AI 主题研究

关注产业链协同发展布局的投资机会

从近期行业趋势来看，美国芯片巨头、云服务商、AI 模型与应用厂商正在深化战略协同、提升全产业链的效率并深化商业化变现探索。Sora 2 上线加速 AI 视频应用生态发展，而应用生态繁荣也反向驱动推理算力需求快速增长。我们仍坚定看好 AI 主题的投资机会，并建议关注在产业链协同发展中的投资机会。中国互联网软件公司中，建议关注云业务增长有望维持强劲、芯片端有良好布局且 AI 相关新业务营收持续起量的的阿里巴巴（BABA US）、百度（BIDU US），及 AI 拉动广告和云业务较快增长的腾讯（700 HK）、可灵变现进展积极的快手（1024 HK）。海外互联网软件公司中，建议关注云业务营收持续快速增长的微软（MSFT US），及 AI 变现取得实质进展、估值仍有提升空间的派拓网络（PANW US）、Datadog（DDOG US）。

- **中美 AI 产业链的全链协同与循环生态。**美国 AI 领域当前芯片巨头、云服务商、AI 模型与应用领导者（OpenAI、xAI、Meta 等）正在深化战略协同，投资合作形成 AI 基建、算力需求、模型创新迭代与应用落地的正向循环。这一循环在降低大模型与应用厂商算力采购成本、精准匹配其算力需求的同时，也帮助云服务提供商与芯片企业实现资源的合理调度，提升 AI 基建 ROI 与利用效率。我们认为全产业链协同的商业模式或有望在中国市场复制，推动芯片企业、云服务商以及 AI 大模型厂商与行业解决方案提供商的深度协作，通过长期协议、投资入股等方式，提升全链资源调度效率的同时解决算力需求瓶颈，推动行业快速高效发展。

- **AI 应用生态繁荣驱动推理算力需求快速增长。**2025 年 9 月 OpenAI 正式上线视频与语音生成模型 Sora 2，有望加速 AI 视频应用生态发展。AI 应用生态繁荣推动全球主流大模型的调用量快速增长。根据 OpenRouter 数据，全球主流大模型周调用量由 2025 年初的 4997 亿增长至 2025 年 9 月的约 4.9 万亿 Token，9 个月内增长近 10 倍，主要得益于：1）AI 应用的快速落地：编程、广告营销、搜索等多个场景的 AI 应用落地带来 Token 调用量的快速增长；2）智能体应用渗透率提升；3）图像与视频大模型的调用增加。根据 IDC 数据，中国智能算力规模预计将由 2025 年的 1,037 EFLOPS 增长至 2028 年的 2,782 EFLOPS，2025-2028E 复合增长率达到 39%。在 AI 算力需求快速增长以及 AI 应用渗透率提升的驱动下，中国 AI 推理芯片市场规模预计将从 2025 年的 3106 亿元增长至 2029 年的 1.38 万亿元，复合增速为 45%，其中专为推理设计的 NPU 占比将从 2025 年的 19%增长至 2029 年的 29%（据 IDC，CIC）。

- **关注 AI 发展浪潮中产业链协同的投资机会。**报告中我们对中国主要 AI 相关互联网公司的 AI 业务布局及进展进行了梳理，同时梳理了其近两年对外投资及协同主业布局的情况。我们观察到行业 2022 年起逐步减少对外的大规模投资，更加关注投资与主业的协同发展，布局更偏向于支持前沿科技、丰富 AI 相关的应用场景、及推进核心主业的国际扩张。硬件方面，中国互联网企业在芯片领域的投资布局以产业链协同整合和推理芯片突破为核心。通过投资存储、网络、光电子等产业链关键环节构建完整生态，并重点发力边缘 AI 芯片、车载芯片及大模型推理 GPU。

优于大市
（维持）

中国软件 & IT 服务行业

贺赛一, CFA
(852) 3916 1739
hesaiyi@cmbi.com.hk

陶冶, CFA
franktao@cmbi.com.hk

陆文韬, CFA
luwentao@cmbi.com.hk

马泽慧
(852) 3761 8838
joannama@cmbi.com.hk

相关行业报告:

1. AI 主题研究 - 政策加速 AI 应用与商业化落地 - 29 Aug
2. 软件 & IT 服务 - 海外云厂商：营收增速环比加快，利润率表现分化 - 6 Aug
3. AI 主题研究 - 模型能力持续提升，商业化持续推进 - 27 Jun
4. AI 主题研究 - 模型调用成本下降，应用生态有望逐步繁荣 – 26 Feb
5. 软件 & IT 服务 - 2025 展望：关注 AI 应用营收起量趋势 – 12 Dec

相关公司报告:

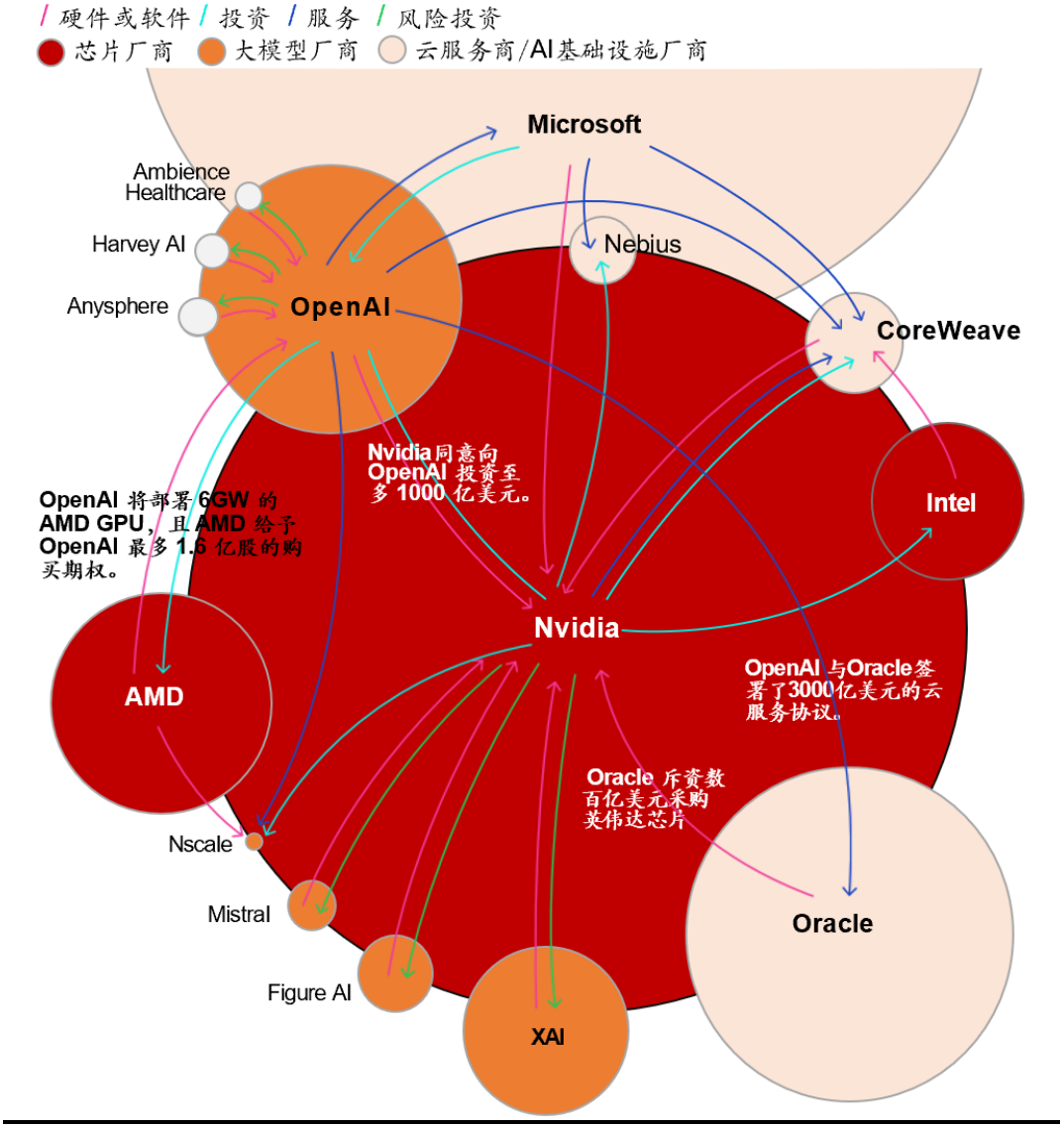
1. Alibaba (BABA US) - Quick commerce investment peaked out; accelerating AI adoption – 10 Oct
2. Baidu (BIDU US) - 3Q25 preview: expecting inline results and more updates from new businesses – 15 Oct
3. Salesforce (CRM US) - Inline 2QFY26 results; intact long-term outlook for AI and data cloud – 5 Sep
4. SenseTime (20 HK) - 1H25 review: strong Gen AI revenue momentum with improving margin – 29 Aug
5. Palo Alto Network (PANW US) - Solid rev growth and profitability expansion likely to sustain in FY26 – 20 Aug
6. Tencent (700 HK) - 2Q25 results: strong games and marketing businesses; AI drove business growth – 14 Aug
7. Kingdee (268 HK) - Recovery in revenue growth in line with expectation, with more updates on AI – 13 Aug
8. Datadog (DDOG US) - Solid 2Q results with better-than-expected operating efficiency gains – 8 Aug
9. Microsoft (MSFT US): Results a solid beat; cloud acceleration better than expectation driven by strong demand – 1 Aug

目录

中美 AI 产业链的全链协同与循环生态

美国 AI 领域当前正通过芯片巨头（NVIDIA、AMD、Intel 等）、云服务商与 AI 基础设施服务商（微软 Azure、谷歌云、Oracle 等）、AI 模型与应用领导者（OpenAI、xAI、Meta 等）的深度绑定与协同，举全行业之力投资合作形成 AI 基建、算力需求、模型创新迭代与应用落地的正向循环。这一循环在降低大模型与应用厂商算力采购成本、精准匹配其算力需求的同时，也帮助云服务提供商与芯片企业实现资源的合理调度，提升 AI 基建 ROI 与利用效率。我们认为，美国 AI 产业链的全链合作协同有望进一步提升 AI 全行业资源利用效率，进一步加速 AI 生态繁荣。同时，我们仍坚定看好中国 AI 主题的投资机会，并建议关注在产业链协同发展中的投资机会。

图 1：美国：AI 的全链协同与循环生态



资料来源：彭博，招银国际环球市场

注：1）箭头代表资金流转方向；2）本图为示意图，并未穷尽所有投资案例

在 OpenAI 与 AMD 官宣合作前，2025 年 9 月，英伟达宣布向 OpenAI 投资至多 1000 亿美元，用于建设搭载数百万块 NVIDIA 芯片的数据中心（如“Vera Rubin”平台），算力规模超 10 GW，合作期限至 2028 年。同时 OpenAI 承诺优先采购 NVIDIA 芯片，并与 NVIDIA 联合优化 AI 基础设施。英伟达和 OpenAI 宣布 1000 亿美元投资协议的第二天，OpenAI 被证实已与甲骨文达成另一项 3000 亿美元的协议，用于在美国建设数据中心。甲骨文则斥资数十亿美元为这些设施购买英伟达芯片，资金形成积极良性循环。

此外，英伟达支持的高性能云计算公司 CoreWeave（英伟达持股 7%）已与 OpenAI 达成达成最新合作协议，年内第三次扩大租赁合作，为其下一代先进模型训练提供算力支持。此次合作合同价值高达 65 亿美元，叠加此前协议，CoreWeave 与 OpenAI 的合同总额已达 224 亿美元。同时，2025 年 9 月，NVIDIA 与 CoreWeave 签订了新订单，初始价值达 63 亿美元。该协议确立了公司向其客户销售预留云计算容量的安排，并允许英伟达获取任何剩余未售出的云计算容量，形成“投资 - 采购 - 再投资”的循环。

2025 年 10 月，据彭博社报道，微软已承诺向多家包括 Nebius、CoreWeave 等在内的新云（Neocloud）服务提供商投资超 330 亿美元，据悉此次交易有望为微软带来超 10 万颗英伟达 GB300 芯片的使用权，有助于缓解 AI 算力不足的问题。微软一直是 CoreWeave 最重要的客户之一。据报道，微软在 2023 年至 2030 年间将花费近百亿美元从 CoreWeave 租用服务器来运行人工智能模型，这一金额占 CoreWeave 与客户签署的 170 亿美元合同总额的一半以上。

此外，黄仁勋已确认英伟达计划参与马斯克的 xAI 公司的股权融资，投资至多 20 亿美元。此轮融资总额为 200 亿美元，其中约 75 亿美元为股权融资，最高 125 亿美元为债务融资，并将通过 SPV（特殊目的载体）实现。该 SPV 将被用于购买英伟达的处理器，而 xAI 会在未来五年出租这些芯片；OpenAI、甲骨文、软银联合发起 “Stargate” 项目，计划投资 5000 亿美元建设 10GW AI 数据中心，首批选址德克萨斯州阿比林、新墨西哥州等，甲骨文负责硬件采购（以 NVIDIA 芯片为主），软银则提供资金与基建支持；微软向 OpenAI 追加“多年期、数十亿美元” 投资，延续 Azure 作为 OpenAI 独家云服务商的地位，并整合 OpenAI 技术到 Office、GitHub Copilot 等产品中；谷歌云成为首批采用 NVIDIA GB300 NVL72 服务器的云服务商，并与 NVIDIA 联合优化谷歌开源模型 Gemma，同时 NVIDIA 采用谷歌 DeepMind 的 SynthID 技术为 AI 生成内容加水印；Meta 投资近 90 亿美元采购 35 万块 NVIDIA H100 GPU，并计划将 GPU 总数增至 100 万块，用于训练 Llama 系列大模型及 AGI 研究；英伟达与英特尔达成了历史性联盟，NVIDIA 向 Intel 投资 50 亿美元，双方联合开发数据中心和 PC 端的 AI 芯片，整合 NVIDIA 的 GPU 技术与 Intel 的 x86 CPU 架构。

长期来看，我们认为全产业链协同的商业模式或有望在中国市场复制，推动芯片企业（如华为昇腾、寒武纪等）、云服务商（阿里云、腾讯云等）以及 AI 大模型厂商与行业解决方案提供商的深度协作，通过长期协议、股权联动等方式，有望降低我国算力投入风险、提升全链资源调度效率的同时解决算力需求瓶颈，推动全行业健康良性发展。

在中国的 AI 产业链中，我们目前已经看到了协同进一步深化的早期趋势。举例而言，商汤与寒武纪 2025 年 10 月签署战略合作，重点推进软硬件的联合优化，共同构建开放产业生态。合作主要分两大方面：1）芯片适配方面，双方将积极推进最新型号的软硬件产品适配，联合打造面向算力市场的服务方案；2）一体机方面，双方将聚焦企业服务等垂直行业场景，紧密结合各自软硬件能力，打造面向垂直领域的一体机解决方案。

美国主要互联网及云计算厂商 AI 业务发展

海外互联网及云计算厂商 AI 业务进展积极：1）Microsoft：截至 4QFY25，Copilot 系列月活用户超 1 亿；Azure 云营收增速重新加速：4QFY25 Azure 及其他云服务收入同比增长 39%；2）Amazon：2Q25 AWS 实现收入 309 亿美元，同比增长 17.5%（1Q25：16.9%），AI 相关营收持续以三位数增速扩张；3）Google：2Q25 谷歌云订单环比增长 18%，同比增长 38%，达到 1060 亿美元，得益于客户对 AI 产品的强劲需求；4）Meta：优化 AI 驱动的广告推荐模型，使 Instagram 和 Facebook 的广告转化分别提升 5%和 3%。

图 2：美国主要互联网及云计算厂商 AI 业务发展

公司	AI 进展
Microsoft	**1）Copilot 产品矩阵变现持续推进：**截至 4QFY25，Copilot 系列月活用户超 1 亿，AI 功能覆盖 8 亿月活用户；Microsoft 365 Copilot 企业部署加速（如巴克莱扩至 10 万员工）； **2）Azure 云营收增速重新加速：**4QFY25 Azure 及其他云服务收入同比增长 39%，超过此前公司 34%-35%的指引； **3）消费者端 AI 创新和游戏联动：①Windows 与 Edge：Copilot Mode 整合搜索、创作和实时屏幕分析，Windows 11 全面预装 Copilot Vision，推动换机潮；②LinkedIn AI 化：招聘与销售代理推动评论量增长 30%，视频上传量增 20%，会员数达 12 亿。**
Amazon	**1）云业务营收稳健增长，AI 云持续带来增量贡献：**2Q25 AWS 实现收入 309 亿美元，同比增长 17.5%（1Q25：16.9%），AI 相关营收持续以三位数增速扩张； **2）自研芯片优势持续显现：**Trainium2 芯片成为 Anthropic 新一代 Claude 模型及 Amazon Bedrock 的算力支柱，推理成本比行业 GPU 低 30%-40%，第三代芯片已启动研发； **3）物流智能化：**AI 机器人 DeepFleet 提升仓储效率 10%，百万机器人协同作业降低 15%单位处理成本，支持 Prime 会员当日/次日达订单量同比增长 30%； **4）消费者端 AI 产品商业化：**Alexa+：下一代生成式 AI 助手，支持多设备联动和复杂指令执行（如“布置晚餐派对”），非 Prime 会员订阅价 19.99 美元/月，用户活跃度与调用量显著提升。
Google	**1) AI 提升广告转化效果：**AI MAX and Search 推动搜索广告转化效果提高 14%； 2) AI 竞价工具 Smart Bidding Exploration 平均提升 19%的转化效果。2Q25 已有超 200 万广告主在广告投放中使用 Google 的 AI 素材生成工具，同比增长 50%。 **3) AI 云业务加速增长：**2Q25 谷歌云收入同比增长 32%至 136 亿美元，主要得益于 GCP 核心产品和 AI 产品的强劲增长。2Q25 谷歌云订单环比增长 18%，同比增长 38%，达到 1060 亿美元，得益于客户对 AI 产品的强劲需求。2Q25 超过 2.5 亿美元的订单数量同比增长一倍，GCP 新客户数量环比增长 28%。
Meta	**1）Meta 优化 AI 驱动的广告推荐模型，**使 Instagram 和 Facebook 的广告转化分别提升 5%和 3%； **2）公司利用 AI 优化内容推荐系统，**Facebook 和 Instagram 用户使用时长分别增加 5%和 6%； 3）生成式 AI 广告创意工具 Advantage+ Creative 的用户群体不断扩大，目前已有约 200 万广告主正使用视频生成、视频扩展等生成式 AI 功能。

资料来源：公司资料，招银国际环球市场

中国主要互联网及云计算厂商 AI 业务发展及产业链投资梳理

我们观察到中国主要的互联网厂商自 2022 年起逐步减少了对外的大范围投资布局，更加关注核心主业的协同发展，同时整体布局更偏向于支持前沿科技及探索 AI 相关的应用场景及核心主业的国际扩张。在硬件方面，中国互联网企业在芯片领域的投资布局以产业链协同整合和推理侧芯片突破为核心。通过投资存储、网络、光电子等产业链关键环节（如 ReRAM 存算一体、DPU 智能网卡等）构建完整生态，并重点发力边缘 AI 芯片、车载推理芯片及大模型推理 GPU。

■ 阿里巴巴：围绕“消费”及“AI+云”两大核心战略开展布局

在投资端和业务发展端，阿里巴巴均围绕“消费”及“AI+云”两大核心战略展开布局。在业务端，阿里巴巴于 FY26（财年结于 3 月）对公司财报口径列报的业务单元进行了重新划分，从此前的“1+6+N”架构（即 1 个集团+6 大业务集团+N 个其他业务）变更为四大业务板块：阿里巴巴中国电商集团、阿里国际数字商业集团、云智能集团及所有其他业务。此前

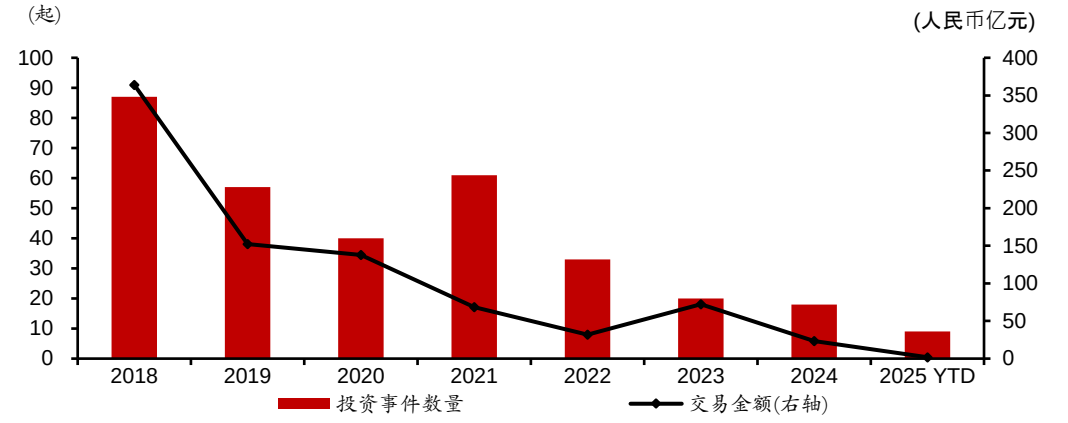
2023 年底，公司曾对外公布夸克、1688、闲鱼、钉钉为阿里四个战略级创新业务，公司将持续布局资源、推动发展。

图 3：阿里巴巴：集团业务板块划分

资料来源：公司资料，招银国际环球市场

在投资端，阿里巴巴自 2022 年起逐步收缩了对外的投资，更加专注主业的发展，逐步收缩退出非核心业务，强化股东回报。据鲸准数据，公司自 2022 年起对外投资事件数持续下降，2025 年 YTD 对外投资数量为 9 起，较 2025 年的 18 起有进一步的收缩，总对外投资体量也有进一步的下滑。

图 4：阿里巴巴：对外投资事件数量及交易金额



资料来源：鲸准，招银国际环球市场

而总结 2024-2025 年的投资布局来看，阿里巴巴主要的对外投资集中于与“消费”及“AI+云”相关的领域，主要的投资标的包括全球有电商发展潜力国家的电商平台，AI、机器人、大模型等前沿科技探索方向等。

图 5：阿里巴巴：近两年对外投资情况梳理及被投公司简介

被投公司简称	一句话简介	交易时间	交易轮次
扬杨得熠文化创意	潮玩艺术生活品牌	2025-09-18	战略投资
造父智能	L4 级自动驾驶技术研发商	2025-09-17	战略投资
爱诗科技	AI 视频大模型研发商	2025-09-10	B 轮
宇树科技	四足机器人与动力系统部件研发商	2025-06-19	C+轮
圣达信教育	升学规划服务提供商	2025-05-02	战略投资
两氢一氧	跨境电商品牌	2025-04-01	并购
链企智能	AI 商业搜索和标讯服务平台	2025-03-27	战略投资
源络科技	机器人研发商	2025-02-12	A 轮

喵街	本地生活信息平台	2025-01-13	并购
原力聚合	信息系统集成服务商	2024-12-16	天使轮
地平线	AI 芯片研发商	2024-10-16	PreIPO
星动纪元	通用人形机器人研发商	2024-10-16	Pre-A 轮
神漫文化	动漫影视产品制作商	2024-09-18	A 轮
Connectly.ai	美国营销服务商	2024-09-12	B 轮
源络科技	机器人研发商	2024-09-06	战略投资
月之暗面	下一代跨模态大模型研发商	2024-08-06	B 轮
百川智能	人工智能底层大模型技术提供商	2024-07-25	A+轮
逐际动力	通用足式机器人研发商	2024-07-15	A 轮
OceanBase	原生分布式数据库研发商	2024-06-29	战略投资
瀚博半导体	AI 视觉芯片研发商	2024-06-14	C 轮
来未来科技	企业数字化转型服务商	2024-06-03	B 轮
iSlide	辅助 PowerPoint 一键美化插件工具	2024-05-31	战略投资
精准学	AI 教育服务提供商	2024-05-29	B 轮
Lazada	新加坡在线购物平台	2024-05-21	战略投资
ABLY Corp	韩国电商平台运营商	2024-04-23	战略投资
MiniMax	AI 大模型研发商	2024-03-04	B 轮
Konvy.com	泰国美妆电商平台	2024-01-08	B 轮

资料来源：鲸准，招银国际环球市场

在 AI 芯片产业链方向，阿里巴巴的主要投资包括：清微智能、瀚博半导体、墨芯人工智能、沐创集成电路、原子半导体、长鑫存储，其中清微智能是可重构计算芯片领域的领军企业，致力于提供从端侧到云侧的芯片产品及解决方案，长鑫存储持续推动国内存储芯片产业发展。

图 6：阿里巴巴：对外投资半导体企业情况梳理

序号	投资公司		轮次	投资时间
1	清微智能	AI 芯片设计（FPGA/ASIC 融合架构），适用于 AI 推理场景的灵活部署	D 轮	2025.01
2	瀚博半导体	AI 芯片设计（GPU/DPU），提供一站式独立训练推理解决方案	C+轮	2024.06
3	墨芯人工智能	AI 芯片设计（稀疏计算 ASIC），主要聚焦于大模型推理芯片领域	战略投资	2024.04
4	沐创集成电路	安全芯片（密码算法芯片）	B 轮	2023.11
5	原子半导体	高性能嵌入式存储芯片	Pre-A 轮	2022.03
6	长鑫存储	存储芯片（DRAM 制造），为 AI 训练服务器提供高带宽内存支持	战略投资	2022.02

资料来源：IT 桔子，招银国际环球市场

■ **腾讯：布局全栈 AI 云+AI 应用，打造智能化增长引擎**

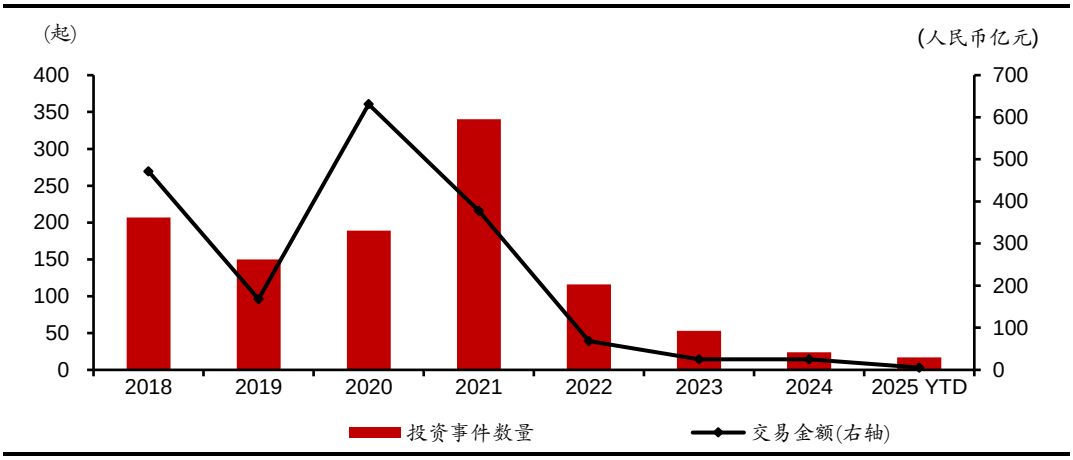
腾讯目前已打造了全栈云+AI 应用的产品矩阵。AI 云方面，腾讯提供基础设施、平台、应用层的全栈云计算解决方案；AI 应用方面，公司目前有超百款应用完成 AI 升级，包括微信、腾讯会议、腾讯文档等，此外公司还打造了元宝、ima 知识库、QQ 浏览器等 AI 原生应用。

图 7：腾讯：全栈云+AI 产品矩阵

资料来源：公司资料，招银国际环球市场

腾讯自 2021 年以来逐渐减少对外投资，同时出售部分已经成熟的投资业务，将资金投入到 Capex 以及公司股票回购。根据鲸准数据，腾讯 2025 年 YTD 主要投资数量共 17 笔，相较 2021 年全年峰值 340 笔有明显回落。

图 8: 腾讯：对外投资事件数量及交易金额



资料来源：鲸准，招银国际环球市场

从 2024-2025 年的投资来看，腾讯对外投资覆盖多个领域，包括 AI、游戏、生物医药、消费等多个行业。

图 9：腾讯：近两年对外投资情况梳理及被投公司简介

被投公司简称	一句话简介	交易时间	交易轮次
民为生物	创新生物药及器械研发商	2025-09-19	A+轮
Born	德国 AI 游戏社交产品研发商	2025-09-11	A 轮
赛丽科技	无晶圆芯片设计服务商	2025-08-05	战略投资
Uzum	乌兹别克斯坦数字服务生态系统提供商	2025-08-05	B 轮
华橙网络	智能家居品牌	2025-07-17	并购
维立志博	原创性抗体新药研发商	2025-07-17	PreIPO
科思明德	医疗器械研发商	2025-07-04	B 轮
宇树科技	四足机器人与动力系统部件研发商	2025-06-19	C+轮
智达星空	创新装备研发商	2025-03-27	战略投资
艺展弘程	演出经纪服务商	2025-03-26	并购
智元机器人	人形机器人研发商	2025-03-24	B 轮
伯希和户外	户外运动及装备品牌商	2025-03-11	战略投资
集益威半导体	高端 IC 设计服务商	2025-02-27	C+轮
GameScience 游戏科学	游戏开发商	2025-02-05	A 轮
古茗	茶饮连锁品牌	2025-02-04	PreIPO
迅驰迅捷	新能源汽车充电桩研发运营商	2025-01-09	A 轮
鲲为科技	超声成像技术研发商	2025-01-06	A+轮
礼邦医药	肾脏药物研发商	2024-12-26	D 轮
阶跃星辰	AI 大模型服务商	2024-12-23	B 轮
云鲸智能	家用机器人研发商	2024-12-16	E 轮及以后
Manus AI	通用型 AI Agent 产品研发商	2024-12-01	Pre-A 轮
英派药业	以合成致死为机制的抗癌新药研发商	2024-11-15	E 轮及以后
智谱 AI	AI 知识智能技术开发商	2024-11-01	战略投资
Insighta	早期癌症检测服务商	2024-10-15	战略投资
TrueLayer	英国银行应用程序接口服务提供商	2024-10-10	E 轮及以后
食气生化	工业废气处理解决方案提供商	2024-08-31	Pre-A 轮
月之暗面	下一代跨模态大模型研发商	2024-08-06	B 轮
费曼动力	电催化研发商	2024-07-30	A 轮
百川智能	人工智能底层大模型技术提供商	2024-07-25	A+轮
傲爵数码	互联网信息服务提供商	2024-07-19	并购
Chainbase	Web3 开发者平台	2024-07-18	A 轮
银之心	信息系统集成服务商	2024-07-05	并购
老铺黄金	古法手工金器品牌	2024-06-20	PreIPO
晟斯生物	生物大分子新药研发商	2024-05-30	D 轮
跃赛生物	多能干细胞药物研发商	2024-04-01	A 轮
Monzo	移动手机银行	2024-03-05	E 轮及以后
信诺维	生物医药研发商	2024-02-20	E 轮及以后
网元圣唐	网络游戏开发公司	2024-02-03	并购
腾讯出行	出行服务提供商	2024-02-02	战略投资
明略科技	大数据整体解决方案提供商	2024-01-25	E 轮及以后
圣方医药研发	临床 CRO 研究解决方案提供商	2024-01-19	A+轮

资料来源：鲸准，招银国际环球市场

在 AI 芯片产业链方向，腾讯的主要投资包括：云豹智能、星空科技、集益威半导体、费曼动力、无问芯穹、燧原科技、合见工软、轻舸光电、光舟半导体、长鑫存储等企业，其中长鑫存储是国内存储芯片领域的关键力量，合见工软在 EDA 工具领域为芯片设计提供支持，燧原科技专注于人工智能算力芯片研发，无问芯穹核心业务是打造连接多种大模型与多种芯片的“M×N”异构 AI 基础设施新范式。

图 10: 腾讯：对外投资半导体企业情况梳理

序号	被投公司	芯片产业链环节	轮次	投资时间
1	云豹智能	数据中心基础设施（DPU 芯片设计），加速 AI 工作负载的数据处理和通信	C 轮	2025.04
2	星空科技	半导体设备（光刻机、键合机等高端装备制造）	B 轮	2025.03
3	集益威半导体	模拟芯片设计，用于 AI 芯片的时钟同步、信号转换等模块	B 轮	2025.02
4	无问芯穹	异构算力整合平台，支持不同芯片的协同训练与推理	A 轮	2023.12
5	燧原科技	云端算力芯片（推出的高性能 AI 加速卡支持千亿级参数大模型的实时推理）	战略投资	2023.09
6	合见工软	EDA 工具链（数字芯片设计全流程工具）	战略投资	2023.06
7	轻舟光电	光通信模块（支持 AI 训练集群的超高速数据传输）	A 轮	2022.11
8	光舟半导体	光通信模块（光芯片、光互连解决方案）	A+轮	2022.08
9	原子半导体	高性能嵌入式存储芯片	Pre-A 轮	2022.03
10	长鑫存储	存储芯片（DRAM 制造），为 AI 训练服务器提供高带宽内存支持	战略投资	2022.02

资料来源：IT 桔子，招银国际环球市场

■ 百度：关注广告业务增长新驱动、Robotaxi 及 AI 新业务发展

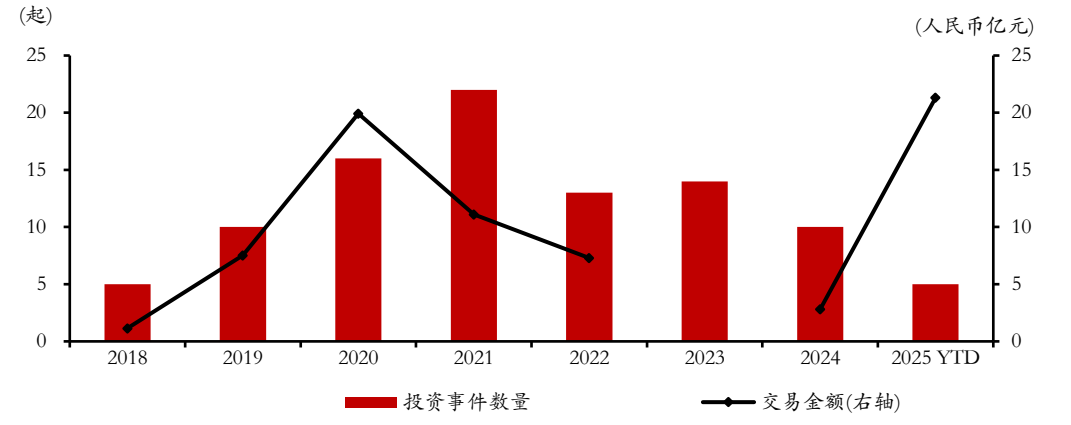
百度目前在 AI 方面的主要应用和探索主要包括：1）生成式 AI 对搜索结果的改造提升用户体验，并在中长期提升广告效果，带来新的广告形式的变现（如按销售付费）；2）AI 相关云业务的营收变现持续起量；3）个人云方面，AI 订阅相关营收持续起量（如百度网盘、百度文库 AI 相关的功能驱动订阅进一步提升）；4）昆仑芯；5）Robotaxi 业务中长期的国际化拓展给公司带来增量。

图 11：百度：AI 应用及布局

资料来源：机器之心，公司资料，招银国际环球市场
注：数据截至 2025 年 7 月

尽管百度核心广告业务的复苏可能仍需时日，但 Agent 相关广告、数字人等新的广告增长驱动力正在培育当中。我们认为，核心广告业务及在手现金是支撑公司估值的关键因素，而未来估值进一步重估的机会可能来自以下三方面：1）云业务运营指标有更清晰的能见度；2）自动驾驶出租车及 AI 相关新业务释放更多进展；3）在手现金的使用方式进一步优化以提高股东回报。

图 12：百度：对外投资事件数量及交易金额



资料来源：鲸准，招银国际环球市场

在投资布局方面，百度 2024 年至今合计对外投资 15 起，主要范围集中在 AI 相关应用领域。历史上，在 AI 芯片产业链方向，百度的主要投资包括：无问芯穹、亘存科技、赛昉科技、识光芯科 Sophoton、微核芯、星云智联，其中赛昉科技在 RISC-V 架构芯片领域布局以推动开源芯片生态建设，亘存科技在存储芯片领域持续发力，生态投资覆盖从指令集创新到存储国产化的技术护城河。

图 13：百度：近两年对外投资情况梳理及被投公司简介

被投公司简称	一句话简介	交易时间	交易轮次
生数科技	生成式人工智能基础设施及应用提供商	2025-09-19	A 轮
广州奕凌网络	一站式音视频会议解决方案提供商	2025-02-27	并购
极睿科技	全链路电商内容生成引擎研发商	2025-02-25	B+轮
YY 直播	语音视频直播平台	2025-02-25	并购
小浆搏浪	短视频软件研发商	2025-01-13	战略投资
SEELE AI	多模态大模型和应用公司	2024-11-18	Pre-A 轮
水母智能	可商用智能设计交付平台	2024-11-08	战略投资
镜象科技	心理健康服务商	2024-10-24	Pre-A 轮
地平线	AI 芯片研发商	2024-10-16	PreIPO
CreativeFitting	新一代 AIGC 超级内容平台	2024-07-01	A 轮
生数科技	生成式人工智能基础设施及应用提供商	2024-06-05	Pre-A 轮
光魔科技	互联网数据服务商	2024-05-07	战略投资
Funmangic	游戏社交开发商	2024-03-06	Pre-A 轮
百应科技	AI 解决方案提供商	2024-02-05	战略投资
幻量科技	材料信息设计和工程技术研发商	2024-02-01	战略投资

资料来源：鲸准，招银国际环球市场

图 14：百度：对外投资半导体企业情况梳理

序号	被投公司		轮次	时间
1	无问芯穹	异构算力整合平台，支持不同芯片的协同训练与推理	Pre-A 轮	2023/12/5
2	亘存科技	存储芯片（存算一体 NVM）	A 轮	2023/9/19
3	赛昉科技	处理器设计（RISC-V CPU IP）	战略投资	2023/3/23
4	识光芯科	光电子器件（SPAD-SoC 芯片），提升 AI 推理的输入数据质量	Pre-A 轮	2023/1/9
5	微核芯	处理器设计（RISC-V 服务器芯片）	战略投资	2023/1/3
6	星云智联	网络芯片（DPU / 智能网卡）	战略投资	2022/2/15

资料来源：IT 桔子，招银国际环球市场

■ **字节跳动：深度协同流量生态与 AI 技术能力**

字节跳动的 AI 业务布局核心竞争力体现在流量生态与技术落地的深度协同，其商业化路径正逐步清晰。在 C 端市场，豆包作为核心流量入口持续巩固领先地位，2025 年 6 月发布的豆包 1.6 版本支持多模态交互（文本 + 图像 + 语音），据火山引擎官方披露数据，豆包大模型日均 Tokens 使用量 2025 年 9 月已超 30 万亿。据 QuestMobile 发布的 2025 年春季数据，截至 2025 年 3 月，豆包 MAU 已突破 1.1 亿。通过“AI 全家桶”策略，覆盖通用对话、角色扮演、文生图、视频生成等场景。在视频生成赛道，即梦 AI 依托 Seaweed 模型正快速起量。在 AI 硬件领域，其推出的 AIoT 硬件如 Ola Friend 耳机 2025 年出货量已超百万台，年底有望突破千万台，形成了软件工具与硬件载体的 C 端生态。在垂直领域，其医疗板块推出的小荷 AI 医生依托抖音 9 亿用户流量池，实现科普至问诊至购药的闭环转化，并与线下医疗机构形成导流协同；教育领域的 AI 应用 Gauth 海外用户已超 2 亿，提供 AI 家教与付费导师服务，已居海外教育类应用下载榜第二。

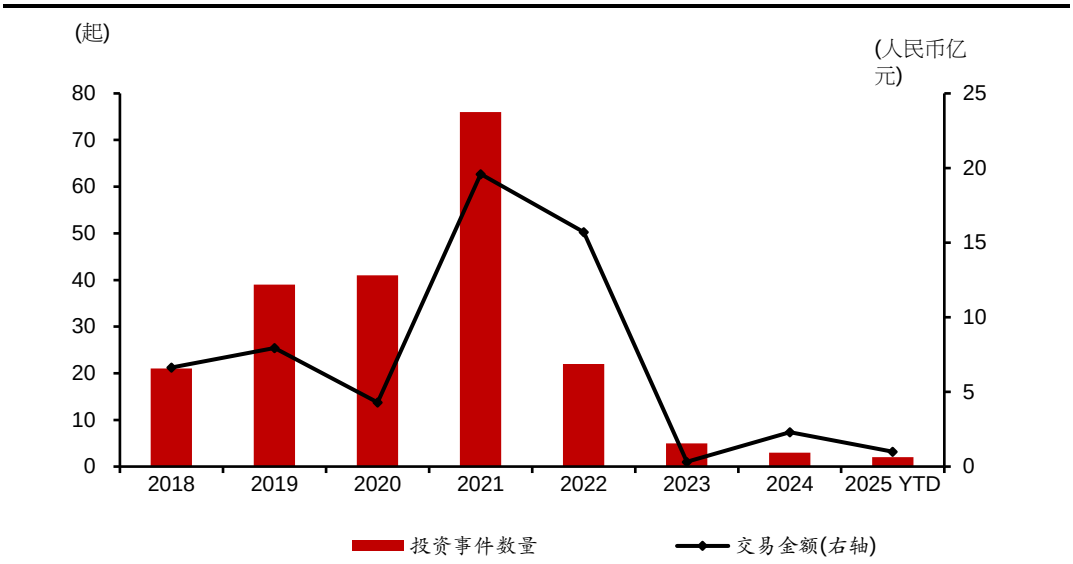
B 端业务方面，火山引擎作为 AI + 云服务的领导者之一，其 MaaS 平台集成豆包、智谱、MiniMax 等 20 余家主流模型，提供精调、评测、推理全托管服务，并为汽车、金融、教育等垂直领域提供定制化行业解决方案。HiAgent 2.0 为字节推出的企业级智能体平台，通过行业模板库帮助企业实现效率提升，同时其推出的抖音 AI 分身、TikTok AI 网红等数字人产品已在直播电商领域验证商业化潜力。

赋能自身业务方向上，在电商领域，抖音电商依托千亿级用户行为数据训练的推荐模型，AI 推荐系统使商品点击率提升 25%、转化率提高 15%，豆包大模型赋能的电商客服系统，支持多轮对话与智能应答，常见问题解决率达 92%，响应速度缩短至 1.2 秒，使得客服成本降低超 30%。在广告领域，抖音、TikTok 等超级 App 积累的用户行为数据反哺 AI 模型迭代，使广告投放触达精准度提升近 30%，同时字节旗下巨量引擎推出的 UBMax 自动化投放平台，通过深度转化模型实现一键托管，CTR 预测准确率达 89%，使得广告主 ROI 平均提升 25%，人力成本降低 50%以上；2025 年升级的 AI 创意生成 2.0 系统，可自动生成视频脚本、动态海报等素材，制作效率得到大幅提升。

在战略方向上，字节跳动于 2025 年初启动代号“Seed Edge”的 AGI 长期研究项目，聚焦自动定理证明、多模态推理等领域，目标 2030 年前实现初级 AGI。目前，Seed-OSS-36B 模型已开源，支持原生 512K 上下文长度，并引入“可控思维预算”机制，在 AIME24 数学推理评测中得分 91.7 分，接近人类竞赛水平。在算力基建领域，其与台积电合作开发自研芯片，计划 2026 年量产 5nm 训练和推理芯片，目标在相同成本下实现四倍于英伟达 H100 的算力性能。

从 2024-2025 年的投资布局来看，字节跳动主要的对外投资集中于“AI + 前沿场景”，“消费”及“AI 芯片产业链”相关的领域。

图 15：字节跳动：对外投资事件数量及交易金额



资料来源：鲸准，招银国际环球市场

图 16：字节跳动：近两年对外投资情况梳理及被投公司简介

被投公司简称	一句话简介	交易时间	交易轮次
影眸科技	人工智能与元宇宙结合应用服务商	2025-08-28	A+轮
影眸科技	人工智能与元宇宙结合应用服务商	2025-01-13	A 轮
联动优势电子商务	金融科技服务商	2024-04-03	并购
Oladance	开放式音频设备生产商	2024-04-01	并购
昕原半导体	存储芯片研发商	2024-03-13	B+轮

资料来源：鲸准，招银国际环球市场

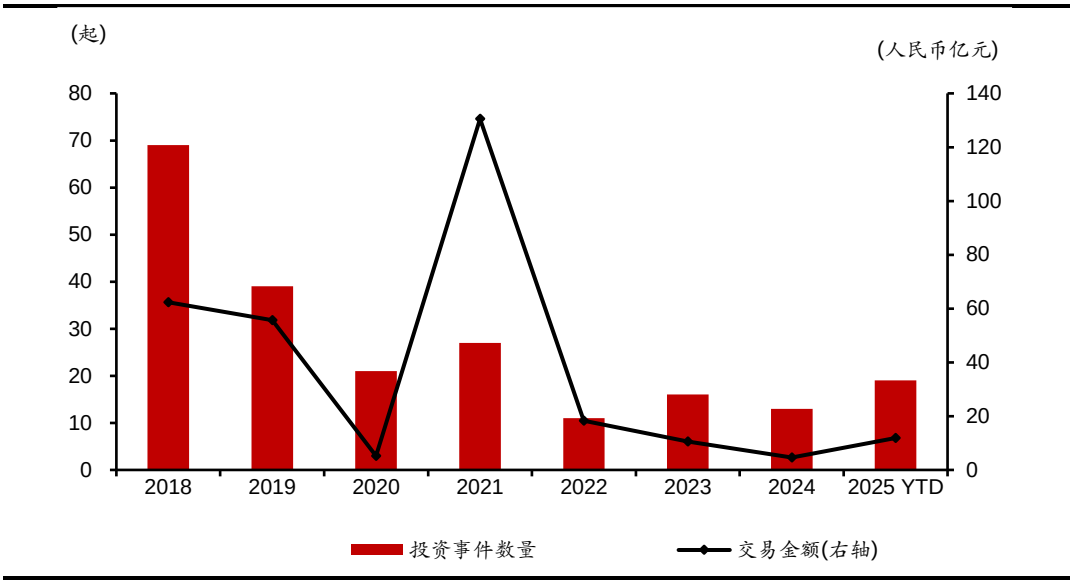
■ **蚂蚁集团：聚焦普惠场景释放 AI 价值**

蚂蚁集团此前在 2025 年 WAIC 大会集中展示了 AI 融入医疗、金融、生活场景的最新产品，其中 AI 健康管家 AQ 成为关注焦点。据公司数据，自 2024 年 9 月在支付宝小程序运行、2025 年 6 月推出独立 App 以来，AQ 累计服务用户数已超 1 亿人次。AQ 基于蚂蚁医疗大模型开发，为用户提供包括问健康、读报告、就诊咨询等超 100 项 AI 功能，直连全国超 5000 家医院、近百万医生、269 个专科医生智能体，并适配多款主流可穿戴和慢病管理设备，帮用户进行个性化健康管理。同时，蚂蚁集团在大会期间还展示了金融、生活等场景的创新应用，包括新版 AI 理财助理“蚂小财”、“AI 出行助手”、“百宝箱”一站式智能体开发平台等。

另外，蚂蚁集团在持续致力于探索 AI 智能上限、构建 AI 时代支付能力，且在积极开源、对行业开放共建 AGI 能力：1）大模型层面，蚂蚁百灵大模型家族开源发布了语言、推理、多模态三大模型系列，目前已应用于健康、金融、生活服务场景；2）全自动分布式深度学习系统 DLRover 为蚂蚁自研，可实现大模型周均有效训练时长占比大于 99%，且此技术也已经开源。面向未来，蚂蚁集团正：1）积极布局智能体领域，为 AI 大模型释放生产力做探索；2）持续搭建 AI 时代支付基础设施，推出了支付 MCP、AI 打赏、AI 钱包等产品，通过各个平台或内部研发平台，服务到各类 AI 时代的应用，推进数字支付走向数智支付。

在全球化方面，蚂蚁国际以“AI 安全性、金融服务专业性、全栈式 AI 平台”三大核心能力，向世界输出中国金融科技 AI 创新。截至目前，蚂蚁国际旗下 Alipay+、万里汇（WorldFirst）等产品现已服务 200 多个国家和地区，连接超 1 亿商户与 17 亿消费者账户。

图 17: 蚂蚁集团：对外投资事件数量及交易金额



资料来源：鲸准，招银国际环球市场

从对外投资布局的角度来看，蚂蚁 2025 年 YTD 合计对外投资 19 起，主要布局范围包括 AI 的各类应用场景及金融科技相关的场景。

图 18: 蚂蚁集团：近两年对外投资情况梳理及被投公司简介

被投公司简称	一句话简介	交易时间	交易轮次
虎鲸大药房	药品零售服务商	2025-10-14	并购
九识智能	无人物流车研发商	2025-10-13	战略投资
未来智能	AI 硬件研发商	2025-10-13	A 轮
首形科技	具身智能仿生机器人研发商	2025-09-29	A+轮
烨知芯科技	集成电路芯片研发商	2025-09-03	战略投资
昕原半导体	存储芯片研发商	2025-09-02	C 轮
灵心巧手	灵巧手技术研发商	2025-08-07	天使轮
造父智能	L4 级自动驾驶技术研发商	2025-06-23	战略投资
钛虎机器人	机器人高端硬件与机器人整体解决方案提供商	2025-06-23	战略投资
宇树科技	四足机器人与动力系统部件研发商	2025-06-19	C+轮
耀才证券金融	证券投资服务提供商	2025-04-27	并购
星尘智能	具身智能机器人研发商	2025-04-10	A 轮
旷视科技	AI 行业应用解决方案提供商	2025-04-02	E 轮及以后
超维无际	人工智能基础软件开发商	2025-03-26	天使轮
深度赋智	全自动 AI 中台提供商	2025-03-03	战略投资
星海图	具身智能机器人研发商	2025-02-20	A 轮
钱塘征信	个人征信机构	2025-02-05	战略投资
路比车险	UBI 车险与车联网平台	2025-01-09	C 轮
清微智能	人工智能芯片及解决方案提供商	2025-01-03	C 轮
云合智网	企业网络服务提供商	2024-12-17	A+轮
原力聚合	信息系统集成服务商	2024-12-16	天使轮
边塞科技	AI 大模型服务商	2024-11-20	并购
中海储能	储能电池控制系统嵌入式软件设计商	2024-10-28	战略投资
星海图	具身智能机器人研发商	2024-09-30	Pre-A 轮
好大夫在线	在线问诊综合服务平台	2024-08-15	并购

秘塔科技	智能语义数据服务提供商	2024-08-08	A 轮
JBD	Micro LED 微显示器制造商	2024-05-29	Pre-B 轮
爱诗科技	AI 视频大模型研发商	2024-04-24	A+轮
墨芯人工	AI 芯片设计商	2024-04-10	B 轮
Anext Bank	新加坡数字银行运营商	2024-03-26	战略投资
燕巢云算	云服务软件开发商	2024-01-17	战略投资
MultiSafepay	荷兰支付技术解决方案提供商	2024-01-09	并购

资料来源：鲸准，招银国际环球市场

图 19: 蚂蚁集团：对外投资半导体企业情况梳理

序号	被投公司	公司简介	轮次	时间
1	烨知芯	AI 芯片设计（端侧 AI 芯片），主要面向推理芯片市场，其产品适用于智能眼镜、手机、机器人、自动驾驶等终端设备	未披露	2025.08
2	昕原半导体	存储芯片（ReRAM 新型存储技术）	未披露	2025.08
3	云合智网	网络芯片（智能网卡、DPU）	A+轮	2024.12
4	清微智能	AI 芯片设计（FPGA/ASIC 融合架构），适用于 AI 推理场景的灵活部署	D 轮	2025.01
5	墨芯人工智能	AI 芯片设计（稀疏计算 ASIC），主要聚焦于大模型推理芯片领域	B 轮	2024.04
6	沐创集成电路	安全芯片（密码算法芯片）	B 轮	2023.11

资料来源：IT 桔子，招银国际环球市场

■ 商汤：1+X 战略布局，打造行业领先的大装置、大模型与 AI 应用

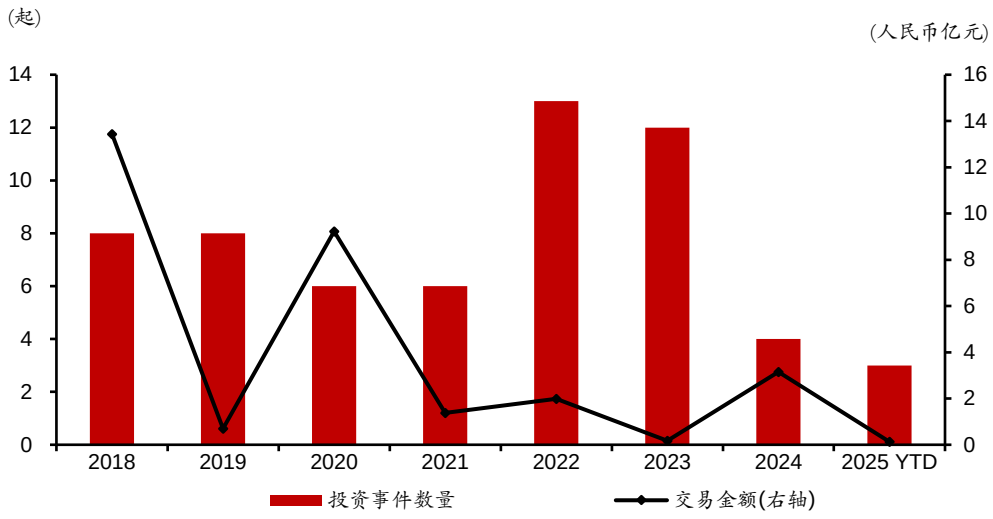
商汤在 2024 年完成战略重组，建立起“1+X”的组织架构与 AI 业务布局，其中：1）“1”代表核心业务，将专注于大装置（AI 云服务）、大模型及 AI 应用业务；2）“X”代表商汤其他生态企业，覆盖智能汽车、家庭机器人、智能医疗、智能零售等垂直 AI 赛道。

图 20：商汤：1+X 组织架构

资料来源：公司资料，招银国际环球市场

商汤自 2021 年以来投资亦逐渐减少，公司聚焦生成式 AI 业务发展，专注于打造行业领先的 AI 应用、多模态大模型以及 AI 应用。

图 21: 商汤：对外投资事件数量及交易金额



资料来源：鲸准，招银国际环球市场

商汤 2024-2025 年期间的投资主要聚焦在 AI 领域，覆盖智能驾驶、语音识别、多模态 AI 内容、空间智能等多个赛道。

图 22：商汤：近两年对外投资情况梳理及被投公司简介

被投公司简称	一句话简介	交易时间	交易轮次
天瞳威视	人工智能软硬件和智能驾驶主动感知集成解决方案提供商	2025-09-03	E 轮及以后
云知声	语音识别及语言处理技术研发商	2025-06-20	PreIPO
辉羲智能	创新车载智能计算平台	2025-01-08	A 轮
想法流	多模态 AI 内容平台提供商	2024-12-12	A 轮
智慧图	实体商业化数字经营平台	2024-05-24	战略投资
福建数产名商	互联网数据服务商	2024-05-16	战略投资
特斯联	空间智能解决方案提供商	2024-04-09	D 轮

资料来源：鲸准，招银国际环球市场

在 AI 芯片产业链方向，商汤的主要投资包括：曦望 Sunrise、北电数智、寒武纪行歌、清微智能，曦望 Sunrise 专注于高性能 GPU 研发，寒武纪行歌主要专注于智能驾驶芯片研发。

图 23：商汤：对外投资半导体企业情况梳理

序号	被投公司	芯片产业链环节	轮次	投资时间
1	曦望 Sunrise	AI 芯片设计（GPU / 多模态推理芯片）	天使轮	2025/1/7
2	北电数智	数据安全与 AI 协同（可信数据空间技术）	天使轮	2023/7/25
3	寒武纪行歌	AI 芯片设计（车载 AI 芯片）	战略投资	2022/7/1
4	清微智能	AI 芯片设计（FPGA/ASIC 融合架构），适用于 AI 推理场景的灵活部署	B 轮	2022/3/24

资料来源：IT 桔子，招银国际环球市场

OpenAI DevDay 2025: AI 应用生态加速繁荣

2025 年 10 月 6 日，OpenAI 在开发者大会（DevDay 2025）中发布多项重磅产品，涵盖智能体开发工具、多模态模型及生态平台升级，推动 AI 应用普及化。在本次大会中，Sam Altman 在演讲中公布，ChatGPT 每周用户已经从 1 亿达到超 8 亿，开发人员从 200 万翻倍至 400 万，API 平台从每分钟处理 3 亿 token 激增 20 倍，至每分钟 60 亿 token。在 ChatGPT 8 亿周活的用户基数和开发门槛持续降低同步驱动下，AI 原生应用生态有望迎来爆发。

图 24：OpenAI AI 开发者生态与影响力

资料来源：OpenAI，招银国际环球市场

本次大会重磅发布了 App Inside ChatGPT 系统，使得用户可直接在 ChatGPT 对话中调用第三方应用（如 Booking 订酒店、Figma 设计图表），同时推出 Apps SDK，开发者可构建交互式应用并触达 8 亿用户，支持 TypeScript 开发交互式应用，首批接入 Coursera、Spotify 等 7 家平台，计划年底开放应用商店目录并建立商业化分成机制。OpenAI 正逐步将 ChatGPT 打造成“万能平台”。

大会发布了 AgentKit 智能体开发套件，包含 Agent Builder，Chat Kit，Evals for Agents 几大核心组件，Agent Builder 支持拖拽式设计工作流，在演示中 8 分钟即可完成 AI Agent 开

发（含界面设计、逻辑部署），ChatKit 可将智能体嵌入应用，HubSpot、Evernote 等企业已用于客服、知识管理等场景，而新增的评估功能可以支持自动优化提示词及第三方模型测试。

大会中正式发布了编程工具 Codex，最新的 Codex 新增 Slack 集成功能，支持团队频道内直接分配编码任务。内部数据显示，OpenAI 工程师使用 Codex 后每周代码合并请求（PR）数量提升 70%，日活跃使用量增长超 10 倍。

在 API 迭代方面，大会中宣布了 GPT-5 Pro，Sora 2 & Sora 2 Pro 及 Real-Time Mini 三套模型的升级。Sora 2 API 开放视频生成能力，支持 720P（0.1 美元/秒）和 Sora 2 Pro 1080P（0.5 美元/秒）两种规格，生成速度较初代提升 300%，适用于创意到专业级场景，并新增物理引擎模拟功能，可生成符合力学规律的特效场景。GPT-5 Pro API 正式开放，专注高精度推理，推理速度较前代提升 40%，支持 40 万 token 上下文，支持纯语音交互，现场演示显示代码修改、灯光控制等复杂任务可通过语音指令完成。而轻量级语音模型 GPT-Realtime-Mini 和画图模型 GPT-Image-1-Mini 的发布，进一步推动了多模态应用门槛的降低。

在生态合作方面，OpenAI 宣布了与 AMD 达成 AI 芯片合作，其将合作部署 6GW 算力的 Instinct GPU 集群以保障算力需求，同时 OpenAI 可获得 AMD 1.6 亿股认股权证（行权后持股或达 10%，认股权证将分阶段解锁：首批随 1GW 算力部署生效，后续与算力规模扩大、AMD 股价目标及 OpenAI 技术里程碑挂钩）。在充足算力芯片供应保障下，AI 应用有望加速爆发迎来生态繁荣。

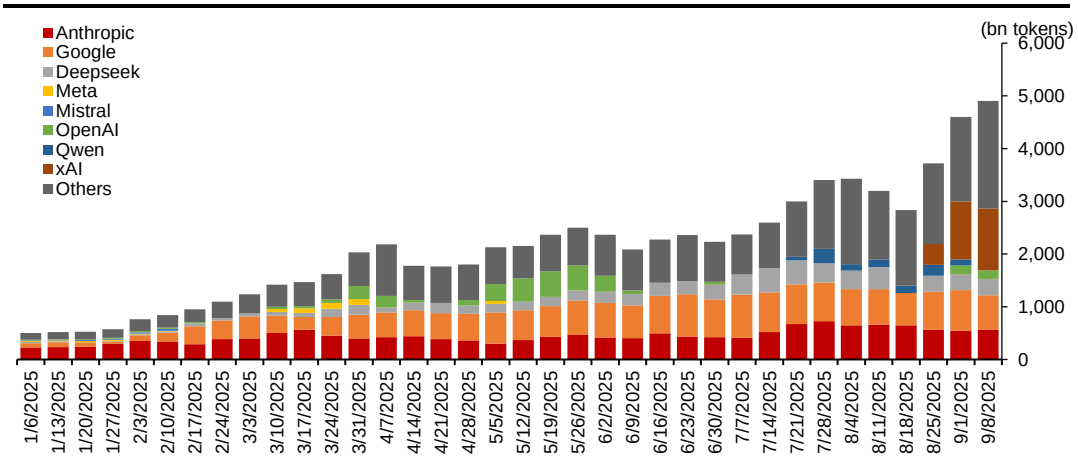
此外，10 月 13 日，OpenAI 进一步宣布了与博通的合作计划：双方拟于 2026 年推出定制数据中心芯片，并规划部署由 OpenAI 主导设计的 10GW 级 AI 加速器集群。具体来看，二者将联合研发集成博通加速器与以太网解决方案的一体化系统，核心聚焦算力的纵向堆叠（性能提升）与横向扩容（规模扩展）能力。其中，博通将承担 AI 加速器及网络系统机架的部署落地责任，项目周期计划自 2026 年下半年启动，至 2029 年底完成全量交付。该系统机架将全面基于博通以太网及配套连接技术实现扩展，核心目标为承接全球 AI 算力需求的爆发式增长，部署场景将覆盖 OpenAI 自有数据中心及合作方数据中心网络。

AI 应用生态繁荣驱动推理算力需求快速增长

■ AI 应用发展推动模型调用量与算力需求快速提升

全球主流大模型的调用量快速增长，根据 OpenRouter 数据，全球主流大模型周调用量由 2025 年初的 4997 亿增长至 2025 年 9 月的约 4.9 万亿 Token，9 个月内增长近 10 倍，我们认为主要得益于：**1）AI 应用的快速落地：**编程、广告营销、搜索等多个场景的 AI 应用落地带来 Token 调用量的快速增长，谷歌平台总 Token 调用量在 2025 年 5 月/7 月/10 月分别达到 480 万/960 万/1,300 万亿，主因 AI Overview 和 Gemini App 等 AI 应用渗透率的快速提升；**2）智能体应用渗透率提升：**由于智能体调用通常会将一个任务拆分成多个步骤并涉及多个工具的使用，智能体调用的 Token 量相较于 ChatBot 对话调用的 Token 量有数倍的提升，智能体应用渗透率的提升也将推动 Token 调用量的快速增长；**3）图像与视频大模型的调用增加：**图像与视频大模型训练及推理的 Token 调用量相较文本大模型均有数倍的增加，随着未来图像与视频大模型的持续迭代以及使用率提升，预计总 Token 调用量及算力需求均会有显著增长。

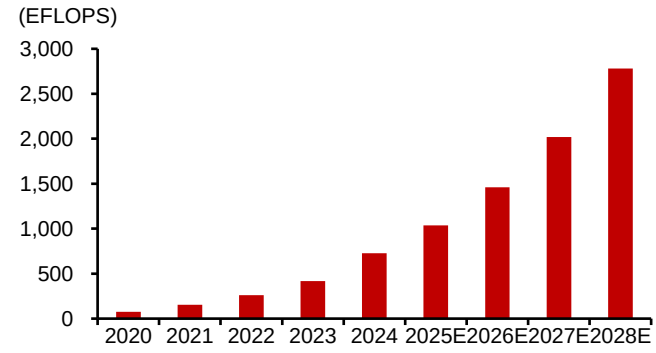
图 25：全球主流大模型：调用量趋势



资料来源：OpenRouter，招银国际环球市场

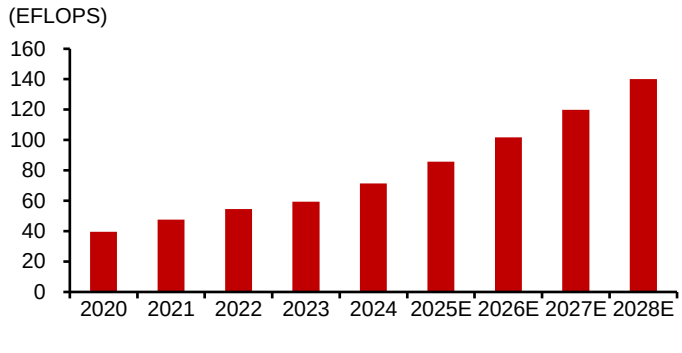
聚焦中国市场，根据 IDC 数据，中国智能算力（FP16）/通用算力（FP64）规模预计将由 2025 年的 1,037/86 EFLOPS 增长至 2028 年的 2,782/140 EFLOPS，2025-2028E 复合增长率分别达到 39%/18%。

图 26：中国智能算力规模



资料来源：IDC，招银国际环球市场

图 27：中国通用算力规模

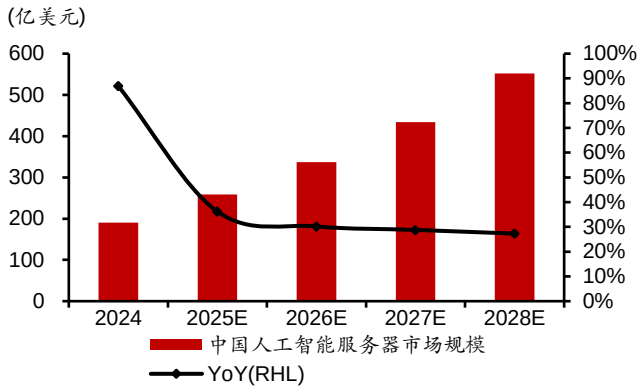


资料来源：IDC，招银国际环球市场

■ 算力需求增长带动推理芯片需求

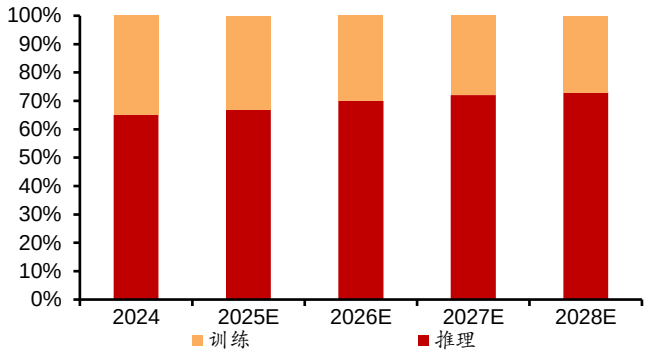
在 AI 算力需求快速增长的带动下，IDC 预计中国智能服务器市场规模将从 2025 年的 259 亿美元增长至 2028 年的 552 亿美元，复合增速为 29%。其中，随着 AI 应用及智能体的渗透率持续提升，推理端的算力需求增长将成为服务器需求增长的主要驱动力，IDC 预计 AI 服务器的推理负载占比将从 2024 年的 65%增长至 2028 年的 73%。

图 28：中国智能服务器市场规模



资料来源：IDC，招银国际环球市场

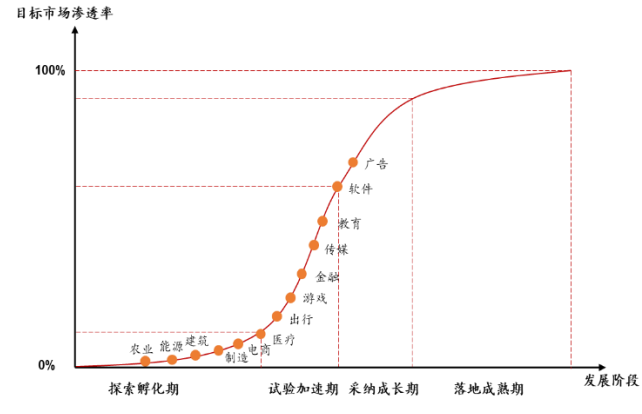
图 29：中国智能服务器工作负载占比



资料来源：IDC，招银国际环球市场

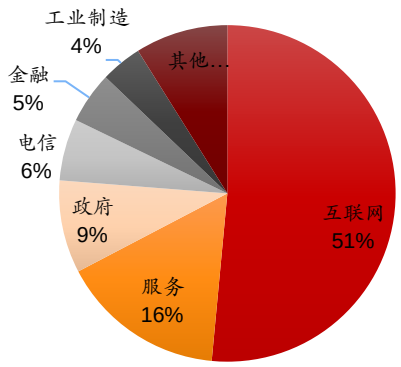
当前阶段，软件等科技行业的 AI 应用渗透率较高，互联网行业占中国 AI 服务器出货量 51%。传统行业如农业、能源、建筑、制造等行业的 AI 应用渗透率仍然较低（<10%），未来传统行业 AI 应用落地以及渗透率的提升有望带动增量算力需求。

图 30：主要行业 AI 应用渗透率趋势



资料来源：甲子光年，招银国际环球市场

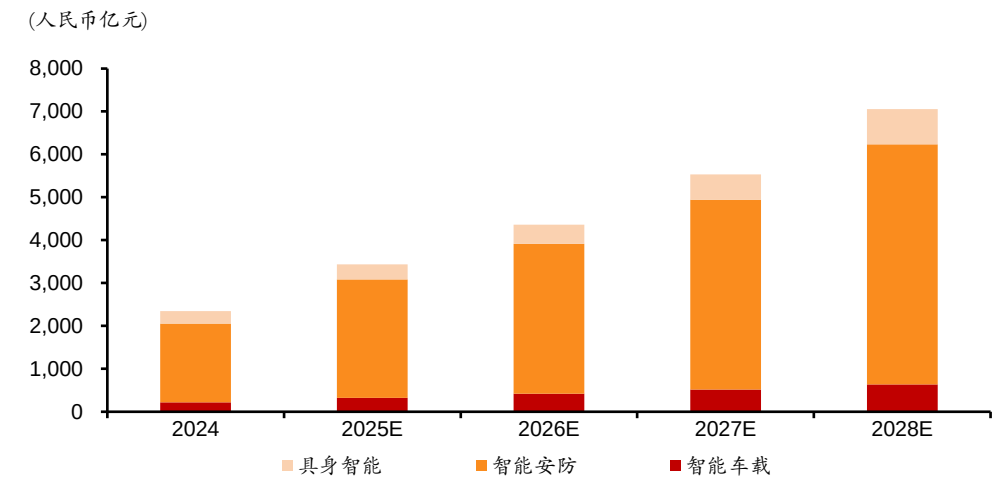
图 31：中国 AI 服务器出货量行业分布（2024）



资料来源：甲子光年，招银国际环球市场

包括智能安防、智能车载、具身智能等端侧 AI 行业未来有望迎来快速发展，带动算力与 AI 芯片需求。根据弗若斯特沙利文与头豹研究院数据，中国智能安防/具身智能/智能车载行业 2028 年市场规模有望达到 5,598/815/640 亿元，2025-2028 年复合增速分别为 25%/27%/33%。

图 32: 端侧 AI 行业快速增长

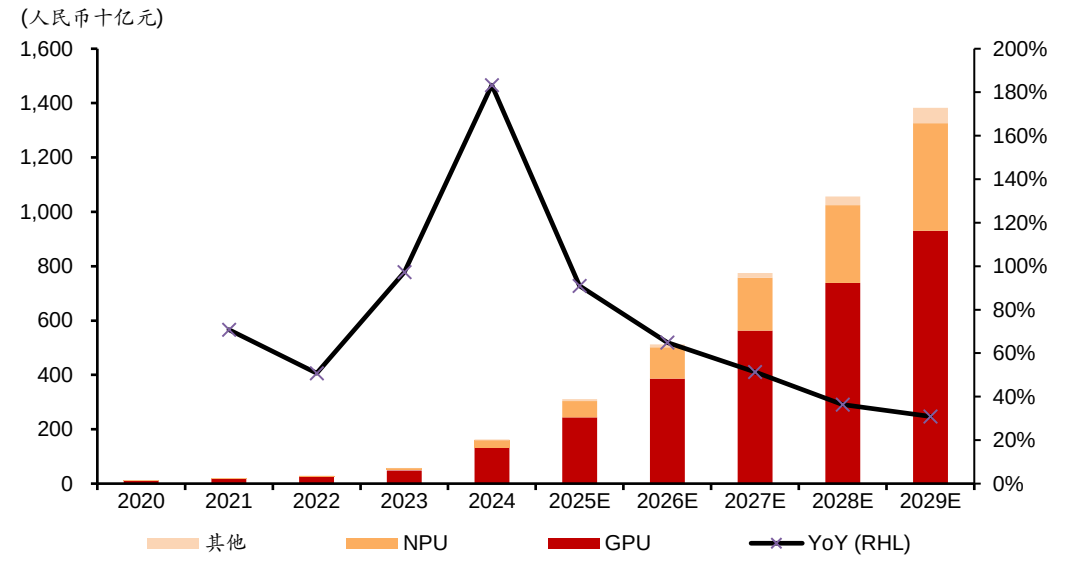


资料来源：弗若斯特沙利文，头豹研究院，招银国际环球市场

■ 国内推理芯片产业链有望迎来快速发展

目前大部分 AI 模型的推理运算仍然由 GPU 完成，但展望未来，专门面向 AI 推理而设计和优化的 NPU 有望占据 AI 推理芯片更大的市场份额。相较于 GPU，NPU 在 AI 推理计算中具备更显著的功耗和成本优势。根据 CIC 和 IDC 预测，中国 AI 推理芯片市场规模将从 2025 年的 3106 亿元增长至 2029 年的 1.38 万亿元，复合增速为 45%。其中专为推理设计的 NPU 占比将从 2025 年的 19%增长至 2029 年的 29%。

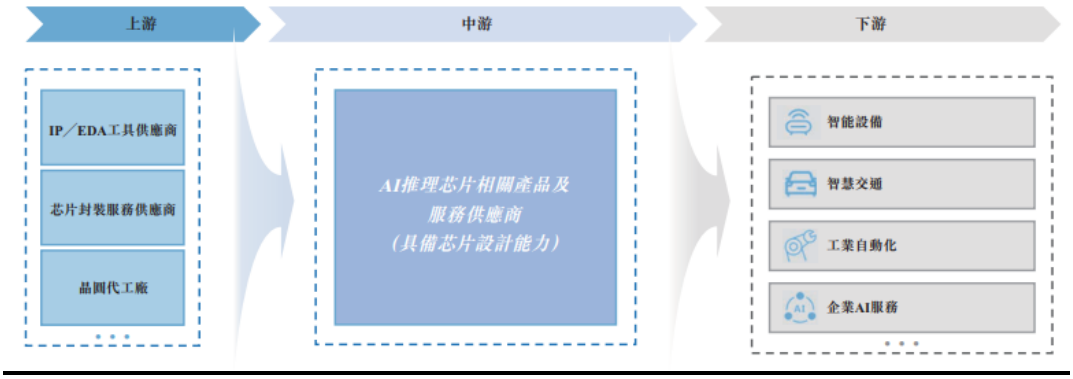
图 33: 中国：AI 推理芯片市场规模



资料来源：CIC，IDC，招银国际环球市场

中国 AI 推理芯片产业链上游主要包括 IP/EDA 工具供应商（ARM、Synopsys、Cadence、芯原股份）、芯片封装和测试公司（长电科技、通富微电、华天科技）以及晶圆体代工厂（台积电、中芯国际等）。中游则主要涉及 AI 推理芯片相关产品及服务供应商，利用创新能力和技术积累进行电路设计与验证，包括领先的 GPU 及 NPU 设计厂商（英伟达、AMD、寒武纪、云天励飞等）。下游则主要包括各个不同场景与行业的客户。

图 34：中国：AI 推理芯片产业链



资料来源：CIC，招银国际环球市场

在当前 AI 发展阶段，大部分模型的训练和推理仍然由 GPU 完成，但随着针对推理场景研发的 ASIC 芯片持续发展（包括 TPU、NPU 等），在推理场景中已经有很多 ASIC 芯片展现出更好的能效比。

海外市场，谷歌、亚马逊、Meta 等头部 AI 和云厂商持续加大自研 ASIC 芯片布局，以提升云端 AI 推理效率。谷歌 2025 年推出专为 AI 推理研发的 TPU v7，FP8 峰值算力达到 4614 TFlops，峰值能耗比达到 29.3 TFlops/W，能效比显著好于英伟达 B200 芯片。

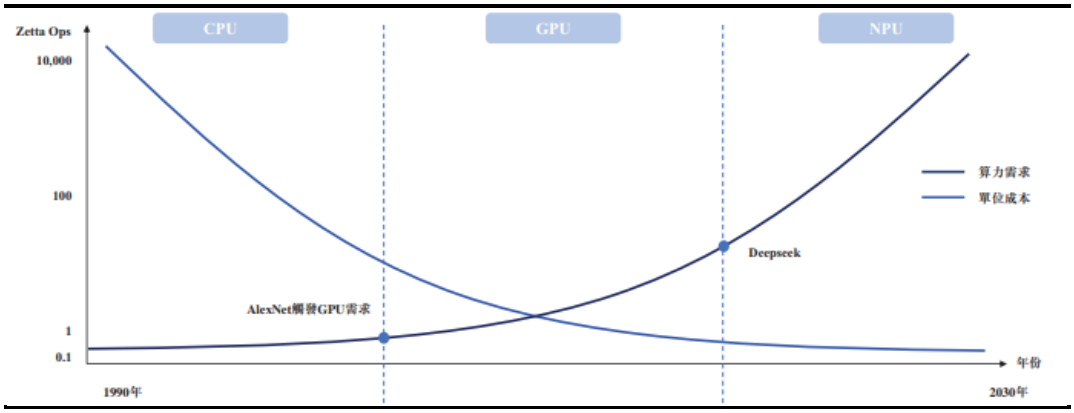
国内市场，华为及寒武纪等头部芯片厂商已推出具备较高能效比的 ASIC 推理芯片，能够较好的覆盖主流大模型的推理场景。此外，部分初创芯片与 AI 公司亦有在 ASIC 推理芯片领域布局，在低成本的推理场景（如边缘和终端推理）实现了对 GPU 芯片的替代。

图 35：AI 芯片对比

厂商	芯片	类型	制程	算力(FP16)	功耗	主要覆盖场景
海外						
英伟达	B200	GPU	4nm	2250TFLOPS	1000W	前沿模型训练、主流模型训练、前沿模型推理
英伟达	H20	GPU	4nm	148TFLOPS	400W	主流模型训练、主流模型推理
谷歌	TPU v7	ASIC	/	2300TFLOPS	160W	前沿模型推理、主流模型推理
Amazon	Trainium3	ASIC	3nm	1310TFlops	730W	前沿模型推理、主流模型推理
Meta	MTIA V2	ASIC	5nm	177TFLOPS	90W	前沿模型推理、主流模型推理
国内						
华为	昇腾 910B	ASIC	7nm	320TFLOPS	310W	主流模型训练、主流模型推理
寒武纪	MLU370-X8	ASIC	7nm	96TFLOPS	150W	主流模型训练、主流模型推理
壁仞科技	BR100	GPU	7nm	1024TFLOPS	550W	主流模型训练、主流模型推理
摩尔线程	MTT S4000	GPU	/	100TFLOPS	450W	主流模型训练、主流模型推理
百度	昆仑芯 P800	GPU	/	/	450W	主流模型训练、主流模型推理
天数智芯	天垓 150	GPU	7nm	192 TFLOPS	350W	主流模型训练、主流模型推理
燧原	云燧 i20	ASIC	12nm	128 TFLOPS	/	主流模型推理、边缘与终端推理
沐曦	曦思 N100	GPU	7nm	80 TFLOPS	70W	主流模型推理、边缘与终端推理
云天励飞	DeepEdge 10	ASIC	14nm	8 TFLOPS	40W	边缘与终端推理
晶晨股份	晶晨 A311D	ASIC	12nm	5 TFLOPS	/	边缘与终端推理

资料来源：公司资料，招银国际环球市场

图 36: NPU 等 ASIC 芯片有望在推理场景中逐渐替代 GPU



资料来源：CIC，招银国际环球市场

Sora 2：加速 AI 视频应用生态发展

2025 年 9 月 30 日，OpenAI 正式上线视频与语音生成模型 Sora 2，并且推出了由 Sora 2 支持的视频社交应用 Sora App。OpenAI 将 Sora 模型类比为视频模型的 GPT-1.0 时刻，而 Sora 2 模型则来到了 GPT-3.5 时刻，相较于 Sora 模型在多方面能力上实现提升：1）更加遵循真实物理定律，改进了过往视频模型为了过度贴合提示词而不遵守物理定律的问题；2）可控性显著提升，模型能够在保持视频环境一致的前提下遵循复杂的提示词或者用户的修改要求；3）通用的视频+音频生成系统，可以生成复杂且具备高真实度的背景声与人物发言等。

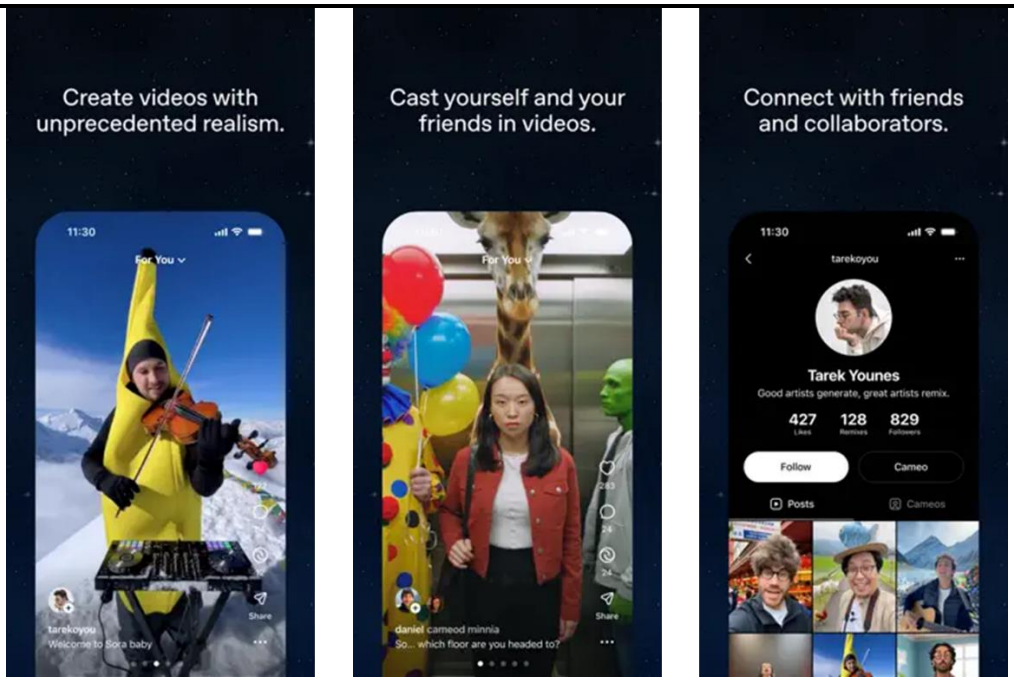
图 37: Sora 2 模型 vs Sora 模型

公司	视频效果演示	Sora 2	Sora
物理规律遵循度	 Prompt: a group people playing volleyball	物理定律遵循度高	视频模型经常过于乐观，它们会为了执行文本提示词，导致物体变形并扭曲现实。
可控性	 Prompt: in the style of a studio ghibli anime, a boy and his dog run up a grassy scenic mountain with gorgeous clouds, overlooking a village in the distant background	可控性强	视频模型可控性较低，修改提示词后可能会导致视频环境和风格发生较大变化
音频生成	 Prompt: underwater scuba diver, sounds of the coral reef	可生成复杂且具备高真实度的背景声与人物发言等	仅生成视频，无法生成音频

资料来源：公司资料，招银国际环球市场

除了 Sora 2 模型，OpenAI 还发布了视频社交应用 Sora App，除了常见的视频信息流功能以外，Sora App 还结合 AI 提供新的视频社交玩法 Cameos：用户只要在应用中完成一次简短的音视频录制（用于验证身份并捕捉用户形象），就能以高保真度置身于 Sora 生成的视频中。

图 38: Sora App 主要功能



资料来源：公司资料，招银国际环球市场

Sora App 在美国及加拿大发布后的一周内快速登顶了美国 iOS 免费应用榜。从 Sora App 目前的产品定位来看（C 端用户为主，视频生成主要用于 C 端用户间的社交互动），Sora App 或将影响社交行业的用户习惯和市场格局，短期可能对于传统社交平台（由用户产出内容为主的社交平台，如 Instagram/TikTok 等）带来一定的竞争压力，但长期的竞争影响或较为有限，AI 或难以颠覆头部社交平台目前已经形成的竞争壁垒，包括用户规模、社交网络效应以及用户数据优势，同时社交平台在长期亦可跟进相关的 AI 视频生成功能以消除在 AI 应用方面的差异。

另外，Sora App 对于 AI 原生视频创作应用（如可灵 AI、Runway）的竞争压力亦较为有限：1）Sora App 主要面向普通 C 端用户，而可灵 AI 主要面向专业创作者；2）Sora App 的产出视频主要用于用户间的社交互动，而可灵 AI 旨在为创作者提供视频生成工具以用于社交、广告、专业视频等多场景。

图 39: Sora 2 模型 vs Sora 模型

	Sora App	可灵 AI	社交/短视频平台(Instagram/TikTok 等)
用户群	C 端用户为主	PUGC 和 B 端用户为主	C 端用户为主
视频生态	AI 视频生成	AI 视频生成	用户产出为主
使用场景	社交娱乐	社交娱乐、广告营销，专业内容制作	社交娱乐及内容消费
变现方式	探索中，或考虑 IP 变现收入分成模式	订阅制会员为主	广告、会员、电商等

资料来源：公司资料，招银国际环球市场

我们持续看好中长期全球 AI 创意应用的市场空间。我们测算 2027 年全球创意应用市场规模有望达到 546 亿美元，其中 AI 创意应用渗透率有望达到 11%，对应全球 AI 创意工具市场规模有望达到 58 亿美元。

图 40：AI 创意应用市场规模测算

创意工具 TAM (亿美元)	2024	2025E	2026E	2027E
创意工具 TAM	**391.6**	**437.9**	**489.3**	**546.1**
YoY		11.8%	11.7%	11.6%
AI 创意工具市场规模		**18.6**	**40.1**	**58.0**
AI 渗透率		4.3%	8.2%	10.6%

职业创意工作者	2024	2025E	2026E	2027E
全球职业创意工作者 (百万)	68.0	71.4	75.0	78.7
付费渗透率	85%	85%	85%	85%
付费用户数（百万）	57.8	60.7	63.7	66.9
月 ARPPU (美元)	44.0	45.0	46.0	47.0
创意工具 TAM - 职业创意工作者	**305.2**	**327.7**	**351.8**	**377.4**
AI 创意工具市场规模 - 职业创意工作者		**13.1**	**26.4**	**37.7**
AI 渗透率		4%	8%	10%

业余创意爱好者	2024	2025E	2026E	2027E
全球业余创意爱好者 (百万)	900.0	927.0	954.8	983.5
付费渗透率	10%	11%	12%	13%
付费用户数（百万）	90.0	102.0	114.6	127.8
月 ARPPU (美元)	8.0	9.0	10.0	11.0
创意工具 TAM - 业余创意工作者	**86.4**	**110.1**	**137.5**	**168.8**
AI 创意工具市场规模 - 业余创意工作者		**5.5**	**13.7**	**20.3**
AI 渗透率		5%	10%	12%

资料来源：公司资料、国际劳工组织、招银国际环球市场

我们看好可灵 AI 在全球 AI 视频应用赛道维持领先，主因其 1）技术优势：快手 9 月推出可灵 2.5 Turbo 视频生成模型，对比 Seedance 1.0/Veo3-Fast 的文生视频效果胜负比达到 1.6/2.1，上线后快速登顶 Artificial Analysis 文生视频榜单；2）产品优势：可灵 AI 自产品发布以来已经迭代超 30 次，产品能力与体验始终保持领先；3）用户规模及商业化优势：可灵 AI 创作者已达 4500 万，发布至今已生成 2 亿视频和 4 亿图片，2Q25 可灵 AI 营收已突破 2.5 亿元，商业化的跑通让公司能够形成产品迭代-收入增长的正循环。

图 41: Artificial Analysis：文生视频榜单

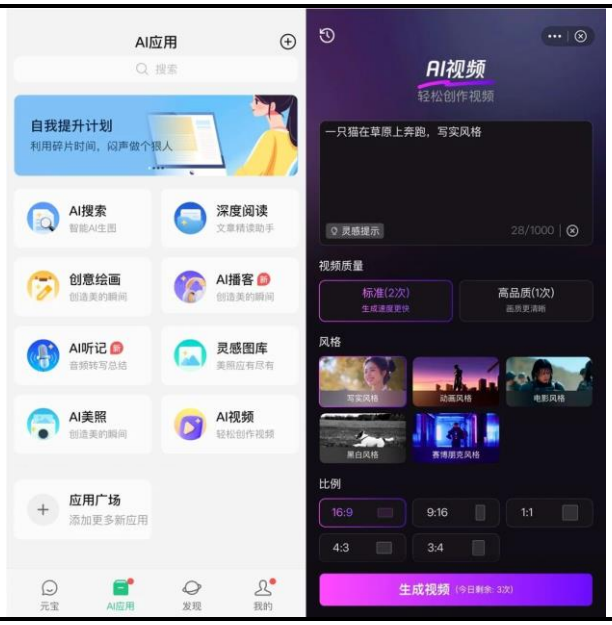
	Creator	Model	ELO	95% CI	Appearances	Release Date
1	Kuaishou KlingAI	Kling 2.5 Turbo 1080p	1,245	-13/+12	4,070	Sept 2025
2	Google	Veo 3 (No Audio)	1,229	-9/+9	8,833	Jul 2025
3	Luma Labs	Ray 3	1,215	-11/+12	4,386	Sept 2025
4	MiniMax	Hailuo 02 Standard	1,203	-9/+10	6,673	Jun 2025
5	Alibaba	Wan 2.5 Preview	1,189	-11/+12	4,057	Sept 2025
6	Bytedance	Waver 1.0	1,187	-8/+7	13,610	Aug 2025
7	Google	Veo 3 Fast Preview (No Audio)	1,187	-7/+7	14,598	Jun 2025
8	MiniMax	Hailuo 02 Pro	1,184	-9/+9	6,789	Jun 2025
9	PixVerse	PixVerse V5	1,184	-9/+10	6,826	Aug 2025
10	ByteDance Seed	Seedance 1.0	1,174	-10/+10	6,285	Jun 2025

资料来源：Artificial Analysis，招银国际环球市场

随着文生视频/图生视频模型能力的提升，国内互联网各赛道公司包括腾讯、阿里巴巴和百度等或也将持续加强 AI 视频应用的布局，从而提升用户体验，提高内容创作效率。同时，当前基于 Diffusion 框架的视频模型的推理成本和算力需求对比文本大模型要高出数倍，AI 视频应用的普及或将带来推理算力需求的快速增长。

腾讯混元视频生成大模型是目前开源模型中参数最多、性能最强的文生视频大模型之一。模型参数量 130 亿，可供企业与个人开发者免费使用，模型目前上线了腾讯元宝 APP。未来随着视频大模型能力提升，文生视频能力有望接入到更多腾讯生态应用，从而提升用户体验。

图 42: 腾讯：混元 AI 视频生成



资料来源：元宝，招银国际环球市场

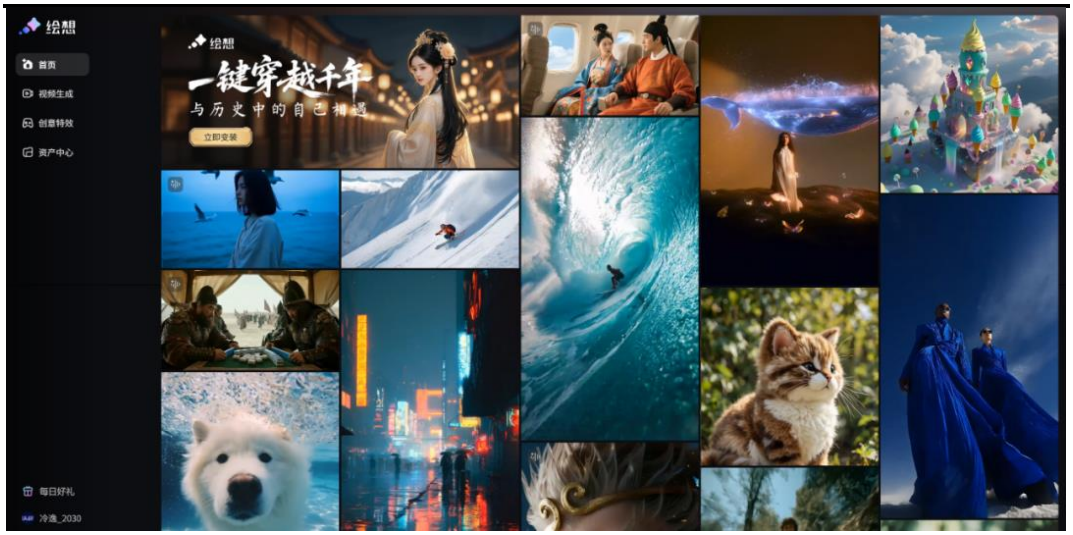
阿里巴巴云栖大会发布通义万相 Wan2.5-preview 系列模型涵盖文生视频、图生视频、文生图和图像编辑四大模型，其中视频生成模型能生成和画面匹配的人声、音效和音乐 BGM，首次实现音画同步的视频生成能力，进一步降低电影级视频创作的门槛。此外，本次模型发布还升级了图像生成能力，可生成中英文文字和图表，支持图像编辑功能，输入一句话即可完成 P 图。

图 43: 阿里巴巴：通义万相 Wan 模型家族系列产品图谱

资料来源：公司资料，招银国际环球市场

百度蒸汽机（MuseSteamer）是行业首个中文音视频一体化生成的 I2V 模型，支持环境音效和多角色语音的一体化生成。同时，百度蒸汽机的视频生成技术能实现语音与唇形、表情、动作的毫秒级对齐，即使在复杂场景下也依然稳定。此外，百度蒸汽机首创多模态潜在空间规划技术（Latent Multi-Modal Planner），在该技术支持下，蒸汽机能够自主协调多角色身份、情感与互动逻辑，保障叙事连贯性。

图 44: 百度：应用蒸汽机模型的“绘想”平台



资料来源：公司资料，招银国际环球市场

免责声明及披露

分析员声明

负责撰写本报告的全部或部分内容之分析员，就本报告所提及的证券及其发行人做出以下声明：（1）发表于本报告的观点准确地反映有关于他们个人对所提及的证券及其发行人的观点；（2）他们的薪酬在过往、现在和将来与发表在报告上的观点并无直接或间接关系。

此外，分析员确认，无论是他们本人还是他们的关联人士（按香港证券及期货事务监察委员会操作守则的相关定义）（1）并没有在发表研究报告 30 日前处置或买卖该等证券；（2）不会在发表报告 3 个工作日内处置或买卖本报告中提及的该等证券；（3）没有在有关香港上市公司内任职高级人员；（4）并没有持有有关证券的任何权益。

招银国际环球市场投资评级

买入 ：股价于未来 12 个月的潜在涨幅超过 15%
持有 ：股价于未来 12 个月的潜在变幅在-10%至+15%之间
卖出 ：股价于未来 12 个月的潜在跌幅超过 10%
未评级 ：招银国际证券并未给予投资评级

招银国际环球市场行业投资评级

优于大市 ：行业股价于未来12 个月预期表现跑赢大市指标
同步大市 ：行业股价于未来12 个月预期表现与大市指标相若
落后大市 ：行业股价于未来 12 个月预期表现跑输大市指标

招银国际环球市场有限公司

地址: 香港中环花园道 3 号冠君大厦 45 楼 **电话: (852) 3900 0888** **传真: (852) 3900 0800**

招银国际环球市场有限公司(“招银国际环球市场”)为招银国际金融有限公司之全资附属公司 (招银国际金融有限公司为招商银行之全资附属公司)

重要披露

本报告内所提及的任何投资都可能涉及相当大的风险。报告所载数据可能不适合所有投资者。招银国际环球市场不提供任何针对个人的投资建议。本报告没有把任何人的投资目标、财务状况和特殊需求考虑进去。而过去的表现亦不代表未来的表现，实际情况可能和报告中所载的大不相同。本报告中所提及的投资价值或回报存在不确定性及难以保证，并可能会受目标资产表现以及其他市场因素影响。招银国际环球市场建议投资者应该独立评估投资和策略，并鼓励投资者咨询专业财务顾问以便作出投资决定。

本报告包含的任何信息由招银国际环球市场编写，仅为本公司及其关联机构的特定客户和其他专业人士提供的参考数据。报告中的信息或所表达的意见皆不可作为或被视为证券出售要约或证券买卖的邀请，亦不构成任何投资、法律、会计或税务方面的最终操作建议，本公司及其雇员不就报告中的内容对最终操作建议作出任何担保。我们不对因依赖本报告所载资料采取任何行动而引致之任何直接或间接的错误、疏忽、违约、不谨慎或各类损失或损害承担任何的法律责任。任何使用本报告信息所作的投资决定完全由投资者自己承担风险。

本报告基于我们认为可靠且已经公开的信息，我们力求但不担保这些信息的准确性、有效性和完整性。本报告中的资料、意见、预测均反映报告初次公开发布时的判断，可能会随时调整，且不承诺作出任何相关变更的通知。本公司可发布其它与本报告所载资料及/或结论不一致的报告。这些报告均反映报告编写时不同的假设、观点及分析方法。客户应该小心注意本报告中所提及的前瞻性预测和实际情况可能有显着区别，唯我们已合理、谨慎地确保预测所用的假设基础是公平、合理。招银国际环球市场可能采取与报告中建议及/或观点不一致的立场或投资决定。

本公司或其附属关联机构可能持有报告中提到的公司所发行的证券头寸并不时自行及/或代表其客户进行交易或持有该等证券的权益，还可能与这些公司具有其他投资银行相关业务联系。因此，投资者应注意本报告可能存在的客观性及利益冲突的情况，本公司将不会承担任何责任。本报告版权仅为本公司所有，任何机构或个人于未经本公司书面授权的情况下，不得以任何形式翻版、复制、转售、转发及或向特定读者以外的人士传阅，否则有可能触犯相关证券法规。

如需索取更多有关证券的信息，请与我们联络。

对于接收此份报告的英国投资者

本报告仅提供给符合(I)不时修订之英国 2000 年金融服务及市场法令 2005 年(金融推广)令(“金融服务令”)第 19(5) 条之人士及(II) 属金融服务令第 49(2) (a) 至(d) 条(高净值公司或非公司社团等)之机构人士，未经招银国际环球市场书面授权不得提供给其他任何人。

对于接收此份报告的美国投资者

招银国际环球市场不是在美国的注册经纪交易商。因此，招银国际环球市场不受美国就有关研究报告准备和研究分析员独立性的规则的约束。负责撰写本报告的全部或部分内容之分析员，未在美国金融业监管局（“FINRA”）注册或获得研究分析师的资格。分析员不受旨在确保分析师不受可能影响研究报告可靠性的潜在利益冲突的相关 FINRA 规则的限制。本报告仅提供给美国 1934 年证券交易法 (经修订) 规则 15a-6 定义的 “主要机构投资者”，不得提供给其他任何个人。接收本报告之行为即表明同意接受协议不得将本报告分发或提供给任何其他人。接收本报告的美国收件人如想根据本报告中提供的信息进行任何买卖证券交易，都应仅通过美国注册的经纪交易商来进行交易。

对于在新加坡的收件人

本报告由 CMBI (Singapore) Pte. Limited （CMBISG）（公司注册号 201731928D）在新加坡分发。CMBISG 是在《财务顾问法案》（新加坡法例第 110 章）下所界定，并由新加坡金融管理局监管的豁免财务顾问公司。 CMBISG 可根据《财务顾问条例》第 32C 条下的安排分发其各自的外国实体，附属机构或其他外国研究机构篇制的报告。 如果报告在新加坡分发给非《证券与期货法案》（新加坡法例第 289 章）所定义的认可投资者，专家投资者或机构投资者，则 CMBISG 仅会在法律要求的范围内对这些人士就报告内容承担法律责任。 新加坡的收件人应致电（+65 6350 4400）联系 CMBISG，以了解由本报告引起或与之相关的事宜。