于 GPU 芯片，性价比提升 30%-40%，且亚马逊会将提升用于自研 ASIC 的资本开支比例；

图18：Meta MTIA v2 芯片架构

资料来源：Meta，国信证券经济研究所整理

图19：亚马逊发布 Trainium 3

资料来源：AWS，国信证券经济研究所整理

谷歌授权自研芯片托管，扩大 TPU 使用范围。据 The Information 报道，谷歌已经开始将其自研的 AI 芯片 TPU 部署在较小型云服务商运营的数据中心中。当前谷歌已接触包括 CoreWeave 和 Crusoe 在内的多家公司，商讨托管 TPU 的事宜。目前，谷歌已与总部位于伦敦的 Fluidstack 达成协议，后者将在纽约新建的数据中心中安装谷歌的芯片。在此之前，TPU 仅用于谷歌自家的服务（例如 Gemini AI 模型），

或通过 Google Cloud 有选择性地提供给一些公司（如 Apple 和图像生成器 Midjourney）。通过允许第三方服务商托管 TPU，谷歌正在扩大其芯片的使用范围，并降低对英伟达的依赖。

表1：AI ASIC 芯片与英伟达 GPU 对比

指标	Trainium2（Trn2-16 芯片）	Trainium2-Ultra（64 芯片）	TPUv6e	GB200	H100
理论 BF16 稠密算力（TFLOP/s/芯片）	667	667	918	2500	989
理论 FP8 稠密算力（TFLOP/s/芯片）	1299	1299	1836	5000	1978
HBM 容量（GB/芯片）	96	96	32	192	80
HBM 带宽（GB/s/芯片）	2900	2900	1640	8000	3350
算术密度（BF16 FLOP/Byte）	230	230.9	559.8	312.5	295.2
每字节 HBM FLOP（BF16 FLOP/Byte）	6947.9	6947.9	28687.5	13020.8	12362.5
稀疏性支持（Structured Sparsity）	4:16，4:12，4:8，2:8，2:4，1:4，1:2	同左	无	2:04	2:04
芯片功耗（瓦特）	500	500	600	1200	700
散热技术	风冷	风冷	风冷	直连液冷（DLC）	主要风冷 / 部分液冷
Scale Up 技术	NeuronLink v3（PCIe Gen5.0）	NeuronLink v3（PCIe Gen5.0）	TPU ICI	NVLink 5	NVLink 4
Scale Up 带宽（GB/s/芯片，单向）	512	640	448	900	450
Scale Up 世界规模（芯片数）	16	64	256	72	8
Scale Up 拓扑结构	4×4 2D 环形（对称带宽）	4×4×4 3D 环形（Z 轴带宽减半）	16×16 2D 环形（对称带宽）	18 通道（NVLink SHARP 网络）	4 通道（NVLink SHARP 网络）
每芯片邻居数	4	6	4	任意到任意	任意到任意
每个 Scale Up 世界的总 HBM 容量	1536 GB	6144 GB	8192 GB	13824 GB	640 GB
每个 Scale Up 世界的物理服务器数	1	4	32	18	1

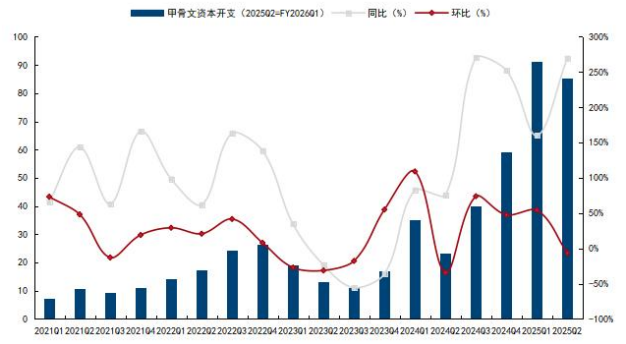
资料来源：SemiAnalysis，国信证券经济研究所整理

算力需求飙升，全球资本开支加速

甲骨文业绩超预期，OpenAI 与英伟达合作加大 AI 投资

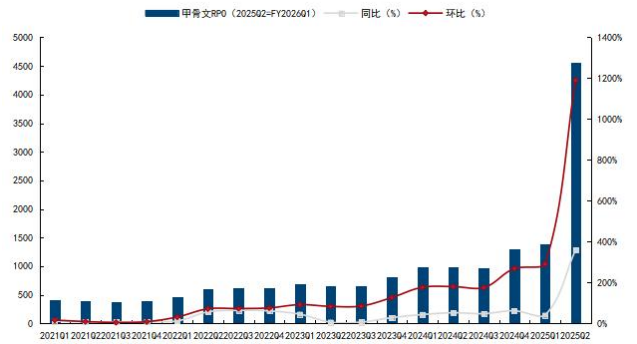
星际之门项目推进 AI 基建，推动甲骨文业绩超预期。2025 年 1 月，OpenAI 宣布启动星际之门项目，打算在未来四年内投资 5000 亿美元，为 OpenAI 在美国建设新一代 AI 基础设施，Arm、微软、英伟达、甲骨文和 OpenAI 是最初的主要技术合作伙伴。2025 年 7 月，甲骨文与 OpenAI 已达成协议，将在美国增加 4.5GW 的星际之门数据中心容量，此次与甲骨文的额外合作将使星际之门人工智能数据中心容量超过 5GW，运行超 200 万个芯片。受下游算力需求推动，甲骨文 FY2026 财报披露其剩余履约义务（RPO）飙升至 4550 亿美元，同比增长 359%，其中仅第一季度新增即达 3170 亿美元。公司上修全年 CapEx 指引至约 350 亿美元，且预计公司云基础设施（OCI）2026 财年将增长 77%，并在未来四年内持续高速增长，对应 2026-2029 财年分别为 320/730/1140/1440 亿美元，超市场预期。

图20：甲骨文资本开支高速增长（单位：亿美元）



资料来源：公司财报，国信证券经济研究所整理

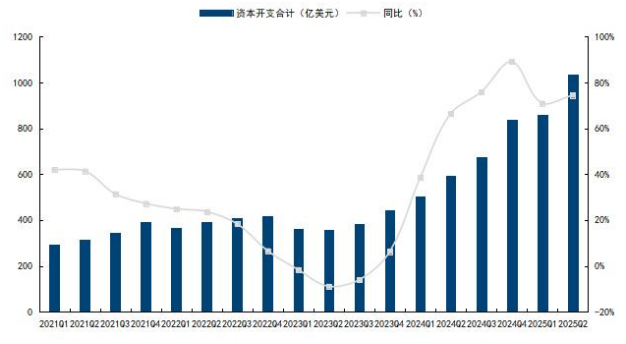
图21：甲骨文 RPO 高速增长（单位：亿美元）



资料来源：公司财报，国信证券经济研究所整理

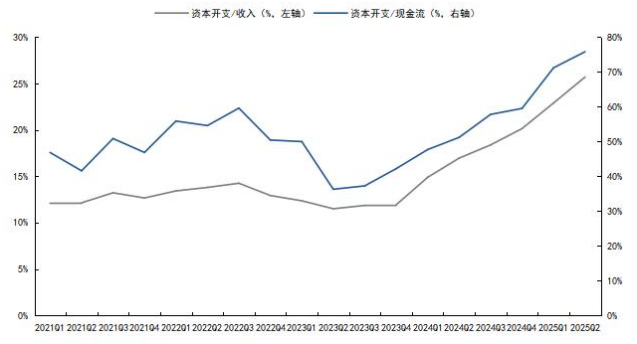
英伟达与 OpenAI 合作，加大算力投入。2025 年 9 月，英伟达和 OpenAI 宣布达成合作，为 OpenAI 的下一代 AI 基础架构部署至少 10 吉瓦的英伟达系统，用于训练和运行其下一代模型，从而实现超级 AI 部署。同时，英伟达计划在新系统部署期间向 OpenAI 投资高达 1000 亿美元，第一阶段预计将于 2026 年下半年上线。根据协议，OpenAI 还将与英伟达合作，成为其 AI 工厂增长计划的首选战略计算和网络合作伙伴，双方将携手优化 OpenAI 模型和基础架构软件以及英伟达硬件和软件的路线图。当前全球资本开支竞赛进入白热化，大厂不断加码硬件端投入。2025Q2，全球大厂（微软、谷歌、亚马逊、Meta 以及甲骨文）合计资本开支达 1035.58 亿美元，同比增长 74.47%，占总收入比重达历史最高的 25.7%，占经营性现金流比重达 75.81%。我们认为，OpenAI 与英伟达的合作将进一步推动全球 AI 投资，相关产业将持续受益。

图22：大厂资本开支合计高速增长



资料来源：公司财报，国信证券经济研究所整理

图23：资本开支占收入、现金流比重达历史峰值



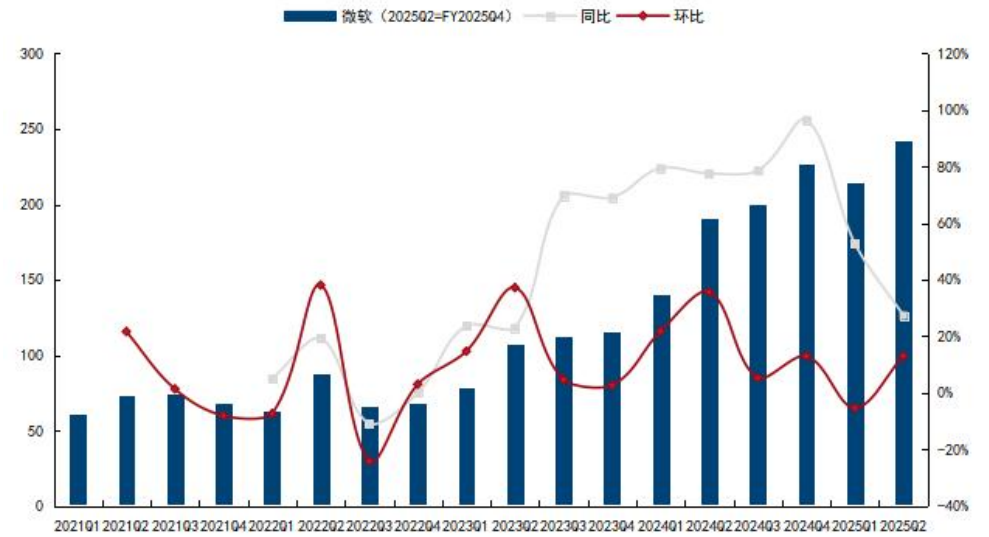
资料来源：公司财报，国信证券经济研究所整理

各大厂大幅提升资本开支，AI 投入进入白热化

分厂商来看，微软 2025Q2 资本支出达 242 亿美元，同比增长 27.37%，环比增长 13.08%，超出市场预期的 231 亿美元。从资本支出占营收比重来看达到 31.66%，创下历史新高，远超过去几年的平均水平。这种高强度的投资在微软历史上前所

未有，反映出公司对 AI 基础设施建设的战略重视。公司明确指出 2025Q3 资本支出预计将超 300 亿美元，年增长率将超 50%。意味着单季度资本支出将达到公司季度营收的近 40%。公司还指引 2026 财年上半年的资本支出增速将快于下半年，全年资本支出预计在 660-720 亿美元区间，显示对 AI 基础设施建设的战略重视。

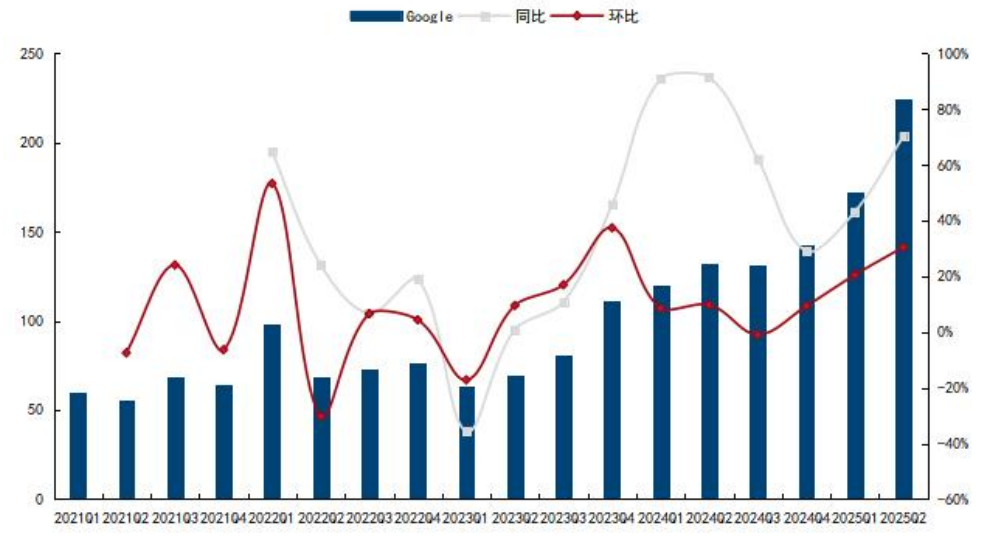
图24：微软资本开支快速增长（单位：亿美元）



资料来源：公司财报，国信证券经济研究所整理

谷歌 2025Q2 资本支出达 224 亿美元，同比增长 70.23%，环比增长 30.5%。从资本支出占营收比重来看，谷歌的投资强度显著提升，达到 23.28%，反映出公司对 AI 基础设施建设的高度重视。谷歌大幅上调全年资本支出指引，从 750 亿美元上调至 850 亿美元，同比增长预计将达 62%，标志着谷歌历史上最大规模的资本支出计划。

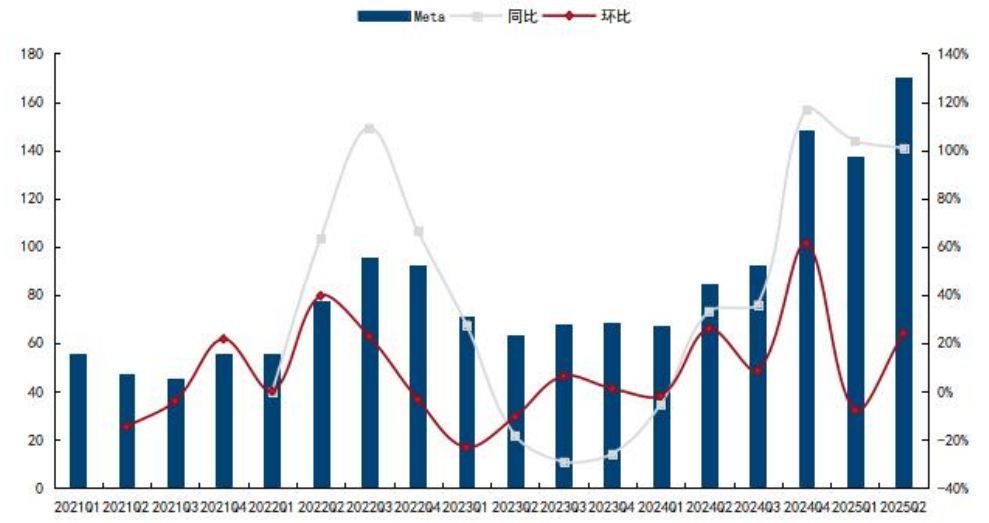
图25：谷歌资本开支快速增长（单位：亿美元）



资料来源：公司财报，国信证券经济研究所整理

2025Q2，Meta 资本支出达 170 亿美元，同比增长 100.85%，环比增长 24.18%。资本支出占营收比重达到 35.8%，创下公司历史最高水平。公司上调全年资本支出指引，将下限从 640 亿美元上调至 660 亿美元，资本支出范围调整为 660 亿美元至 720 亿美元，按中值计算同比增长率达 77%。

图26：Meta 资本开支快速增长（单位：亿美元）



资料来源：公司财报，国信证券经济研究所整理

亚马逊 2025Q2 资本支出达 314 亿美元，同比增长 91.46%，环比增长 29.22%。资本支出占营收比重达到 18.72%，公司预计二季度的资本开支代表了下半年的季度资本开支水平，意味着全年资本支出可能超过 1200 亿美元。

图27：亚马逊资本开支快速增长（单位：亿美元）

资料来源：公司财报，国信证券经济研究所整理

Sora 2 发布，视频生成模型进入 GPT 时刻

OpenAI 发布 Sora 2，性能全面提升

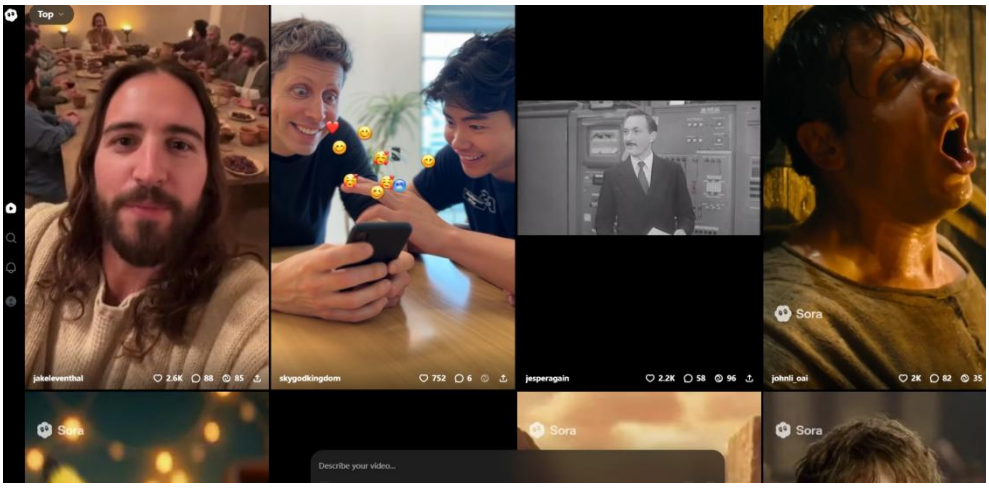
Sora 2 发布，视频生成能力大幅提升。当地时间 2025 年 9 月 30 日，OpenAI 发布最新的旗舰视频与音频生成模型 Sora 2。最初的 Sora 模型在 2024 年 2 月推出，在许多方面堪称视频领域的 GPT-1 时刻。从那时起，Sora 团队便专注于训练具有更先进世界模拟能力的模型。OpenAI 认为 Sora 2 发布直接跨越到了视频领域的 GPT-3.5 时刻，可以做到此前视频生成模型极其困难、甚至不可能做到的事情：如奥运体操动作、在桨板上做后空翻并准确模拟浮力与刚性的动力学效果等，该模型在可控性方面也实现了巨大飞跃，能够遵循跨越多个镜头的复杂指令，同时准确保持世界状态的延续性，在现实风格、电影风格以及动漫风格的视频生成上都表现出色。

图28：Sora 2 生成电影级视频

资料来源：公司官网，国信证券经济研究所整理

发布全新社交软件，拓宽模型应用场景。Sora 2 能够创造复杂的背景声效、语音和音效，并具备高度的真实感。用户可以将现实世界的元素直接注入到 Sora 2 中，例如通过观察 OpenAI 团队成员的视频，模型就能将其插入到任意 Sora 生成的环境中，并且准确还原外貌和声音。这一能力具有高度的通用性，适用于任何人类、动物或物体，这验证了在视频数据上继续扩展神经网络规模以更接近模拟现实的路径。同时，OpenAI 正式发布一款新的社交 iOS 应用，由 Sora 2 驱动，用户可以创作、混合彼此的生成内容，在可定制的 Sora 动态中发现新视频，并通过 cameo（客串）功能把自己或朋友带进作品里。用户只需在应用中完成一次性的视频和音频录制，用于验证身份并捕捉形象，就能直接把自己放入任意 Sora 场景中。Sora iOS 应用已可下载，在美国和加拿大启动首轮上线。

图29：Sora 社交媒体已全面可用

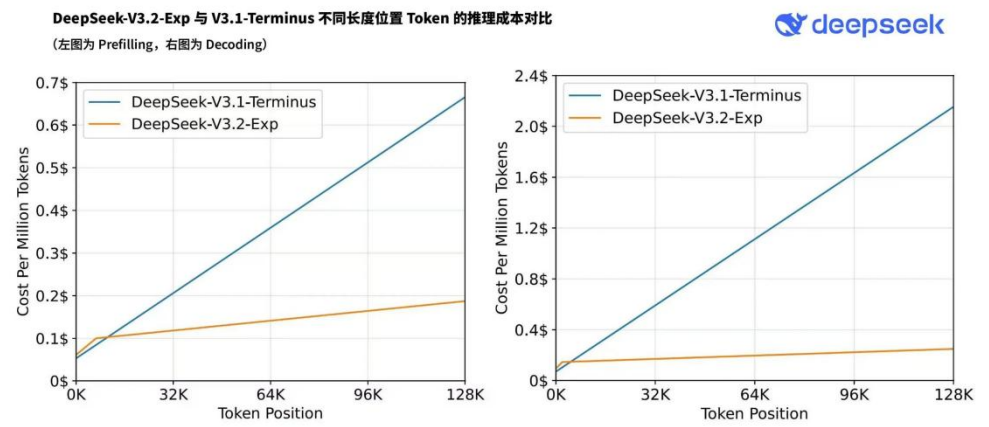


资料来源：公司官网，国信证券经济研究所整理

DeepSeek-V3.2、GLM-4.6 发布，国产模型快速迭代

DeepSeek-V3.2 引入稀疏注意力机制，效率大幅提升。9 月 29 日，DeepSeek 宣布正式发布 DeepSeek-V3.2-Exp 模型，V3.2-Exp 是一个实验性的版本，在 V3.1-Terminus 的基础上引入了 DeepSeek Sparse Attention（一种稀疏注意力机制），在几乎不影响模型输出效果的前提下，可以实现长文本训练和推理效率的大幅提升。为了评估引入稀疏注意力带来的影响，DeepSeek 团队将 DeepSeek-V3.2-Exp 的训练设置与 V3.1-Terminus 进行了严格的对齐。在各领域的公开评测集上，DeepSeek-V3.2-Exp 的表现与 V3.1-Terminus 基本持平，用户调用 DeepSeek API 的成本降低 50%以上。

图30：DSA 实现推理、训练效率的大幅提升



资料来源：公司官网，国信证券经济研究所整理

国产芯片迅速适配，算力生态成熟度提升。DeepSeek-V3.2-Exp 发布后，寒武纪、华为昇腾、海光信息等国产芯片厂商迅速完成适配，华为昇腾宣布基于 vLLM/SGLang 等推理框架快速完成适配部署，支持并开源了所有推理代码和算子实现，寒武纪宣布成功兼容 DeepSeek 最新模型，并开源大规模模型推理引擎 vLLM-MLU 源代码，海光信息表示其 DCU 已实现无缝适配和深度优化，国产算力生态在通用大模型训练、推理任务中的成熟度显著提升。

图31：国产算力厂商与 DeepSeek 快速适配

资料来源：公司官网，国信证券经济研究所整理

9 月 30 日，智谱 AI 正式发布了旗下新一代旗舰模型 GLM-4.6，在多个方面实现了全面提升，包括但不限于：

1）高级编码能力：在公开基准与真实编程任务中，GLM-4.6 代码能力对齐 Claude Sonnet 4；

2）上下文长度：上下文窗口由 128K 增加至 200K，适应复杂的代码与智能体任务；

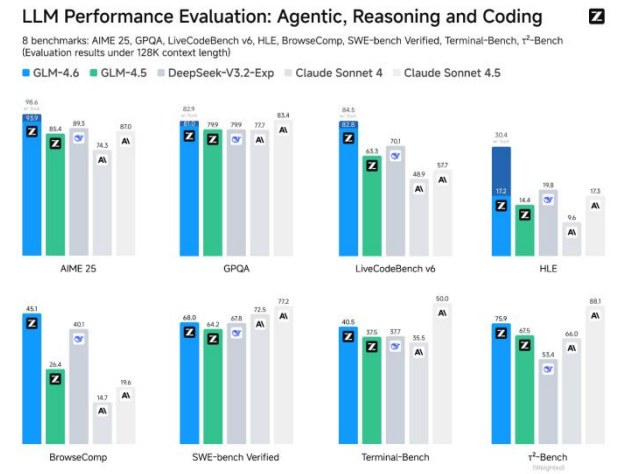
3）推理能力提升，并支持在推理过程中调用工具；

4）增强了模型的工具调用和搜索智能体，在智能体框架中表现更好；

5）更强的写作能力：在文风、可读性与角色扮演场景中更符合人类偏好。

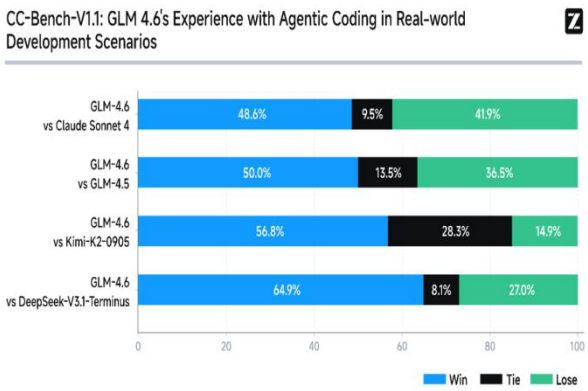
GLM-4.6 模型在基准评测上性能有了全面提升，其中多个基准上胜过了 Claude Sonnet 4/Claude Sonnet 4.5，位居国产模型首位。在平均 token 消耗上，GLM-4.6 比 GLM-4.5 节省了 30%以上。同时，GLM-4.6 已在寒武纪芯片上实现 FP8+Int4 混合量化部署，这是首次在国产芯片投产的 FP8+Int4 模型芯片一体解决方案。该方案在保持精度不变的前提下，可以大幅降低推理成本，为国产生态下大模型本地化运行开创了可行路径。

图32：GLM-4.6 测评性能全面提升



资料来源：公司官网，国信证券经济研究所整理

图33：GLM-4.6 编程能力超越 Claude Sonnet 4



资料来源：公司官网，国信证券经济研究所整理

投资建议：关注算力建设与 AI 应用领域机会

模型方面，阿里云云栖大会中，通义大模型实现七连发，在模型智能水平、Agent 工具调用和 Coding 能力、深度推理、多模态等方面实现多项突破。阿里首次系统阐述了通往 ASI 的演进路线，未来三年阿里将投入超 3800 亿元用于建设云和 AI 硬件基础设施，持续升级全栈 AI 能力。OpenAI 发布最新的旗舰视频与音频生成模型 Sora 2，并正式发布一款新的社交 iOS 应用，由 Sora 2 驱动，用户可以创作、混合彼此的生成内容，视频生成模型进入 GPT 时刻。DeepSeek 宣布正式发布 DeepSeek-V3.2-Exp 模型，引入稀疏注意力机制，实现长文本训练和推理效率的大幅提升。智谱 AI 正式发布 GLM-4.6，在基准评测上性能有了全面提升，其中多个基准上胜过了 Claude Sonnet 4/Claude Sonnet 4.5。当前全球模型快速迭代，模型视频、音频等多模态能力显著提升，可用性快速提高，应用场景有望持续拓宽。

算力方面，随着生成式 AI 和大模型进入推理时代，全球主要云计算厂商纷纷布局自研芯片，ASIC 已成为云厂商在高强度推理场景下降低成本、优化性能、增强生态主导力的重要方向，谷歌已经开始将其自研的 AI 芯片 TPU 部署在较小型云服务商运营的数据中心中，AI 芯片竞争逐步转向大厂自研芯片。甲骨文 FY2026 财报披露其剩余履约义务飙升至 4550 亿美元，同比增长 359%，其中仅第一季度新增即达 3170 亿美元，公司上修全年 CapEx 指引至约 350 亿美元。英伟达和 OpenAI 宣布达成合作，为 OpenAI 的下一代 AI 基础架构部署至少 10 吉瓦的英伟达系统。当前全球 AI 投资加速，2025Q2 海外大厂支合计资本开支达 1035.58 亿美元，同比增长 74.47%，占总收入比重达历史最高的 25.7%。

当前全球 AI 快速推进，算力投资持续加码，模型能力加速迭代，有望在未来加速商业化进程，进而形成产业闭环，推动算力投资持续提升，建议关注 AI 应用及算力相关个股，如海光信息、博通、甲骨文、用友网络、新大陆等。

风险提示

AI 应用落地不及预期、市场需求不及预期、行业竞争加剧、宏观经济波动、新技术研发不及预等。

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