

人工智能研究框架： 大模型白热化，应用加速分化

行业投资评级：强于大市 | 维持

陈涵泊/李佩京/王思

中邮证券研究所 人工智能团队

中邮证券

发布时间：2025-09-03

➤ 大模型：加速多模态研发，闭源模型逐步逆袭开源。

- 多模态技术路线尚未收敛，国内外大模型厂商持续刷新SOAT。图片领域，GPT-4o图像生成功能引发热潮后，谷歌Gemini 2.5 Flash Image登顶多主流图像榜，Meta也从Midjourney授权AI图像模型；视频领域，阿里字节等刷新能力高度，谷歌将视频模型推进至实时交互通用世界模型阶段。
- Deepseek开源浪潮推动模型平权与应用加速，浪潮后闭源模型逐渐维持性能领先，且借开源策略打造生态入口。一方面，预训练模型Scaling Law增速放缓使闭源与开源模型性能差距收窄，但OpenAI、谷歌等闭源厂商凭资源、人才与技术积累仍领先；另一方面，闭源厂商推行“开源+闭源”协同，保留主模型闭源优势的同时，开源上一代或轻量模型，激活开发者并反哺自身闭源模型以维持领先。

➤ 算力：海内外资本开支提振，芯片厂加速推出新版本芯片。

- 海外微软、谷歌、Meta、亚马逊四者合计资本开支在2023Q3后连续多个季度环比上升，同时上调资本开支指引；国内CSP厂商在大模型取得重要突破、比肩海外之时，也需通过大量资本开支缩小与海外差距。
- 为了满足海内外CSP资本开支放量需求，GPU、ASIC竞相角逐新品研发。1) 传统芯片大厂：英伟达芯片架构正在由Blackwell向Rubin过渡；AMD MI350系列对标GB200，MI400系列计划为26年新型服务器“Helios”的基础；华为云CloudMatrix 384通过多卡互联实现性能弯道超车。2) 云端巨头自研ASIC：根据TrendForce集邦咨询，AI Server需求带动北美四大CSP加速自研ASIC芯片，平均1~2年就会推出升级版本；国内阿里、百度、腾讯等亦推出自研ASIC方案。

➤ 应用：海外应用加速分化，聚焦与数据结合的复杂toB场景。

复盘今年美股AI应用表现，行情有所分化，股价涨幅较好的公司体现在：具备显著的行业Know-how壁垒，同时AI赋能业务放量增长。例如，以PLTR为代表的toB企业，其股价走强的核心逻辑在于AI产品深度赋能主业（如AI提升服务效率，或拓展高价值客户场景），直接推动营收规模快速增长，这一特征在toB领域因商业场景更易落地而表现得尤为显著。to C企业中，创意类挤压工具类、AI替代性显现，多模态大模型推动其从“辅助工具”进化为懂需求、能创方案的“创意伙伴”，既强化替代优势，又以低学本、快产出吸引客户，逼近传统专业工具价值核心。

➤ 投资建议：建议关注：1) AI Agent：鼎捷数智、视源股份、泛微网络、致远互联、金蝶国际、用友网络、汉得信息、赛意信息、迈富时、光云科技、税友股份、润达医疗、同花顺、恒生电子、金山办公、万兴科技、福昕软件、科大讯飞、金桥信息、卓易信息、普元软件等；2) AI终端：云天励飞、虹软科技、小米集团、联想集团、科大讯飞、乐鑫科技、中科蓝讯等；3) AI算力：寒武纪、云天励飞、亿都（国际控股）、工业富联、浪潮信息、曙光数创、超讯通信、华丰科技、神州数码、软通动力、烽火通信、广电运通、拓维信息、四川长虹等；4) AI多模态：万兴科技、虹软科技、当虹科技、中科创达、大华股份、海康威视、漫步者、萤石网络、汉仪股份、美图公司、云从科技等。

➤ 风险提示：地缘政治冲与贸易风险、技术迭代不及预期、国内AI Agent商业化进展不及预期、算力供应不及预期、行业竞争加剧风险等。



目录

- 一** | **大模型：加速多模态研发，闭源模型逐步逆袭开源**
- 二** | **算力：海内外资本开支提振，芯片厂加速推出新版本芯片**
- 三** | **应用：海外应用加速分化，聚焦与数据结合的复杂toB场景**
- 四** | **投资建议与风险提示**



大模型：加速多模态研发，闭源模型逐步逆袭开源

- 1.1 多模态尚未，国内需求刷新 SOAT
- 1.2 开源浪潮后，闭源模型性价比

1.1.1 多模态技术尚未收敛，生成视频、多模态实时交互为未来方向

- 多模态大模型融合了多种感知路径和表达形态，能够同时处理文本、图像、语音等多种数据，并进行深度的语义理解和交叉模态处理，具备深度人机交互和全面智能应用的潜力。当前，多模态大模型主要分为理解和生成两种类型，技术路线尚未收敛，成为国内外大模型厂商重点突破方向之一。
- **多模态理解模型**：多模态理解模型对齐视觉特征与文本特征实现跨模态的统一理解，分为以下两类技术路线：1) 一方面，基于语言大模型底座，配合多类外部专家模型共同实现多模态处理；2) 另一方面，通过跨模态特征对齐学习，实现多模态输入的统一和融合，例如OpenAI的CLIP模型通过对比学习，将图像与文本通过各自的与训练模型获得的编码向量在向量空间上对齐，从而理解和推理图像和文本之间的关系。
- **多模态生成**：多模态生成模型基于对不同模态信息的理解，具备文本、图像、视频、语音信息的生成能力，分为两类技术路线：1) 一方面，**DiT结合扩散模型与Transformer优势**，成为视频生成模型主流架构；2) 另一方面，**端到端统一多模态架构**，实现跨模态生成与实时交互响应，例如GPT-4o与Gemini均采用端到端原生多模态单体模型的方式学习文本、视觉、语音等不同模态的统一表征，实现跨模态实时交互响应。

图表1：多模态模型技术路线表

类型	路线	典型
多模态理解	语言大模型调度	微软 Visual ChatGPT
		谷歌 PaLM-E
	跨模态特征对齐	OpenAI CLIP
		微软 KOSMOS
		DeepMind Flamingo
		Salesforce BLIP
多模态生成	扩散模型	Stability AI Stable Diffusion
		OpenAI DALL·E
		OpenAI Sora
		快手 KLING
	端到端理解与生成架构	Runway Gen-3
		谷歌 Gemini
		OpenAI GPT-4o

资料来源：中国信通院《人工智能发展报告（2024年）》，中邮证券研究所
请参阅附注免责声明

1.1.2 国内外大厂积极涌入多模态赛道，图片&视频陆续刷新SOAT

- **图片：**继GPT-4o原生图像生成功能掀起使用热潮后，谷歌最新Gemini 2.5 Flash Image迅速登顶多个主流图像排行榜；Meta也宣布将从初创企业Midjourney授权AI图像模型，以跟随OpenAI与谷歌步伐。
- **3月底，GPT-4o集成多模态原生图像生成功能，效果良好引爆用户使用热情。**GPT-4o原生图像生成，意味着图像生成不再依赖于单独的DALL-E 3模型，而是直接集成到了GPT-4o的核心能力中，这种集成带来了更流畅的用户体验和更强大的图像生成、修改能力，具备更好的文本集成、增强的上下文理解、改进的多对象绑定、多样化风格适应等优势。得益于超预期的效果，GPT-4o原生图像功能上线不到72小时，OpenAI CEO奥特曼宣布ChatGPT图像生成功能开始暂时受限。
- **8月底，谷歌发布了最新的图像生成和编辑模型Gemini 2.5 Flash Image，成为新晋性能冠军，单图片生成成本仅0.039美元。**2.5 Flash Image相较2.0 Flash Image，在图像质量、编辑控制和应用场景上有大幅改进。用户不仅可以对人物和宠物进行精准编辑，保持其特征一致，还能实现多图合成、多轮次修改与风格迁移等复杂操作。在大模型竞技场LMArena的文生图与图像编辑两个场景，谷歌的图像模型均拿下全球第一，在图像编辑榜单上模型表现尤为出色，获得1362的高分。

图表2：GPT-4o吉卜力风格《星际穿越》



资料来源：新智元微信公众号，中邮证券研究所
请参阅附注免责声明

图表3：谷歌Gemini 2.5 Flash Image登顶LMArena文生图与图像编辑

Text-to-Image				Image Edit			
Rank (UB) ↑	Model ↓	Score ↓	Votes ↓	Rank (UB) ↑	Model ↓	Score ↓	Votes ↓
1	gemini-2.5-flash-image-previ...	1147	220,674	1	gemini-2.5-flash-image-previ...	1362	2,521,035
2	imagen-4.0-ultra-generate-pr...	1135	193,895	2	flux-1-kontext-max	1191	357,196
3	gpt-image-1	1129	128,710	3	flux-1-kontext-pro	1174	2,015,530
4	imagen-4.0-generate-preview...	1119	196,696	3	gpt-image-1	1170	1,026,399
5	seedream-3	1077	159,028	5	flux-1-kontext-dev	1152	1,584,400
5	flux-1-kontext-max	1075	78,017	6	qwen-image-edit	1145	1,585,904
7	imagen-3.0-generate-002	1062	256,225	6	seededit-3.0	1142	1,285,080
8	flux-1-kontext-pro	1056	165,377	8	gemini-2.0-flash-preview-ima...	1093	1,700,785
9	qwen-image	1051	79,245	9	bagel	1044	12,774
10	photon	1035	181,590	10	stepix-edit	1017	138,399

资料来源：第一财经微信公众号，中邮证券研究所

1.1.2 国内外大厂积极涌入多模态赛道，图片&视频陆续刷新SOAT

- **视频：**阿里字节等大厂刷新视频生成时长、帧率、一致性等能力高度，谷歌将视频模型迈入支持实时交互的通用世界模型。
- 6月，字节**豆包视频生成模型Seedance 1.0 pro超越Veo3、可灵2.0等多模态模型**，在文生视频、图生视频两项任务上均排名首位（评测榜单Artificial Analysis）。该模型具备三大特性：无缝多镜头叙事、多动作及随心运镜、稳定运动与真实美感。Seedance支持文字与图片输入，可生成多镜头无缝切换的1080P高品质视频。
- 8月，阿里通义万相正式开源全新多模态视频生成模型**通义万相Wan2.2-S2V，单次生成的视频时长可达分钟级，支持73帧长视频生成**，完美匹配口型。7月28日，阿里开源文生视频Wan2.2-T2V-A14B、图生视频Wan2.2-I2V-A14B和统一视频生成Wan2.2-IT2V-5B三款模型。其中，文生视频模型和图生视频模型为业界首个使用MoE架构的视频生成模型。此次开源的Wan2.2-S2V则更加偏向音频驱动，专攻图像+音频，让画面和音频更加契合。
- **8月，谷歌发布首个支持实时交互的世界模型Genie 3**，相较于前两代产品，它不仅能在720p、24 fps的规格下即时生成可交互的三维场景，还把单次交互时长从几十秒延长到数分钟，并首次加入“可提示的世界事件”和“视觉记忆”等关键特性，使得同一房间里的涂鸦、家具位置等细节在玩家多次往返时依然保持一致，真实感提升。

图表4：豆包视频生成模型Seedance 1.0 pro在Artificial Analysis测评登顶

Artificial Analysis Video Arena Leaderboard					Artificial Analysis Video Arena Leaderboard				
Text to Video					Text to Video				
Creator	Model	Arena ELO	95% CI	# Appearances	Creator	Model	Arena ELO	95% CI	# Appearances
ByteDance Seed	Seedance 1.0	1299	-13/+13	4,947	ByteDance Seed	Seedance 1.0	1343	-13/+14	5,454
Google	Veo 3 Preview	1252	-18/+18	8,833	Google	Veo 3 Preview	1238	-11/+11	8,261
Google	Veo 2	1128	-8/+8	12,114	Kuaishou	Kling 2.0	1197	-18/+18	9,748
Kuaishou	Kling 2.0	1114	-9/+9	9,467	Kuaishou	Kling 1.6 (Pro)	1138	-9/+9	10,471
Kuaishou	Kling 1.5 (Pro)	1054	-5/+5	24,211	Runway	Runway Gen 4	1121	-9/+9	21,161
OpenAI	OpenAI Sora	1053	-5/+5	25,251	Google	Veo 2	1118	-9/+9	10,564

Artificial Analysis 文生视频榜单

Artificial Analysis 图生视频榜单

截止2025年6月9日

资料来源：新智元微信公众号，中邮证券研究所

请参阅附注免责声明

图表5：谷歌历次视频生成模型对比

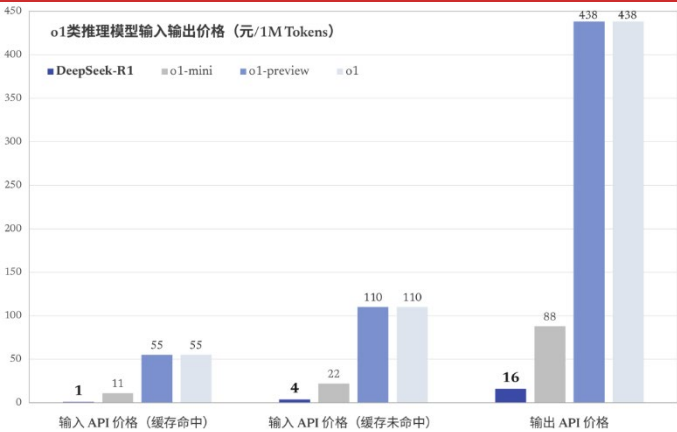
	GameNGen	Genie 2	Veo	Genie 3
分辨率	320p	360p	720p 至 4K	720p
领域	针对特定游戏	3D 环境	通用	通用
控制方式	针对特定游戏	限制的键盘/鼠标操作	视频级描述	导航、可指令触发实时反馈的世界事件
交互时长	几秒	10-20 秒	8 秒	数分钟
交互延迟	实时	非实时	不适用	实时

资料来源：腾讯科技微信公众号，中邮证券研究所

1.2 经历开源浪潮后，闭源模型逐渐开始维持性能领先优势

- 25年开年，DeepSeek多款模型通过算法的系列创新提升算力利用率，以成本、低价、性能等特征显著出圈，带来了模型平权和应用加速。
- DeepSeek以开源方式加速追赶闭源模型，API低价推动“大模型平权”。2024年12月，深度求索推出的DeepSeek-V3以极低的训练成本，实现了与GPT-4o和Claude Sonnet 3.5等顶尖模型相媲美的性能；2025年1月，DeepSeek推出推理模型DeepSeek R1，在后训练阶段大规模使用了RL技术，在仅有极少标注数据的情况下，在数学、代码、自然语言推理等任务上，性能比肩OpenAI o1正式版。R1 API输出16元/百万tokens，而GPT o1为438元/百万tokens。
- 模型能力和推理成本迎来拐点，国内Agent应用企业积极接入DeepSeek API，把握AI应用需求爆发红利。2月，国内万兴科技、易点天下、泛微网络等AI应用厂商相继宣布自身产品接入DeepSeek模型系列，并取得良好的协同赋能。

图表8：R1相比同类推理模型价格显著下降

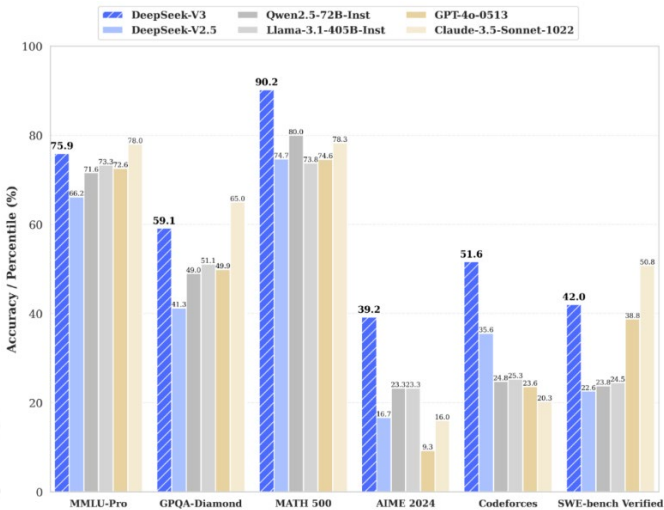


图表6：接入DeepSeek的国内AI应用公司

公司	细分领域	接入DeepSeek时间	赋能情况
万兴科技	视频工具	2025.02.04	完成深度求索（DeepSeek）最新推理大模型DeepSeek-R1的深入适配，涵盖旗下视频创意、绘图创意及文档创意软件业务多款产品。
易点天下	智能营销	2025.02.05	已完成DeepSeek开源推理大模型DeepSeek-R1的私有化部署，这也意味着易点天下旗下KreadoAI、数眼智能、zMaticoo ADX等多款核心产品将融合DeepSeek-R1大模型相关能力。
泛微网络	OA	2025.02.06	泛微公司旗下所有产品全面接入DeepSeek大模型，借助泛微2024年发布的数智大脑XiaoE.AI，可快捷方便为客户搭建“DeepSeek大模型+专业小模型+智能体”的数智底座，并可量身定制更安全、高效、国产化的数智化解决方案，助力组织管理与业务、财务一体化数智运营升级。
金蝶国际	ERP	2025.02.06	DeepSeek全面集成至金蝶云全线SaaS应用及金蝶云苍穹平台，为客户提供更高效、更安全、更低成本的智能解决方案
汉得信息	工业软件	2025.02.06	2024年年中已经完成了和Deepseek-V2的对接，在2025年1月份完成了和R1版本的对接（V3版本的对接随着官网V2模型的升级，已自动对接）。
用友网络	ERP	2025.02.06	用友BIP全面上线以DeepSeek-V3和DeepSeek-R1作为基座大模型的智能服务，企业客户可以通过用友BIP3 R6智能平台的公有云、专属云模式使用相关服务。
法本信息	ITO	2025.02.06	当前已实现FarAI GPTCoder企业级辅助编码平台、FarAI GPTBrain 企业级知识大脑产品、FarAI GPTRecruit 智能招聘平台等多款产品与DeepSeek的深度融合及优化，为多行业的智能化转型提供强大AI动力。

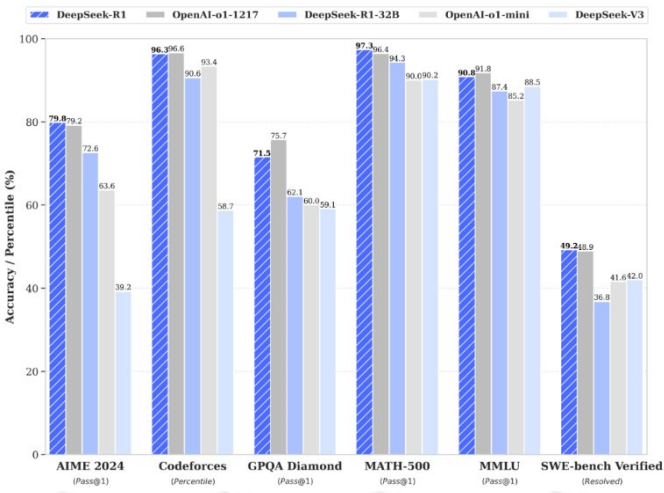
资料来源：各公司官网微信公众号，中国证券报微信公众号，中邮证券研究所

图表7：V3性能媲美GPT-4o和Claude Sonnet 3.5



资料来源：DeepSeek微信公众号，中邮证券研究所

图表9：R1性能比肩o1

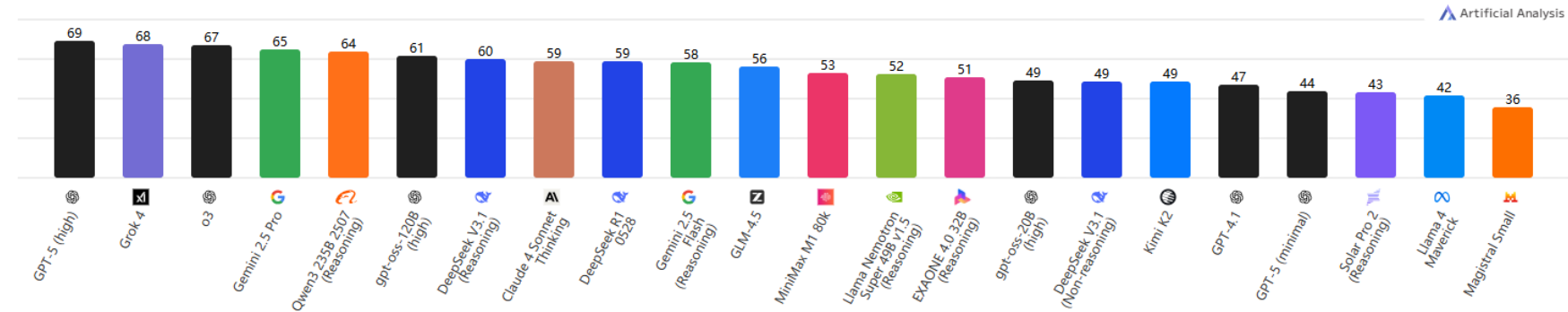


1.2 经历开源浪潮后，闭源模型逐渐开始维持性能领先优势

- 历经开源浪潮后，当前大模型生态呈现两大核心特征：
- 一方面，随着预训练模型Scaling Law（缩放定律）增速放缓，闭源与开源大模型的性能差距有所收窄，但以OpenAI、Claude、xAI、Google为代表的闭源厂商，凭借资源、人才与技术积累仍占据性能领先地位，持续刷新模型能力上限。例如，OpenAI新推出的GPT 5在数学、代码、多模态理解、健康等领域基准测试均获最新SOAT。根据Artificial Analysis编制的人工智能综合指数，GPT 5以69的综合评分排名第一，Grok 4以1分之差位居第二。
- 另一方面，闭源厂商逐步推行“开源 + 闭源”协同策略，在保留主模型闭源优势提高商业价值的同时，通过开源上一代或轻量小模型打造生态入口，既激活开发者推动技术改进，也反哺自身闭源模型，助力其维持领先。例如，OpenAI开源两款全新的大语言模型：gpt-oss-120b和gpt-oss-20b，前者可在单块800GB显存的H100上高效运行，后者转为低延迟、本地化或专业场景设计，仅需16GB内存即可在边缘设备上运行；谷歌开源系列Gemma 3模型（单云和桌面加速器的Gemma 3和Gemma 3 QAT、实时多模态AI的Gemma 3n、轻量化Gemma 3 270M）；xAI在7月发布了Grok4之后，8月开源上一代grok2.5并计划约6个月后年开源Grok3。

图表10：闭源模型相较开源模型仍保持性能领先优势

Artificial Analysis Intelligence Index: Results
Independently conducted by Artificial Analysis



资料来源：Artificial Analysis，中邮证券研究所（注：分析指数综合了七项评估的表现：MMLU-Pro、GPQA Diamond、Humanity's Last Exam、LiveCodeBench、SciCode、AIME 2025和IFBench，旨在全面衡量人工智能在数学、科学、编程和推理方面的能力）

请参阅附注免责声明

图表11：国外大模型龙头厂商开源策略

	闭源模型	开源模型
	GPT 5	gpt-oss-120b gpt-oss-20b
	Grok 4	Grok 2.5
	Gemini 2.5	Gemma 3

资料来源：腾讯科技微信公众号，中邮证券研究所

二

算力：海内外资本开支提振，芯片厂加速推出新版本芯片

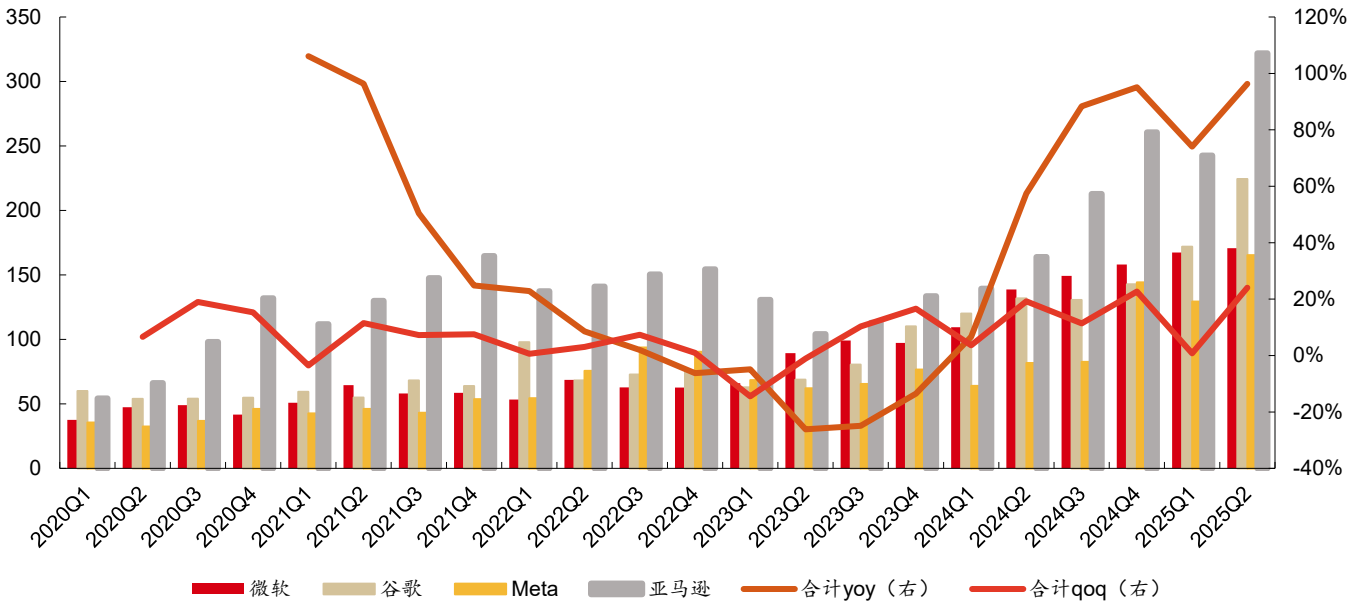
2.1 国内外CSP厂商开启AI军备竞赛，Capex先行提高投入

2.2 GPU、ASIC竞相角逐新品研发，满足海内外CSP放量需求

2.1 国内外CSP厂商开启AI军备竞赛，Capex先行提高投入

- 海外：微软、谷歌、Meta、亚马逊四者合计资本开支在2023Q3后连续多个季度环比上升，同时上调资本开支指引。
- 微软：25Q2（2025财年第四财季）资本开支为242亿美元，同比+27%，Q1同比+53%；其中，不动产和设备相关的现金支出171亿美元，同比+23%，Q1同比+52%。目前，微软正在加速建设数据中心，以满足对AI训练和工具激增的市场需求。公司预计，2026财年第一季度有望将资本开支提升至300亿美元以上。
 - 谷歌：由于云产品和服务的需求强劲且不断增长，25年capex上调至850亿美元，同比+13%，此前2月份预计全年capex为750亿美元；预计26年capex将进一步增加。
 - 亚马逊：2025财年制定了1000亿美元的激进资本支出计划，重点关注人工智能基础设施和数据中心的扩张。
 - Meta：25Q2，将全年capex的最低水平从上季度的640亿美元上调至660亿美元，capex范围在660亿美元至720亿美元之间，25年的总支出将在1140-1180亿美元之间。

图表12：国外CSP龙头资本开支情况（亿美元）

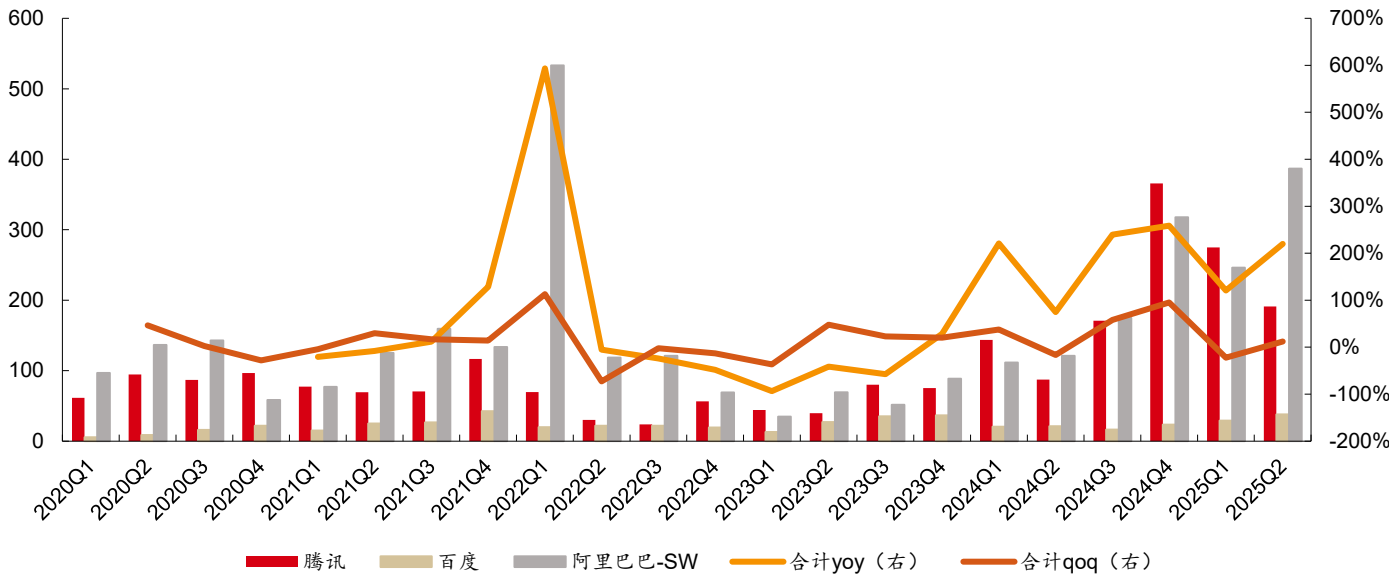


资料来源：Wind，各公司公告，中邮证券研究所

2.1 国内外CSP厂商开启AI军备竞赛，Capex先行提高投入

- 国内：以腾讯、阿里为代表的云厂商由于AI发展滞后于海外，23下半年开始发力，加大对AI领域投资。当前，我国开源大模型取得重要突破，资本开支上或将连续上涨。
- 阿里：2月，阿里CEO吴泳铭表示，将加大投入三大AI领域，未来三年将是云建设最集中的三年，在云和AI的基础设施投入预计将超越过去十年的总和。8月29日，阿里发布FY26Q1财报，capex 386.76亿元，同比+220%，超市场预期，创历史新高，并重申未来三年持续投入3800亿元用于AI资本开支的计划。
- 腾讯：从24年下半年开始，腾讯发觉AI加速发展迹象，加大了对AI芯片的投入。25年腾讯计划进一步加大资本开支，预计会占25年总收入的“低两位数百分比”，意味着25年腾讯capex可能接近千亿水平，而24年全年capex为767.6亿元。
- 百度：自24Q3开始，资本开支连续三个季度环比增长。25Q2，百度资本开支为38亿元，同比+79.41%，环比+31.03%。
- 阿里等资本开支先行，也有望带动华为云、通信运营商等资本开支预期，以应对云端推理需求大幅增加的可能。例如，根据中国产业经济信息网，三大运营商25年capex计划规模合计达到2898亿元，虽各自的投资规模出现下调，但在投入结构上偏向算力、AI，甚至还设立了特别预算或者“不设上限”。

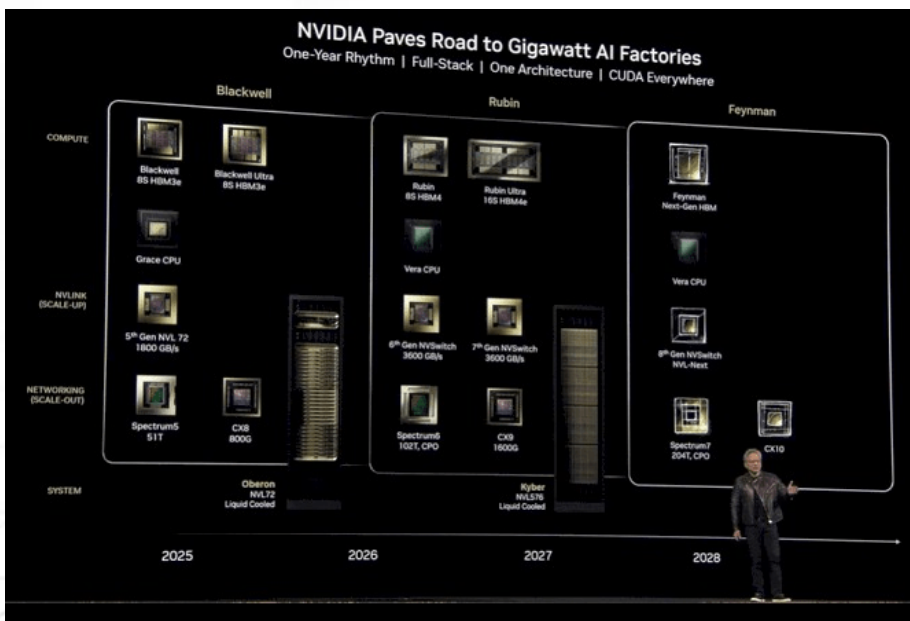
图表13：国内CSP龙头资本开支情况（亿元）



2.2.1 英伟达由B向R卡跨越，AMD布局26年新型MI400系列GPU

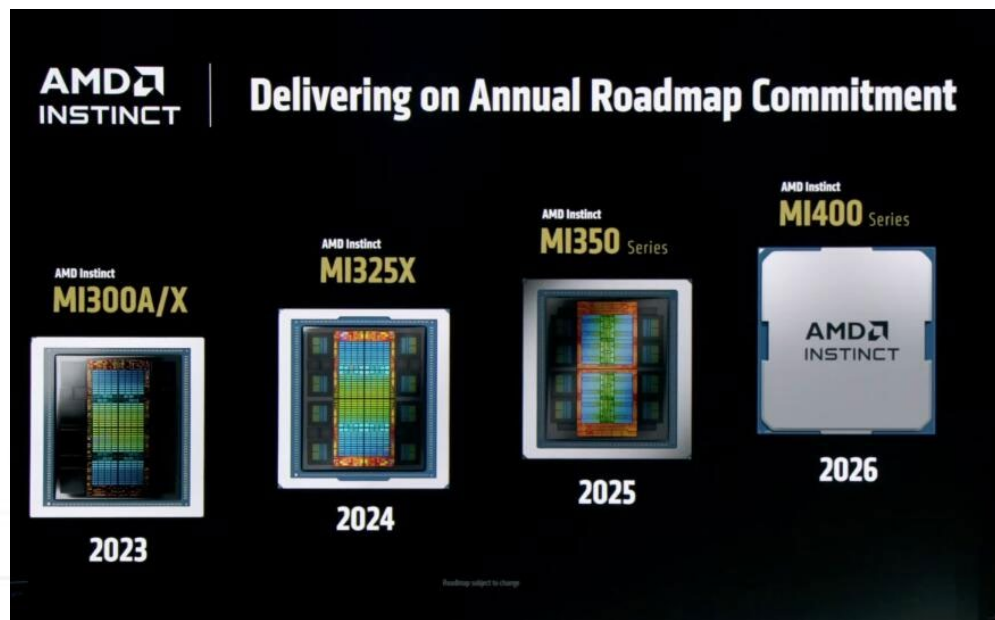
- **英伟达芯片架构正在由Blackwell向Rubin过渡。**根据美股研究社，短期看，英伟达GB300相比GB200在性能、存储、互连、安全与管理等多个方面进行了全面升级，性能增幅高达50%，25Q4或是NVL72规模化出货节点。26年下半年，英伟达将发布Vera Rubin架构，这是基于HBM4内存、基于3nm节点构建的全新芯片系列。Vera Rubin NVL144将提供比GB300高三倍以上的推理计算能力提升，同时提供机架级系统级架构兼容性。在2027年之前，Rubin Ultra设计将继续突破极限。凭借NVL576机架设计，它可以提供高达15 exaFLOPS的FP4吞吐量，大约是新GB300系统计算能力的14倍。
- **AMD MI350系列对标GB200，MI400系列计划为26年新型服务器“Helios”的基础。**6月，AMD正式推出下一代Instinct MI350系列，其中包括MI350X和旗舰MI355 GPU及平台。在FP4和FP6精度，新GPU芯片相对MI300X实现的AI计算性能提升4倍，推理性能提升35倍。此外，AMD预览了基于MI400 GPU的下一代人工智能机架“Helios”。

图表14：英伟达GPU芯片路线图



资料来源：美股研究社微信公众号，中邮证券研究所
请参阅附注免责声明

图表15：AMD GPU芯片路线图



资料来源：电子发烧友微信公众号，中邮证券研究所

2.2.2 华为云CloudMatrix 384通过多卡互联实现性能弯道超车

- 25年4月，华为发布384超节点，凭借其颠覆性的系统架构设计与全栈技术创新，在多项关键指标上实现对英伟达旗舰产品GB200 NVL72的超越，标志着中国在人工智能基础设施领域实现里程碑式突破。
- CloudMatrix 384由384颗昇腾910C芯片组成，这些芯片通过全互联拓扑结构连接。其中的权衡很简单：昇腾芯片的数量是英伟达 Blackwell芯片的五倍，这足以弥补每颗昇腾芯片的性能仅为后者三分之一的差距。
- 完整的CloudMatrix系统现在可以提供300 PFLOPs的密集型BF16计算能力，几乎是GB200 NVL72的两倍；凭借超过3.6倍的总内存容量和2.1倍的内存带宽，为大规模AI训练和推理提供了更高效的硬件支持。

图表16：华为CloudMatrix 384实物图



资料来源：华为，Semianalysis，中邮证券研究所
请参阅附注免责声明

图表17：华为CloudMatrix 384与英伟达GB200 NVL72性能对比图

Key Capabilities - Huawei Ascend 910C Cloud Matrix 384 vs Nvidia GB200 NVL72				
Chip and Package Level				
	Unit	GB200	Ascend 910C	Huawei vs Nvidia
BF16 dense TFLOPS	TFLOPS	2,500	780	0.3x
HBM capacity	GB	192	128	0.7x
HBM bandwidth	TB/s	8.0	3.2	0.4x
Scale Up Bandwidth	Gb/s uni-di	7,200	2,800	0.4x
Scale Out Bandwidth	Gb/s uni-di	400	400	1.0x
System Level				
	Unit	Nvidia GB200 NVL72	Cloud Matrix CM384	Huawei vs Nvidia
BF16 dense PFLOPS	PFLOPS	180	300	1.7x
HBM capacity	TB	13.8	49.2	3.6x
HBM bandwidth	TB/s	576	1,229	2.1x
Scale Up Bandwidth	Gb/s uni-di	518,400	1,075,200	2.1x
Scale Up Domain Size	GPUs	72	384	5.3x
Scale Out Bandwidth	Gb/s uni-di	28,800	153,600	5.3x
All-In System Power ₁	W	145,000	599,821	4.1x
All-in Power per BF16 dense FLOP	W/TFLOP	0.81	2.00	2.5x
All-in Power per memory bandwidth	W per TB/s	251.7	488.1	1.9x
All-in Power per memory capacity	kW/TB	10.5	12.2	1.2x

1. All-in System Power is total cluster power including scale-out networking, storage, etc.

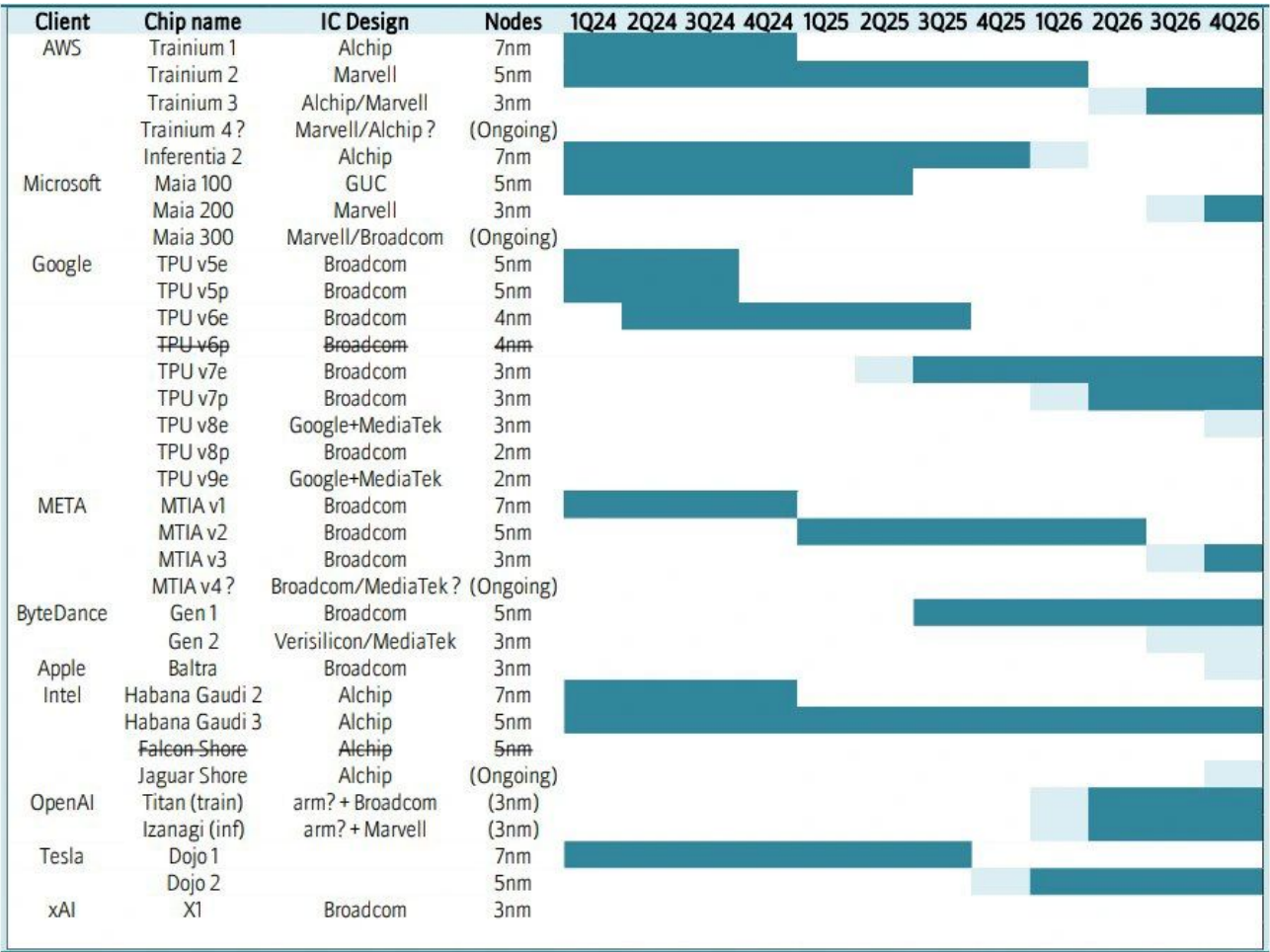
资料来源：华为，英伟达，Semianalysis，中邮证券研究所

2.2.3 云端巨头竞相自研ASIC推理芯片

- 根据TrendForce集邦咨询，AI Server需求带动北美四大CSP加速自研ASIC芯片，平均1~2年就会推出升级版本。CSP为应对AI工作负载规模逐步扩大，同时计划降低对NVIDIA、AMD的高度依赖，因此积极投入ASIC开发进程，以便能控制成本、性能和供应链弹性，进一步改善营运成本支出。
- **谷歌**：已推出TPU v6 Trillium，主打能效比和针对AI大型模型的最佳化，预计2025年将大幅取代现有TPU v5。
 - **亚马逊**：目前以与Marvell（美满电子）协同设计的Trainium v2为主力，其主要支持生成式AI与大型语言模型训练应用，AWS也和Alchip合作Trainium v3开发。
 - **Meta**：成功部署首款自研AI加速器MTIA后，正与Broadcom共同开发下一代MTIA v2，MTIA v2设计特别聚焦能效最佳化与低延迟架构，以确保兼顾推理效能与运营效率。
 - **微软**：目前在AI Server建置仍以搭载NVIDIA GPU的解决方案为主，但也加速ASIC开发，其Maia系列芯片主要针对Azure云端平台上的生成式AI应用与相关服务进行优化，下一代Maia v2的设计也已定案，并由GUC负责后段实体设计及后续量产交付。
 - **国内**：阿里巴巴旗下平头哥(T-head)已推出Hanguang 800 AI推理芯片；百度继量产Kunlun II后，已着手开发Kunlun III，主打高效能训练与推理双支持架构；腾讯除了自家AI推理芯片Zixiao，亦采用策略投资的IC设计公司Enflame（燧原科技）解决方案。

请参阅附注免责声明

图表18：国内外ASIC研发时间预测



资料来源：富邦投顾，零氪1+1微信公众号，中邮证券研究所

三

应用：海外应用加速分化，聚焦与数据结合的复杂toB场景

- 3.1 创新竞争加速，货币化进展成为AI初创公司的关键维度
- 3.2 企业级Agent落地较快，核心在于打破数据孤岛与保障结果准确
- 3.3 美股toB应用：AI业务强协同、AI加速放量者市场表现更优
- 3.4 C端应用：创意类开始挤压工具类，AI的替代性开始显现

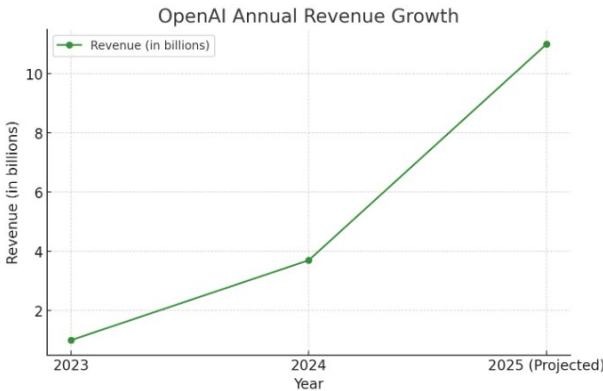
3.1 创新竞争加速，货币化进展成为AI初创公司的关键维度

➤ AI初创公司ARR放量时间显著缩短，货币化速度或为初创公司未来估值提升的重要考量。

■ OpenAI：得益于用户的高速增长与较强的粘性，OpenAI 25年收入有望达到116亿美元，同比翻2倍+，**25年8月ARR已达到120亿美元**。根据NerdyNav，ChatGPT为有史以来增长最快的应用程序，仅5天就达到100万用户，前两个月达到1亿用户，该公司的目标是在2029年实现盈利，收入目标为1250亿美元。

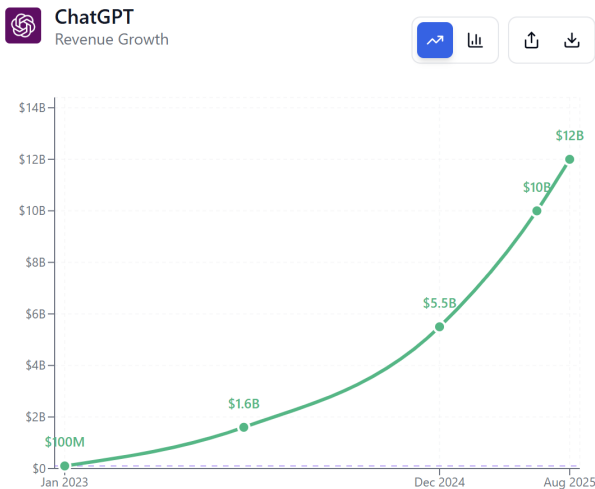
■ Anthropic：根据COATUE，Anthropic年化收入0 → 1、1 → 2、2 → 3亿美元分别经过21、3、2月，创收能力快速提升。根据ARR.club，**25年7月，Anthropic ARR已突破50亿美元**，是去年12月底的5倍。

图表19：OpenAI收入预测



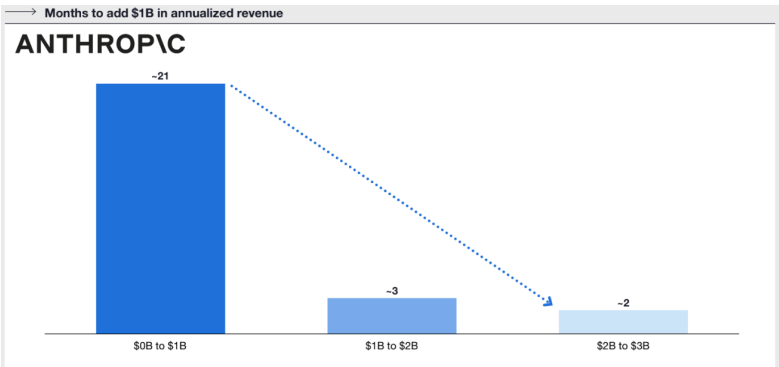
资料来源：NerdyNav，中邮证券研究所

图表21：ChatGPT的ARR变化



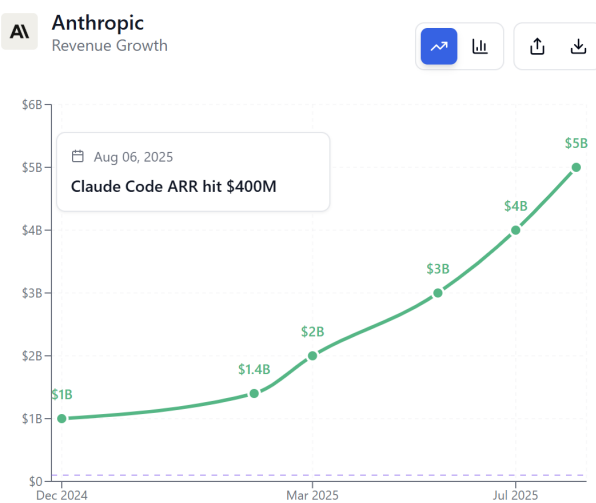
资料来源：ARR.club，中邮证券研究所

图表20：Anthropic年化收入增加1亿美元所需要的月份数



资料来源：COATUE，中邮证券研究所

图表22：Anthropic的ARR变化

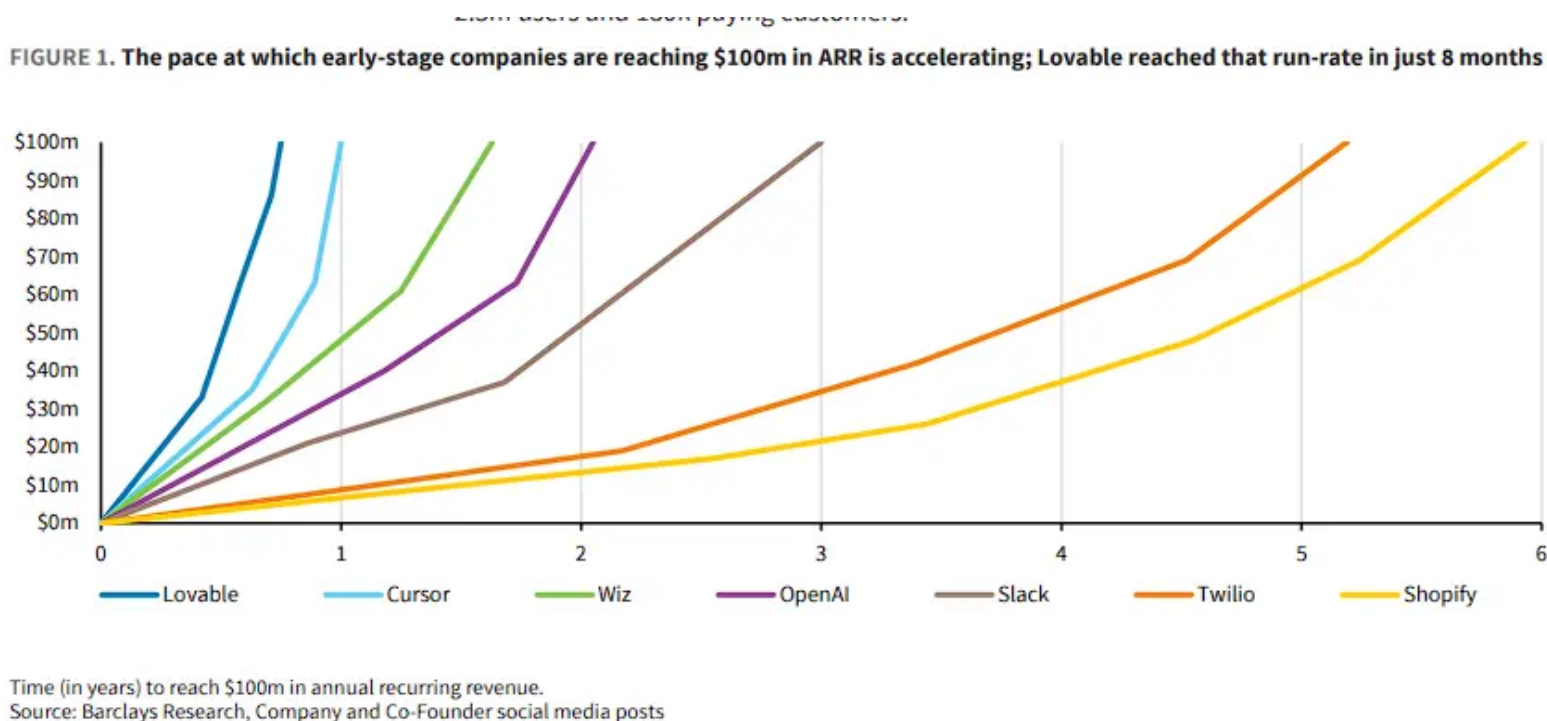


资料来源：ARR.club，中邮证券研究所

3.1 创新竞争加速，货币化进展成为AI初创公司的关键维度

- 根据非凡产研，AI IDE产品Cursor的ARR仅经过21个月就达到1亿美元；
- 据巴克莱研究报告，AI建站工具Lovable在8个月内达到1亿美元ARR，超越了Cursor等知名AI工具的增长速度，并在18亿美元估值下完成2亿美元A轮融资。

图表23：AI建站工具LovableARR突破1亿美元仅经过8个月

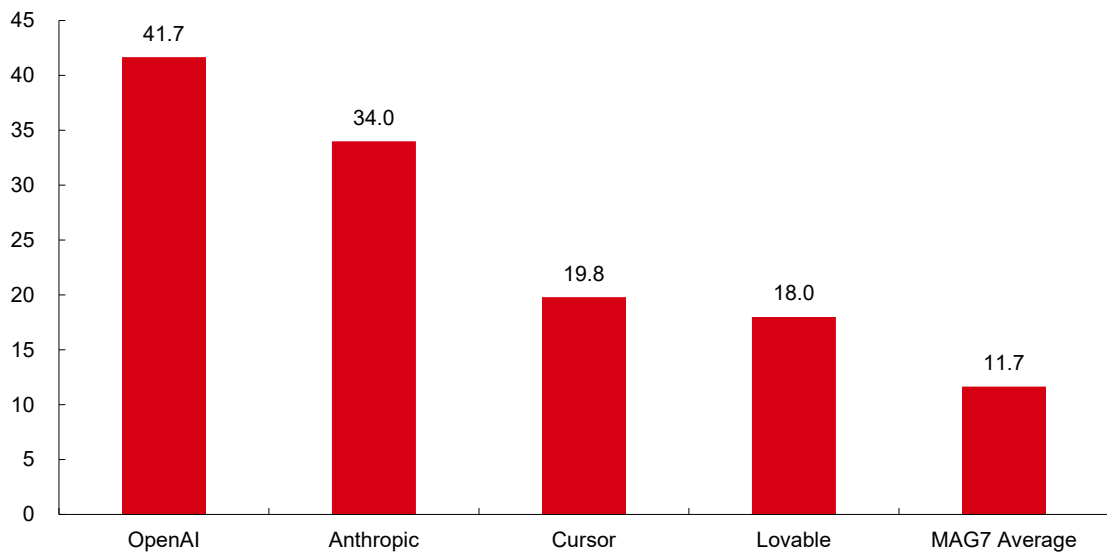


资料来源：巴克莱，华尔街见闻，中邮证券研究所

3.1 创新竞争加速，货币化进展成为AI初创公司的关键维度

- 市场给予货币化进展较快的AI初创公司较高的估值溢价。相比于美股科技七巨头平均11x PS估值，以OpenAI、Anthropic等为代表的初创公司估值为前者的三四倍。

图表25：部分AI初创公司与美股科技七巨头PS对比



资料来源：iFind，ARR.club，凤凰网科技，AIGC开放社区，财联社等，中邮证券研究所（注：美股科技七巨头为英伟达、苹果、特斯拉、微软、亚马逊、谷歌、Meta，估值取自2025年8月28日PS-TTM）

图表24：部分AI初创公司ARR与估值情况

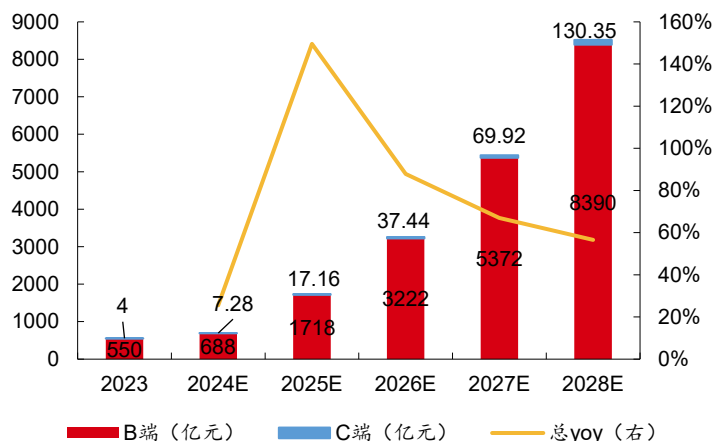
公司	ARR (亿美元)	最新估值 (亿美元)	PS
OpenAI	120	5000	41.67
	2025年8月	2025年8月	
Anthropic	50	1700	34.00
	2025年7月	2025年8月	
xAI	-	1130	-
	-	2025年6月	
Cursor	5	99	19.80
	2025年6月	2025年6月	
Lovable	1	18	18.00
	2025年7月	2025年7月	

资料来源：ARR.club，凤凰网科技，AIGC开放社区，财联社等，中邮证券研究所

3.2 企业级Agent落地较快，核心在于打破数据孤岛与保障结果准确

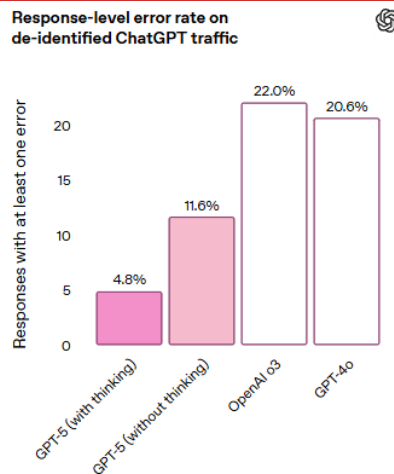
- **toB企业级Agent为主流。**根据头豹研究院的研究报告，2023年中国AI Agent市场规模为554亿元，预计至2028年将达到8520亿元，年均复合增长率高达72.7%。而在这8520亿元中，根据观研天下数据预测，预计2028年B端AI agent市场规模为8390亿元，占比98.5%。
- **对比通用型Agent，企业型Agent落地的关键，在于与业务流程深度结合，确保输出结果的准确可靠。**AI能够理解实际应用场景，能够在特定的业务场景、流程与数据库中稳定工作，使得企业能够通过数字员工提高生产力。
- 我们认为，企业级Agent的价值落地需分层推进“可靠性基础”与“业务流程结合”两大目标。前者依托大模型迭代，如思考模式下GPT-5较GPT-4o、o3事实错误率分别降约45%、80%，通过架构优化、数据扩容与强化学习降低幻觉率，筑牢技术底座；后者更依赖企业将行业隐性know-how转化为Agent可执行参数，及打通ERP、CRM等系统数据孤岛以实现业务流程的闭环，推动Agent从工具升级为流程主体。**基于此，未来Agent将聚焦与业务数据深度绑定的复杂toB场景，最大化释放Agent降本增效的价值。**

图表26：28年B段AI Agent占比98.5%



资料来源：观研天下，中邮证券研究所

图表27：GPT5相比GPT4o、o3错误率下降



资料来源：OpenAI官网，中邮证券研究所

图表28：通用级和企业级Agent对比

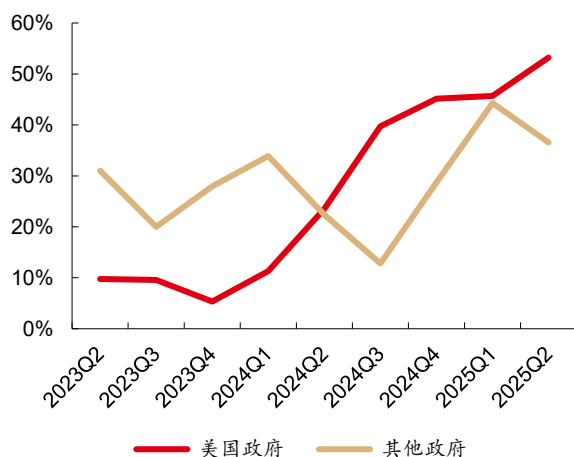
对比维度	通用级Agent (ToC)	企业级Agent (ToB)
容错率	容错率高 用户可以容忍一定程度的幻觉 例：调研100个网站，成功率70%已足够	零失误要求 任何错误可能导致直接的经济损失和信誉危机 例：财务报销、供应链管理不容出错
思维模式	发散思维 自主规划，提供多样化创意结果	收敛思维 业务路径确定化，结果可验证 在特定场景提供特定能力，输出稳定结果
应用范围	通用平台 处理多种任务，研究型导向 广泛覆盖数据、网页、金融等领域	垂直聚焦 针对特定业务场景深度优化 例：深度绑定CRM业务流程和数据
工具与系统集成	开放工具箱 浏览器、虚拟机、搜索引擎、代码解释器等 通过MCP调用各种互联网工具	业务系统整合 需要与ERP、CRM、HRM、财务系统等整合 挑战：对接运行多年可能缺乏API的老系统
使用场景	多样化消费场景 调研报告撰写 创意内容创作 一键生成PPT 旅游攻略规划 休闲娱乐服务	企业核心业务 财务报销流程 供应链管理 客户关系维护 会议纪要自动生成与同步 销售跟进任务创建
价值创造方式	便捷性与体验感 提供便利、节省时间、增强用户体验	闭环业务价值 提高效率、减少错误、创造直接业务价值 例：自动将会议纪要同步到CRM系统

资料来源：甲子光年，中邮证券研究所

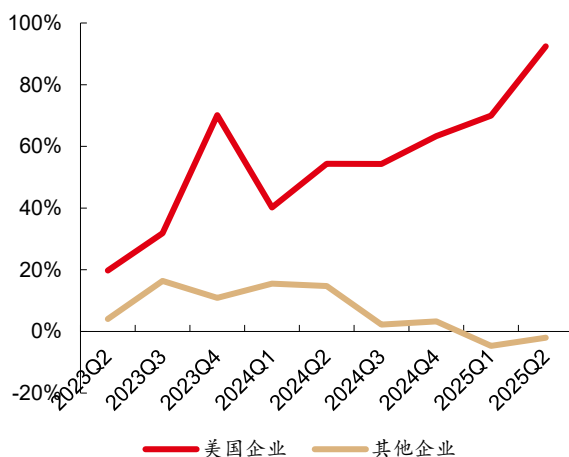
3.3 美股toB应用：AI业务强协同、AI加速放量者市场表现更优

- 复盘今年美股AI应用表现，行情有所分化，股价涨幅较好的公司体现在：具备显著的行业Know-how壁垒，同时AI赋能业务放量增长。例如，以PLTR为代表的toB企业，其股价走强的核心逻辑在于AI产品深度赋能主业（如AI提升服务效率，或拓展高价值客户场景），直接推动营收规模快速增长，这一特征在toB领域因商业场景更易落地而表现得尤为显著。
- Palantir (AI放量：年初至8月18日，股价+109.1%)：2023年推出AIP平台后，开启订单业绩高增期。2025Q2总收入达到10.04亿美元，同比+48%。
 - Applovin (AI放量：年初至8月18日，股价+49.4%)：Axon AI基于价值交付，带动广告收入放量增长。自23Q3开始，Applovin的广告业务收入维持60%以上高速增长。
 - SAP (AI放量：年初至8月18日，股价+12.8%)：Business AI成为订单成交重要手段，驱动云业务快速增长。SAP 25Q2超一半的云订单包含AI用例，实现营收90.27亿欧元，同比增长8.9%，其中云收入51.3亿欧元，同比增长28%。

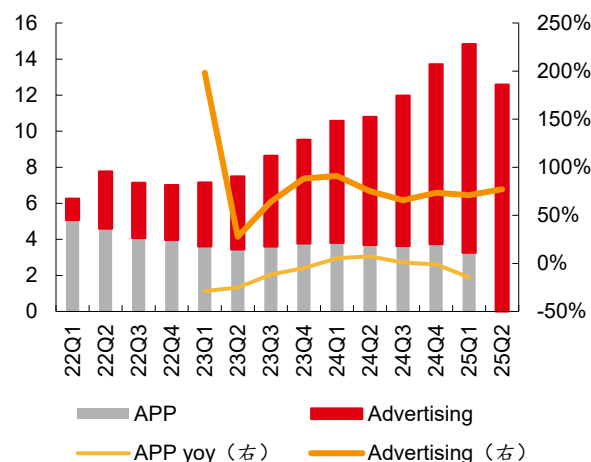
图表29：PLTR来自美国政府的收入增速明显上行



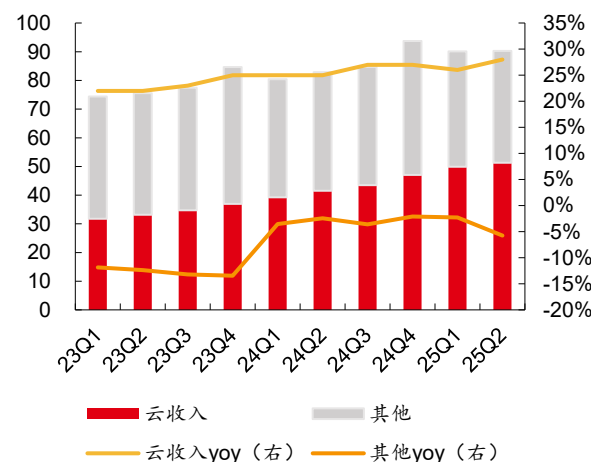
图表30：PLTR来自美国商业企业收入增速明显上行



图表31：AppLovin分业务收入（亿美元）及增速



图表32：SAP分业务收入（亿欧元）及增速



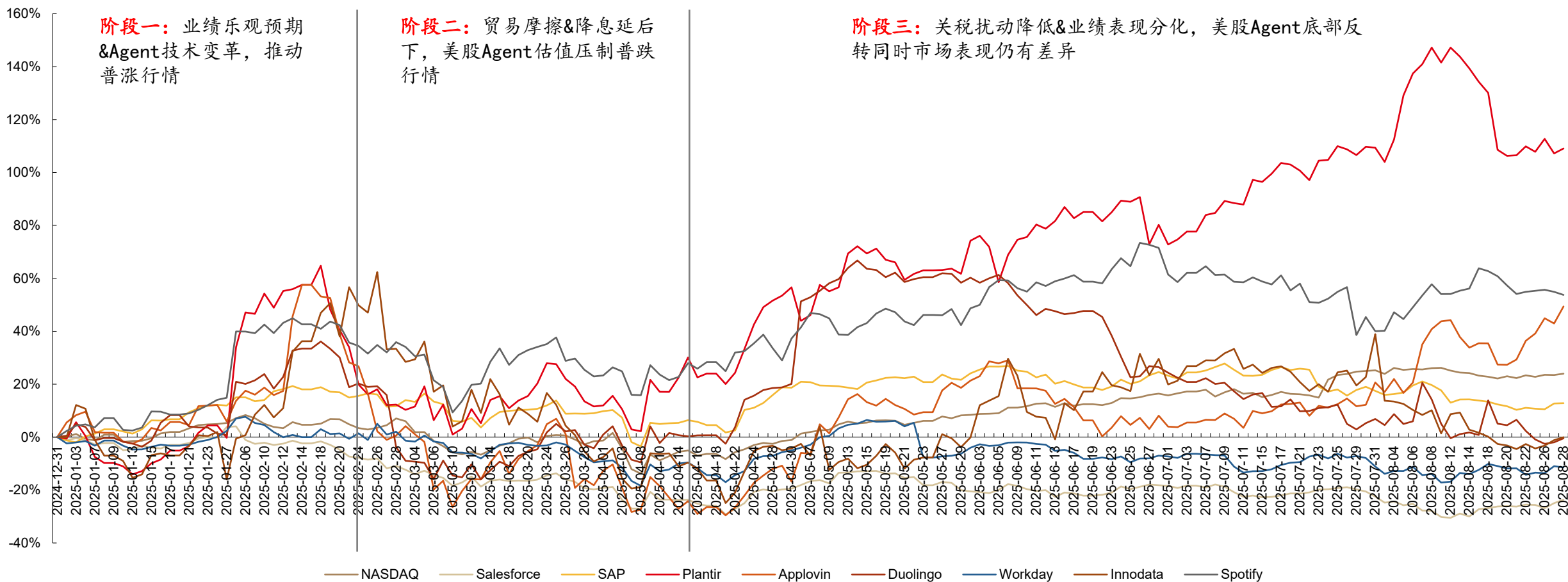
资料来源：彭博，各公司公告，海豚投研等，中邮证券研究所

请参阅附注免责声明

3.3 美股toB应用：AI业务强协同、AI加速放量者市场表现更优

- 整体而言，当前美股Agent板块对AI的定价逻辑已从“讲故事”转向“看实效”：市场更关注AI能否真正转化为营收增长的加速度、利润率改善的能见度，以及与商业模式的适配度，这也成为区分标的表现的核心标尺。

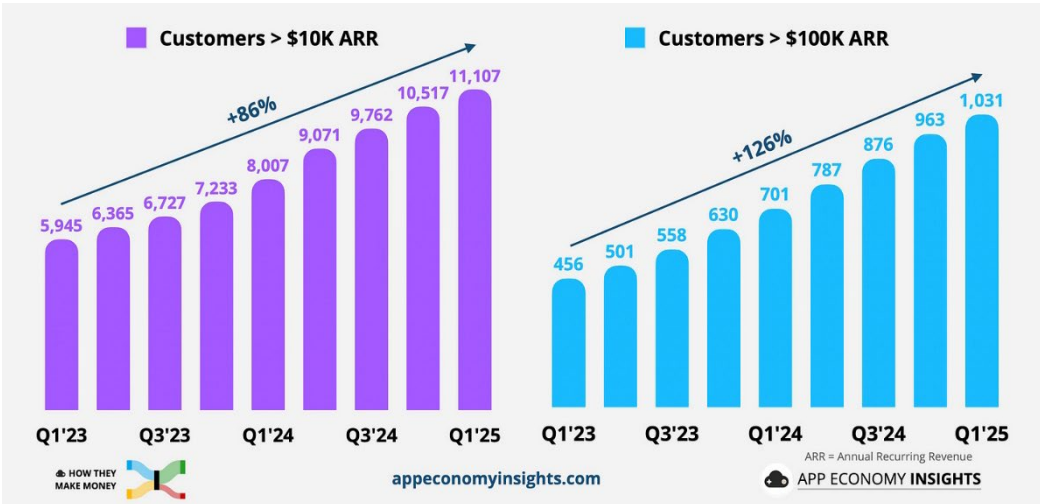
图表33：今年以来美股Agent股价涨幅复盘



3.4 C端应用：创意类开始挤压工具类，AI的替代性开始显现

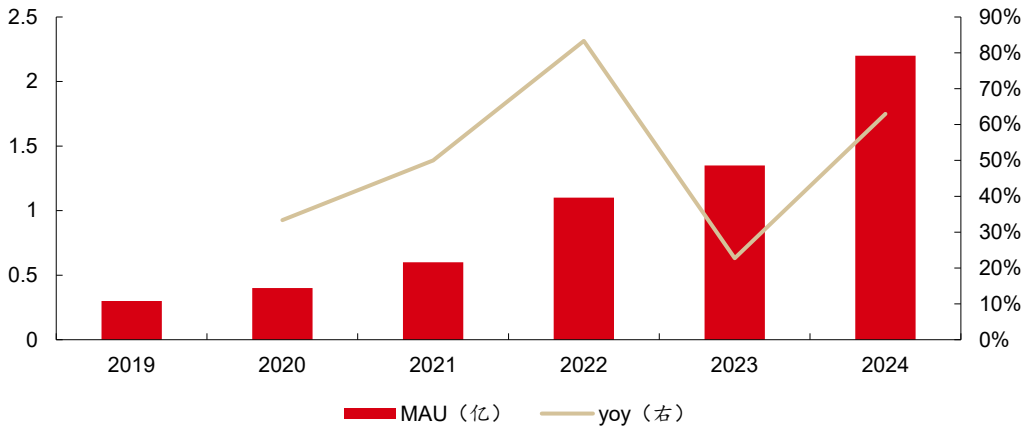
- 多模态大模型发展推动创意类应用从“辅助工具”进化为能理解复杂需求、提供创造性解决方案的“创意伙伴”，这种进化既强化其替代优势，又使其以低学习成本、快产出吸引用户，持续逼近传统专业工具价值腹地。
- 以UI/UX设计为例，Figma、Canva等创意软件的快速增长有望挤占传统工具软件Adobe的市场份额。
- Figma：支持构建的app、网站及数字界面，正将AI深度嵌入产品流程，以提升效率。最具代表性的AI功能Figma Make，支持用户通过提示快速生成可交互原型，并可结合现有设计组件。截至25Q1，Figma MAU超过1300万，其中约三分之二为非设计师，95%的财富500强企业；付费用户高达45万，其中ARR超10万用户达1031家，相对23Q1+126%，ARR超1万用户达11107家，相对23Q1+86%。
- Canva：轻便的设计工具，降低设计门槛。Canva推出视觉套件2.0，引入了AI设计助手、视觉化工作表、图表生成、批量创作等核心功能，极大地拓展了AI在设计与内容创作中的应用领域。根据Demandsage，24年底Canva拥有2.2亿MAU，同比+63%，其中2100万为付费订阅用户；《财富》500强企业中有95%使用Canva。
- Adobe：长期以来主导着设计市场，其产品套件（如Photoshop、Illustrator）在各个设计领域占有重要地位，然而，其相较缓慢的云端化进程，使其在部分新兴市场上面临来自类似Figma的竞争压力。

图表34：Figma ARR超过1、10万的客户数



资料来源：App Economy Insights，中邮证券研究所

图表35：Canva的MAU及增速

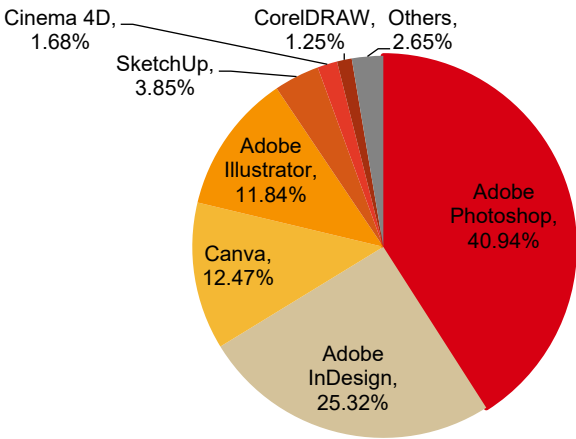


资料来源：Demandsage，中邮证券研究所

3.4 C端应用：创意类开始挤压工具类，AI的替代性开始显现

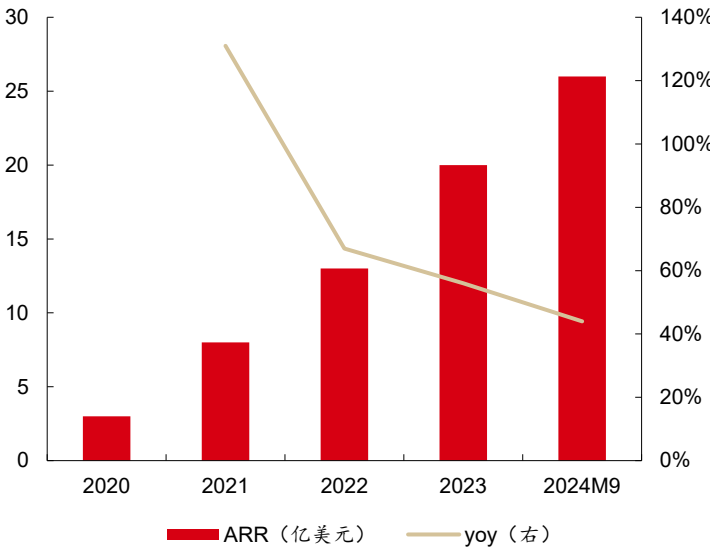
- 以UI/UX设计为例，Figma、Canva等创意软件的快速增长有望挤占传统工具软件Adobe的市场份额。
- 根据Demandsage，在画图工具市场中，Adobe Photoshop占据了40.94%的份额。Canva以12.47%的市场份额位居第三。
- 从ARR增速来看，Figma以及Canva正领先于传统的工具软件Adobe，前两者市场份额有望提升。根据SACRA，24年9月 Canva实现ARR 25.5亿美元，同比+44%；Figma实现ARR 7.43亿美元，同比+31%；Adobe实现ARR 2420亿美元，同比+11%。

图表36：画图工具市场格局



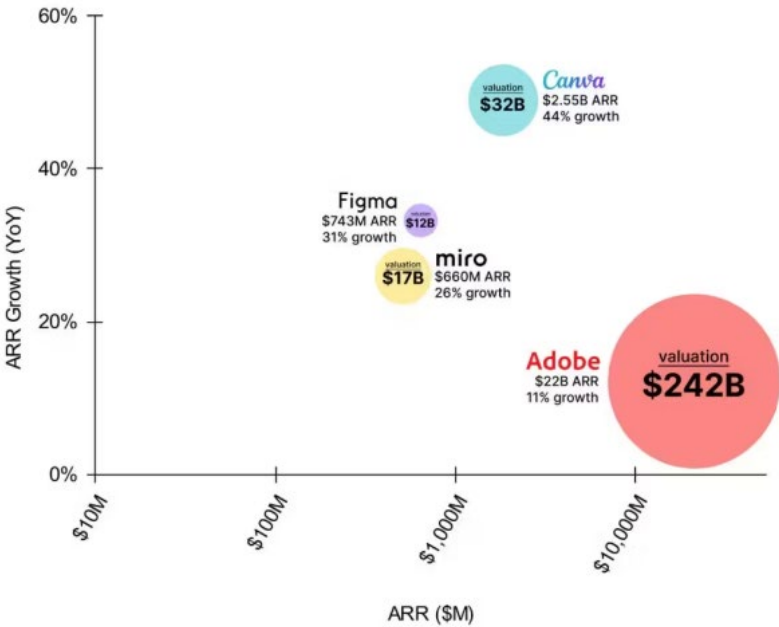
资料来源：Demandsage，中邮证券研究所

图表37：Canva ARR及增速



资料来源：SACRA，中邮证券研究所

图表38：Figma、Canva、Adobe ARR增速对比



资料来源：SACRA，中邮证券研究所

四

投资建议与风险提示

4.1 相关标的

4.2 风险提示

4.1 相关标的

➤ 建议关注：

- 1) AI Agent：鼎捷数智、视源股份、泛微网络、致远互联、金蝶国际、用友网络、汉得信息、赛意信息、迈富时、光云科技、税友股份、润达医疗、同花顺、恒生电子、金山办公、万兴科技、福昕软件、科大讯飞、金桥信息、卓易信息、普元软件等；
- 2) AI终端：云天励飞、虹软科技、小米集团、联想集团、科大讯飞、乐鑫科技、中科蓝讯等；
- 3) AI算力：寒武纪、云天励飞、亿都（国际控股）、工业富联、浪潮信息、曙光数创、超讯通信、华丰科技、神州数码、软通动力、烽火通信、广电运通、拓维信息、四川长虹等；
- 4) AI多模态：万兴科技、虹软科技、当虹科技、中科创达、大华股份、海康威视、漫步者、萤石网络、汉仪股份、美图公司、云从科技等。

4.2 风险提示

- **地缘政治冲与贸易风险：**地缘政治冲突可能导致跨境贸易壁垒升级（如关税提高、技术出口限制），例如某国对核心芯片实施出口管制，会直接影响上游零部件供应。
- **技术迭代不及预期：**AI大模型或AI Agent迭代放缓，影响AI商业化进展和客户付费意愿。
- **国内AI Agent商业化进展不及预期：**美国AI应用虽取得一定进展，而国内或由于客户对AI付费意愿和使用习惯与国外存在差异，商业化进展不及预期；
- **算力供应不及预期：**国内算力供给未能及时填补需求，则可能影响大模型训练及推理的节奏；
- **行业竞争加剧风险：**AI作为新型技术领域，行业竞争加剧将对企业盈利端造成影响。

感谢您的信任与支持!

THANK YOU

陈涵泊

SAC编号: S0120524040004

邮箱: chenhanbo@cnpsec.com

李佩京

SAC编号: S0120524090004

邮箱: lipeijing@cnpsec.com

王思

SAC编号: S1340525080002

邮箱: wangsi@cnpsec.com

分析师声明

撰写此报告的分析师（一人或多人）承诺本机构、本人以及财产利害关系人与所评价或推荐的证券无利害关系。

本报告所采用的数据均来自我们认为可靠的目前已公开的信息，并通过独立判断并得出结论，力求独立、客观、公平，报告结论不受本公司其他部门和人员以及证券发行人、上市公司、基金公司、证券资产管理公司、特定客户等利益相关方的干涉和影响，特此声明。

免责声明

中邮证券有限责任公司（以下简称“中邮证券”）具备经中国证监会批准的开展证券投资咨询业务的资格。

本报告信息均来源于公开资料或者我们认为可靠的资料，我们力求但不保证这些信息的准确性和完整性。报告内容仅供参考，报告中的信息或所表达观点不构成所涉证券买卖的出价或询价，中邮证券不对因使用本报告的内容而导致的损失承担任何责任。客户不应以本报告取代其独立判断或仅根据本报告做出决策。

本报告所载的意见、评估及预测仅为本报告出具日的观点和判断。该等意见、评估及预测无需通知即可随时更改。过往的表现亦不应作为日后表现的预示和担保。在不同时期，中邮证券可能会发出与本报告所载意见、评估及预测不一致的研究报告。

中邮证券及其所属关联机构可能会持有报告中提到的公司所发行的证券头寸并进行交易，也可能为这些公司提供或者计划提供投资银行、财务顾问或者其他金融产品等相关服务。

《证券期货投资者适当性管理办法》于2017年7月1日起正式实施，本报告仅供中邮证券签约客户使用，若您非中邮证券签约客户，为控制投资风险，请取消接收、订阅或使用本报告中的任何信息。本公司不会因接收人收到、阅读或关注本报告中的内容而视其为签约客户。

本报告版权归中邮证券所有，未经书面许可，任何机构或个人不得存在对本报告以任何形式进行翻版、修改、节选、复制、发布，或对本报告进行改编、汇编等侵犯知识产权的行为，亦不得存在其他有损中邮证券商业性权益的任何情形。如经中邮证券授权后引用发布，需注明出处为中邮证券研究所，且不得对本报告进行有悖原意的引用、删节或修改。

中邮证券对于本申明具有最终解释权。

公司简介

中邮证券有限责任公司，2002年9月经中国证券监督管理委员会批准设立，是中国邮政集团有限公司绝对控股的证券类金融子公司。

公司经营范围包括:证券经纪，证券自营，证券投资咨询，证券资产管理，融资融券，证券投资基金销售，证券承销与保荐，代理销售金融产品，与证券交易、证券投资活动有关的财务顾问等。

公司目前已经在北京、陕西、深圳、山东、江苏、四川、江西、湖北、湖南、福建、辽宁、吉林、黑龙江、广东、浙江、贵州、新疆、河南、山西、上海、云南、内蒙古、重庆、天津、河北等地设有分支机构，全国多家分支机构正在建设中。

中邮证券紧紧依托中国邮政集团有限公司雄厚的实力，坚持诚信经营，践行普惠服务，为社会大众提供全方位专业化的证券投、融资服务，帮助客户实现价值增长，努力成为客户认同、社会尊重、股东满意、员工自豪的优秀企业。

投资评级说明

投资评级标准	类型	评级	说明
报告中投资建议的评级标准： 报告发布日后的 6 个月内的相对市场表现，即报告发布日后的 6 个月内的公司股价（或行业指数、可转债价格）的涨跌幅相对同期相关证券市场基准指数的涨跌幅。 市场基准指数的选取：A 股市场以沪深 300 指数为基准；新三板市场以三板成指为基准；可转债市场以中信标普可转债指数为基准；香港市场以恒生指数为基准；美国市场以标普 500 或纳斯达克综合指数为基准。	股票评级	买入	预期个股相对同期基准指数涨幅在 20%以上
		增持	预期个股相对同期基准指数涨幅在 10%与 20%之间
		中性	预期个股相对同期基准指数涨幅在-10%与 10%之间
		回避	预期个股相对同期基准指数涨幅在-10%以下
	行业评级	强于大市	预期行业相对同期基准指数涨幅在 10%以上
		中性	预期行业相对同期基准指数涨幅在-10%与 10%之间
		弱于大市	预期行业相对同期基准指数涨幅在-10%以下
	可转债评级	推荐	预期可转债相对同期基准指数涨幅在 10%以上
		谨慎推荐	预期可转债相对同期基准指数涨幅在 5%与 10%之间
		中性	预期可转债相对同期基准指数涨幅在-5%与 5%之间
		回避	预期可转债相对同期基准指数涨幅在-5%以下

中邮证券研究所

北京

邮箱: yanjiusuo@cnpsec.com

地址: 北京市东城区前门街道珠市口东大街17号

邮编: 100050

上海

邮箱: yanjiusuo@cnpsec.com

地址: 上海市虹口区东大名路1080号大厦3楼

邮编: 200000

深圳

邮箱: yanjiusuo@cnpsec.com

地址: 深圳市福田区滨河大道9023号国通大厦二楼

邮编: 518048



中邮证券
CHINA POST SECURITIES