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

优于大市

核心观点

AI产业持续迭代，Agent成为当下应用最确定性方向。在基础大模型持续迭代的背景下，AI Agent应用水到渠成。当前各互联网厂商以及创新公司持续推出Agent相关产品，Agent已开始逐步进入各个场景的工作流中，成为人机协同新范式。根据MarketsandMarkets最新发布的《AI Agents Market Report 2025》，全球AI Agent（含自主智能体软件与服务）市场规模预计在2025年达到7.9亿美元，将在2030年增至526亿美元，复合增长率约46%。

谷歌推出Gemini 2.5 Pro，再次重回AI舞台中心。谷歌AI token调用猛增，去年同期谷歌AI大模型和API每月处理9.7万亿个Token，现在处理Token数增长至480万亿个。同时，谷歌开源A2A协议，打破系统孤岛，为Agent之间提供了一种标准交互方式，使它们能够相互协作，携手MCP打造AI Agent新生态。在最新的I/O大会上，谷歌发布Gemini 2.5 Pro，以及图片、视频、音频领域的多模态模型，表现惊艳。同时，谷歌推出AI Agent，将通过Gemini API开放给开发者，扩大谷歌AI生态。

阿里和字节持续推出Agent产品，创业公司百花齐放。阿里Qwen3性价比再大幅提升，以DeepSeek-R1三分之一的参数规模，就达成了性能的全面超越，仅需4张H20 GPU便能部署完整功能的Qwen3模型，成为全球最强开源模型。Qwen3原生支持MCP，C端积极探索“心流”和“夸克”产品；B端和亚信科技等合作推动AI本地化落地。字节在多模态领域积极布局，扣子空间开启内测，重点突破复杂任务Agent。同时，大量创新Agent也表现不俗，如Manus作为通用场景Agent，Lovart深度垂直于设计场景，Flowith的画布式交互，实现无限流。多个产品已经形成可观收入。

互联网巨头持续加大AI基础设施投资，算力租赁厂商受益明显。阿里巴巴预计未来三年，将投入超过3800亿元，用于建设云和AI硬件基础设施，总额超过去十年总和。腾讯预计2025年资本开支持续上行，主要满足公司AI相关需求。目前已经有众多上市公司积极在算力租赁布局，部分公司已经披露相关订单，如海南华铁、有方科技、智微智能、协创数据、润建股份等。

投资建议：维持“优于大市”评级。AI Agent应用逐步走向市场，且已探索出商业模式，形成收入。重点关注在AI应用和Agent持续布局的厂商，如金山办公、合合信息、用友网络、税友股份、亚信科技、新大陆等。互联网巨头持续加大AI基础设施投资，算力租赁厂商受益明显，重点关注AI算力租赁产业链，如智微智能等。AI产业进展明显，维持“优于大市”评级。

风险提示：AI终端表现不及预期；下游IT支出收；行业竞争加剧。

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（亿元）	EPS 2025E	EPS 2026E	PE 2025E	PE 2026E
688111	金山办公	优于大市	274.87	1,273.14	4.10	5.01	67.04	54.86
688615	合合信息	优于大市	154.99	216.99	4.86	5.95	31.89	26.05
000997	新大陆	优于大市	29.91	308.69	1.18	1.42	25.35	21.06
603171	税友股份	优于大市	41.80	170.03	0.62	0.87	67.42	48.05
600588	用友网络	优于大市	13.31	454.80	-0.16	0.07	-	190.14
001339	智微智能	优于大市	47.68	119.40	1.20	1.59	39.73	29.99

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

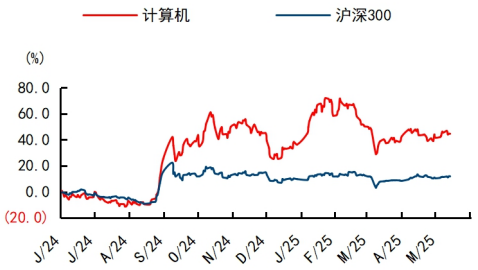
优于大市·维持

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

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

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

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

《稳定币香港政策落地，关注板块投资机会》——2025-06-04
《人工智能专题报告：国内大厂扩张资本开支，算力租赁订单持续落地》——2025-05-21
《计算机行业2025年5月投资策略暨财报总结-大厂布局Agent产品，AI应用快速落地》——2025-05-08
《人工智能行业专题：2025Q1海外大厂CapEx和ROIC总结梳理-20250505》——2025-05-05
《人工智能行业专题：美股大厂Q1业绩追踪，AI持续提供增长动能》——2025-05-01

内容目录

图表目录

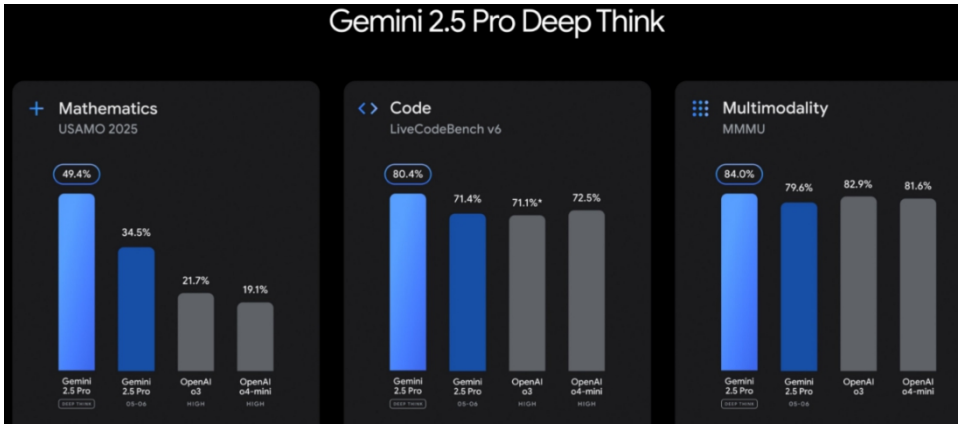
AI 产业持续迭代，Agent 开启应用生态

Google：重回 AI 舞台中心，Agent 生态野心初显

谷歌 I/O 开发者大会推出 Gemini 2.5 Pro，再次重回 AI 舞台中心。在 2025 年 5 月 21 日最新的谷歌 I/O 大会上,Gemini 及一系列产品发布再次证明谷歌 AI 地位。去年同期谷歌 AI 大模型和 API 每月处理 9.7 万亿个 Token，现在处理 Token 数增长至 480 万亿个,谷歌搜索业务的 AI 综述功能月活用户也达到了 15 亿人,Gemini 的 App 每月有 4 亿多活跃用户。5 月 28 日，谷歌宣布升级后的推理大模型 Gemini 2.5 Pro 版本正式可用。而该模型迅速在 LMSYS Arena 榜单中跃居第一，评分 1446 分，远超分数为 1409 分、1405 分的 o3 和 ChatGPT-4o。除了大模型之外，谷歌本次还发布了安卓、AI 眼镜、Agent 系统、视觉生成等多种一系列 AI 相关产品和更新。本次谷歌 AI 发布再次提振市场对 AI 产业信心，也验证在巨头林立的 AI 市场中，谷歌依然在舞台中心。

Gemini 2.5 Pro 技术进步显著，实现推理过程可视化。相比传统模型，Gemini 2.5 Pro 引入了“动态推理架构”，实现推理过程可视化。 Gemini 2.5 Pro 并非仅根据输入内容来生成答案，其会生成多个假设分支，模拟不同决策路径，来选择最优解，将模型决策拆解为可解释的逻辑节点。谷歌还发布了 Deep Think 版本，引入增强型推理机制，在数学、编程和多模态任务中均取得更好的成绩，在 USAMO 2025、LiveCodeBench、MMMU 等多项测试中，Gemini 2.5 Pro 深度思考版本表现均领先 Gemini 2.5 Pro。

图1：Gemini 2.5 Pro Deep Think 版本大幅提升推理水平



资料来源：谷歌官网，国信证券经济研究所整理

同时，谷歌也发布了 Gemini 2.5 Flash 低门槛版。Flash 版本专为速度和低成本而设计，支持边缘计算，在推理、多模态、代码和长上下文等关键基准上都得到了改进。Flash 版本使用的 token 减少了 20-30%，响应速度较之前提升 40%，效率提升了 22%。Gemini 2.5 Flash 专为实时响应场景设计，适用于文档摘要、图像标注、数据分类等高频任务。

图2：Gemini 2.5 Flash 各项数据提升明显

Benchmark		Gemini 2.5 Flash Preview (05-20) Thinking	Gemini 2.0 Flash	OpenAI o4-mini	Claude Sonnet 3.7 64k Ext. Thinking	Grok 3 Beta Extended thinking	DeepSeek R1
Input price	$/1M tokens	$0.15	$0.10	$1.10	$3.00	$3.00	$0.55
Output price	$/1M tokens	$0.60 No reasoning $3.50 Reasoning	$0.40	$4.40	$15.00	$15.00	$2.19
Reasoning & knowledge Humanity's Last Exam (no tools)		11.0%	5.1%	14.3%	8.9%	—	8.6%*
Science GPQA diamond	single attempt (pass@1)	82.8%	60.1%	81.4%	78.2%	80.2%	71.5%
	multiple attempts	—	—	—	84.8%	84.6%	—
Mathematics AIME 2025	single attempt (pass@1)	72.0%	27.5%	92.7%	49.5%	77.3%	70.0%
	multiple attempts	—	—	—	—	93.3%	—
Code generation LiveCodeBench v5	single attempt (pass@1)	63.9%	34.5%	—	—	70.6%	64.3%
	multiple attempts	—	—	—	—	79.4%	—
Code editing Aider Polyglot		61.9% / 56.7% whole / diff-fenced	22.2% whole	68.9% / 58.2% whole / diff	64.9% diff	53.3% diff	56.9% diff
Agentic coding SWE-bench Verified		60.4%	—	68.1%	70.3%	—	49.2%
Factuality SimpleQA		26.9%	29.9%	—	—	43.6%	30.1%
Factuality FACTS Grounding		85.3%	84.6%	62.1%	78.8%	74.8%	56.8%
Visual reasoning MMMU	single attempt (pass@1)	79.7%	71.7%	81.6%	75.0%	76.0%	no MM support
	multiple attempts	—	—	—	—	78.0%	no MM support
Image understanding Vibe-Eval (Reka)		65.4%	56.4%	—	—	—	no MM support
Long context MRCR v2	128k (average)	74.0%	36.0%	49.0%	—	54.0%	45.0%
	1M (pointwise)	32.0%	6.0%	—	—	—	—
Multilingual performance Global MMLU (Lite)		88.4%	83.4%	—	—	—	—

Methodology

Gemini results: All Gemini scores are pass @1 (no majority voting or parallel test time compute unless indicated otherwise). They are all run with the AI Studio API for the model-id gemini-2.5-flash-preview-05-20 and gemini-2.0-flash with default sampling settings. To reduce variance, we average over multiple trials for smaller benchmarks. Vibe-Eval results are reported using Gemini as a judge.

Non-Gemini results: All the results for non-Gemini models are sourced from providers' self reported numbers unless mentioned otherwise below. All SWE-bench Verified numbers follow official provider reports, using different scaffoldings and infrastructure. Google's scaffolding includes drawing multiple trajectories and re-scoring them using model's own judgement.

Thinking vs not-thinking: For Claude 3.7 Sonnet: GPQA, AIME 2024, MMMU come with 64k extended thinking, Aider with 32k, and HLE with 16k. Remaining results come from the non thinking model due to result availability. For Grok-3 all results come with extended reasoning except for SimpleQA (based on xAI reports) and Aider.

Single attempt vs multiple attempts: When two numbers are reported for the same eval higher number uses majority voting with n=64 for Grok models and internal scoring with parallel test time compute for Anthropic models.

Result sources: Where provider numbers are not available we report numbers from leaderboards reporting results on these benchmarks: Humanity's Last Exam results are sourced from https://agi.safe.ai/ and https://scale.com/leaderboard/humanitys_last_exam, AIME 2025 numbers are sourced from https://matharena.ai/. LiveCodeBench results are from https://livecodebench.github.io/leaderboard.html (10/1/2024 - 2/1/2025 in the UI). Aider Polyglot numbers come from https://aider.chat/docs/leaderboards/. FACTS come from https://www.kaggle.com/benchmarks/google/facts-grounding. For MRCR v2 which is not publically available yet we include 128k results as a cumulative score to ensure they can be comparable with previous results and a pointwise value for 1M context window to show the capability of the model at full length.

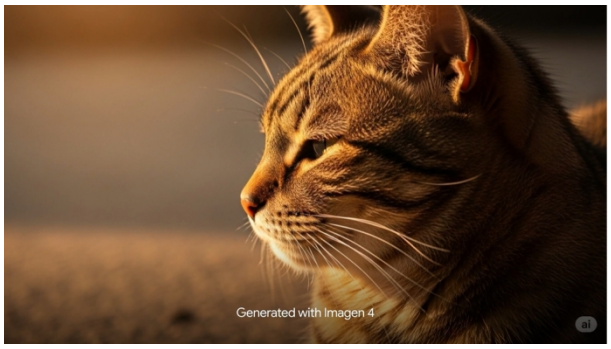
API costs are sourced from providers' website and are current as of May 20th.

* indicates evaluated on text problems only (without images)

资料来源：谷歌官网，国信证券经济研究所整理

谷歌在图片、视频、音频领域均升级明显，多模态能力令人惊喜。文生图领域，谷歌发布 Imagen 4，比上一代快 10 倍，图像细节更精致丰富，分辨率高达 2K，在文字在拼写和排版方面也得到了显著提升。谷歌同时发布了视频生成模型 Veo3，对标 OpenAI Sora；Veo3 是首次可以生成带有音频的视频，在音画同步、画面细节、物理模拟等多方面表现惊艳，已开放给 71 个国家的用户使用。谷歌充分融合了多模态能力，推出 AI 电影创作应用 Flow，将用 Veo + Imagen + Gemini 能力进行融合，展现其电影级的画面内容生成能力。这一系列工具产品有望在内容创作、广告影视、教育等多个行业产生深远影响。除此之外，本次发布会，谷歌还发布了 Project Aura AR 眼镜、Gemini Live、编程智能体 Jules 等多款产品，一扫前期被 OpenAI 压制的阴霾，多项 AI 能力实现反超。

图3：谷歌 Imagen 4 细节提升显著



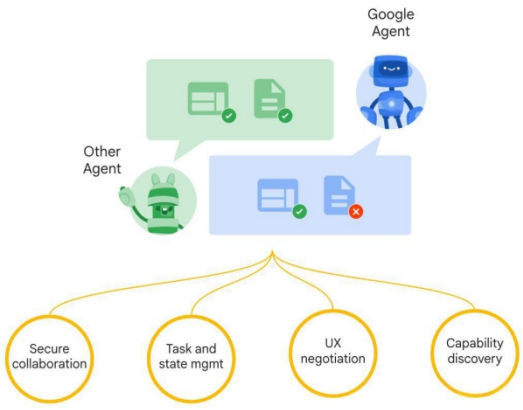
资料来源：谷歌官网，国信证券经济研究所整理

图4：通过 Flow 和 Veo3 制作电影

资料来源：谷歌官网，国信证券经济研究所整理

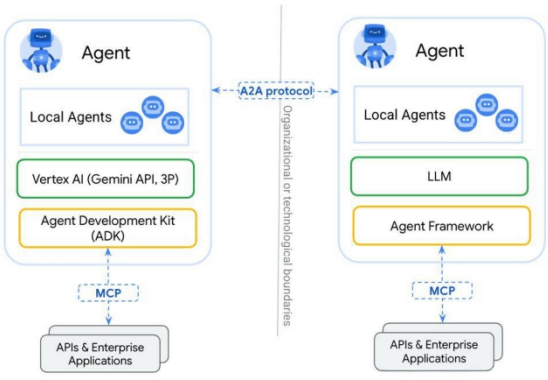
谷歌开源 A2A 协议，携手 MCP 打造 AI Agent 应用新生态。在 2025 年 4 月的 Google Cloud Next 25 大会上，谷歌开源了 Agent2Agent（A2A）协议。相较于 MCP 协议实现的 Agent 与工具和 API 的连接，A2A 协议目标打破系统孤岛，为 Agent 之间提供了一种标准交互方式，使它们能够相互协作，可以在各种底层平台上执行动作。A2A 协议是通过让客户端 Agent 和远程 Agent 之间通信来实现的，客户端负责制定和传达任务，远程端负责执行任务。同时，A2A 还支持 Agent 之间相互发送消息，这些消息可以包含上下文信息、回复、工件或者用户指令，以支持更好的共同完成复杂任务。A2A 一经发布就获得了大量厂商加入，包括埃森哲、波士顿咨询集团、凯捷、科尼、Salesforce、Atlassian、Intuit、MongoDB、甲骨文、SAP、麦肯锡等 50 多家著名企业。Agent 成为 AI 产业发展最确定性趋势，谷歌此举有望复制曾经安卓生态。

图5：A2A 工作原理



资料来源：谷歌官网，国信证券经济研究所整理

图6：Agent 和 Agent 之前的协同工作

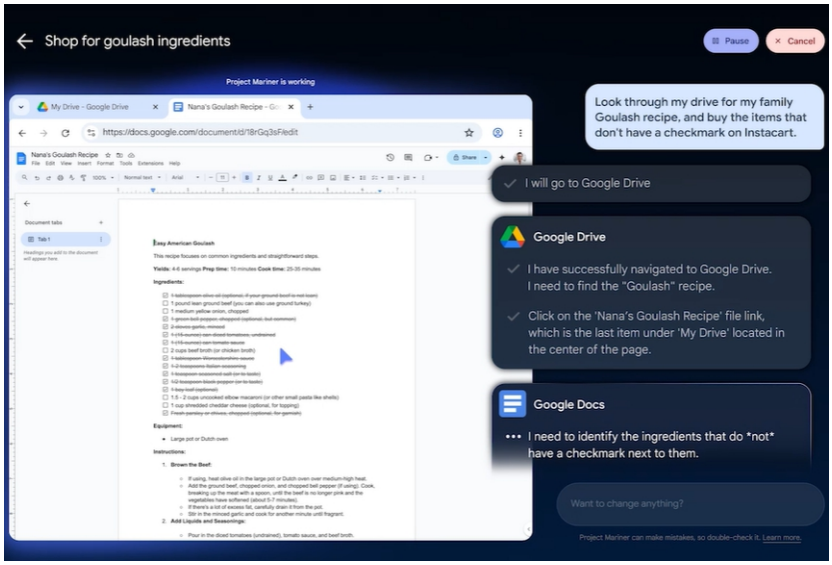


资料来源：谷歌官网，国信证券经济研究所整理

谷歌推出 AI Agent，将通过 Gemini API 开放给开发者，扩大谷歌 AI 生态。本次大会上，谷歌 Project Mariner AI 智能体也即将上线，此前命名为 Jarvis（贾维斯）。与 OpenAI 的 Operator 智能体类似，Project Mariner 也是一个用于网络的 AI 智能体，运行在浏览器中的虚拟机上，根据用户指令，行程规划和目标，并采取行动。目前 Project Mariner 可同时处理 10 项任务，例如用户提出采购

特定食物需求后，其完成“查找食谱 → 生成购物清单 → 在线下单”的全流程。目前，Project Mariner 使用工具现已登陆 Gemini API，Gemini SDK 现在兼容 MCP 协议，智能体模式即将来到 Chrome、搜索和 Gemini 应用。

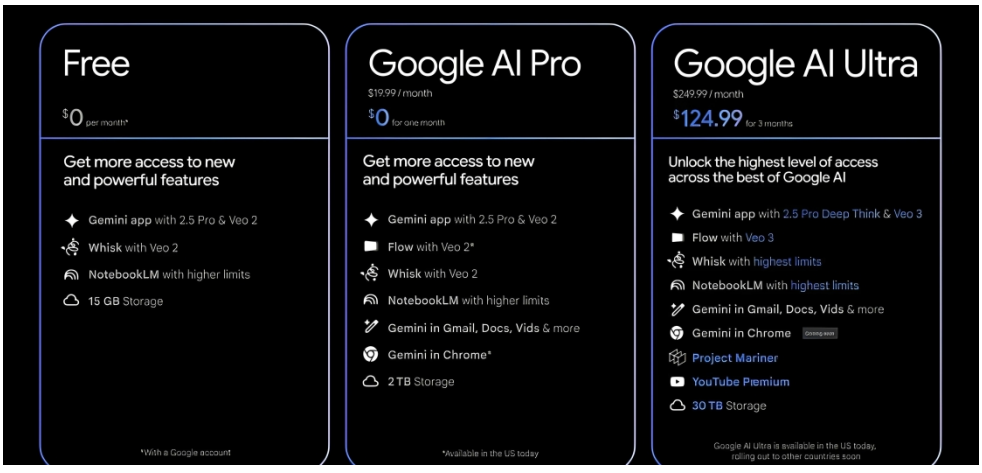
图7：谷歌 AI Agent 产品 Project Mariner



资料来源：谷歌官网，国信证券经济研究所整理

谷歌推出两款 AI 订阅，Ultra 版本定价高于 ChatGPT Pro 近 50 美元。基于本次 I/O 大会众多产品升级和推出，谷歌也推出了两个 AI 订阅等级。Google AI Pro 订阅价格为每月 19.99 美元，包括 Gemini 2.5 Pro、视频生成模式 Veo 2，以及 2TB 云存储。Google AI Ultra 订阅价格为每月 249.99 美元，包括最新的 Gemini 2.5 Pro 深度思考模式、视频生成模式 Veo 3，以及 30TB 云存储。Ultra 版本主要集成了谷歌全栈式 AI 能力，尤其是面向创意专业人士、开发者与学者，价格高于 OpenAI（200 美元/月）和 Anthropic（200 美元/月）。

图8：谷歌 AI 定价

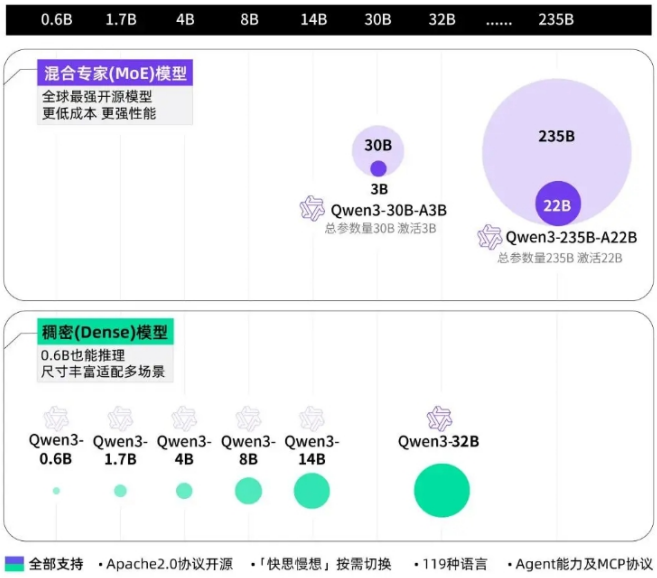


资料来源：谷歌官网，国信证券经济研究所整理

阿里：Qwen3 性价比再大幅提升，BC 两端 Agent 生态加速

阿里 Qwen3 发布，包含六款密集模型与两款混合专家模型。2025 年 4 月 29 日，阿里正式发布 Qwen3，标志着阿里巴巴首次推出混合推理模型。Qwen3 在推理、指令遵循、工具运用及多语言任务方面能力显著提升。Qwen3 六款密集模型，参数量分别为 0.6B、1.7B、4B、8B、14B、32B；两款 MoE 模型分别为 30B 总参数、3B 激活参数；235B 总参数、22B 激活参数。Qwen3 开源全系列模型，一经发布便登顶全球最强开源模型。尤其在性能和成本的上实现了惊人提升，以 DeepSeek-R1 三分之一的参数规模，就达成了性能的全面超越，仅需 4 张 H20 GPU 便能部署完整功能的 Qwen3 模型。

图9：Qwen3 8 款不同尺寸模型



资料来源：IT 之家，国信证券经济研究所整理

Qwen3 预训练的数据规模更大、质量更高。Qwen3 使用的数据量几乎是 Qwen2.5 的两倍，达到了约 36 万亿个 token，涵盖了 119 种语言和方言。数据集包括从网络上收集的数据和从 PDF 文档中提取信息，并且使用 Qwen2.5-VL 从文档中提取文本，用 Qwen2.5 改进提取内容的质量。同时为了增加数学和代码数据的数量，利用 Qwen2.5-Math 和 Qwen2.5-Coder 这两个数学和代码领域的专家模型合成数据，包括教科书、问答对以及代码片段等多种形式的数据。模型的训练分三个阶段，分别是：

阶段一：学习基本的语言技能和通用知识。在超过 30 万亿个 token 上进行了预训练，上下文长度为 4K token。

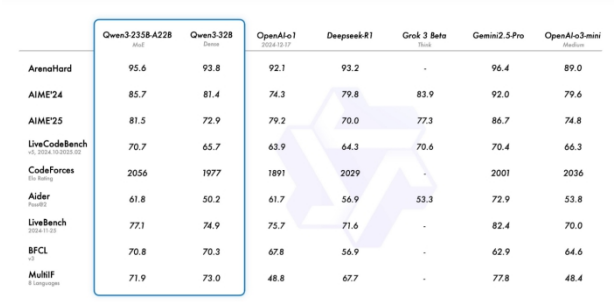
阶段二：通过增加知识密集型数据（如 STEM、编程和推理任务）的比例来改进数据集，随后模型又在额外的 5 万亿个 token 上进行了预训练。

阶段三：使用高质量的长上下文数据将上下文长度扩展到 32K token，确保模型能够有效地处理更长的输入。

由于模型架构的改进、训练数据的增加以及更有效的训练方法，Qwen3 模型整体

性能更优越。在 Dense 基础模型方面，Qwen3 Dense 基础模型的整体性能与参数更多的 Qwen2.5 基础模型相当，例如，Qwen3-1.7B/4B/8B/14B/32B-Base 分别与 Qwen2.5-3B/7B/14B/32B/72B-Base 表现相当。在 STEM、编码和推理等领域，Qwen3 Dense 基础模型的表现甚至超过了更大规模的 Qwen2.5 模型。与顶尖模型相比，旗舰版 Qwen3-235B-A22B 在代码、数学、通用能力等基准测试中，与 Deepseek-R1、OpenAI-o1 和 Gemini-2.5-Pro 等模型比表现出很强的竞争力。小型模型 Qwen3-30B-A3B 在激活参数数量是 QwQ-32B 10%的情况下，表现更胜一筹。

图10：旗舰模型在代码、数学、通用能力等基准测试



	Qwen3-235B-A22B MoE	Qwen3-32B Dense	OpenAI-o1 2024-12-17	Deepseek-R1	Grok 3 Beta Think	Gemini2.5-Pro	OpenAI-o3-mini Medium
ArenaHard	95.6	93.8	92.1	93.2	-	96.4	89.0
AIME'24	85.7	81.4	74.3	79.8	83.9	92.0	79.6
AIME'25	81.5	72.9	79.2	70.0	77.3	86.7	74.8
LiveCodeBench v5, 2024.10-2025.02	70.7	65.7	63.9	64.3	70.6	70.4	66.3
CodeForces Elo Rating	2056	1977	1891	2029	-	2001	2036
Aider Pass@2	61.8	50.2	61.7	56.9	53.3	72.9	53.8
LiveBench 2024-11-25	77.1	74.9	75.7	71.6	-	82.4	70.0
BFCL v3	70.8	70.3	67.8	56.9	-	62.9	64.6
MultiIF 8 Languages	71.9	73.0	48.8	67.7	-	77.8	48.4

资料来源：Qwen，国信证券经济研究所整理

图11：小型 MoE 模型在代码、数学、通用能力等基准测试

	Qwen3-30B-A3B MoE	QwQ-32B	Qwen3-4B Dense	Qwen2.5-72B-Instruct	Gemma3-27B-IT	DeepSeek-V3	GPT-4o 2024-11-20
ArenaHard	91.0	89.5	76.6	81.2	86.8	85.5	85.3
AIME'24	80.4	79.5	73.8	18.9	32.6	39.2	11.1
AIME'25	70.9	69.5	65.6	15.0	24.0	28.8	7.6
LiveCodeBench v5, 2024.10-2025.02	62.6	62.7	54.2	30.7	26.9	33.1	32.7
CodeForces Elo Rating	1974	1982	1671	859	1063	1134	864
GPQA	65.8	65.6	55.9	49.0	42.4	59.1	46.0
LiveBench 2024-11-25	74.3	72.0	63.6	51.4	49.2	60.5	52.2
BFCL v3	69.1	66.4	65.9	63.4	59.1	57.6	72.5
MultiIF 8 Languages	72.2	68.3	66.3	65.3	69.8	55.6	65.6

资料来源：Qwen，国信证券经济研究所整理

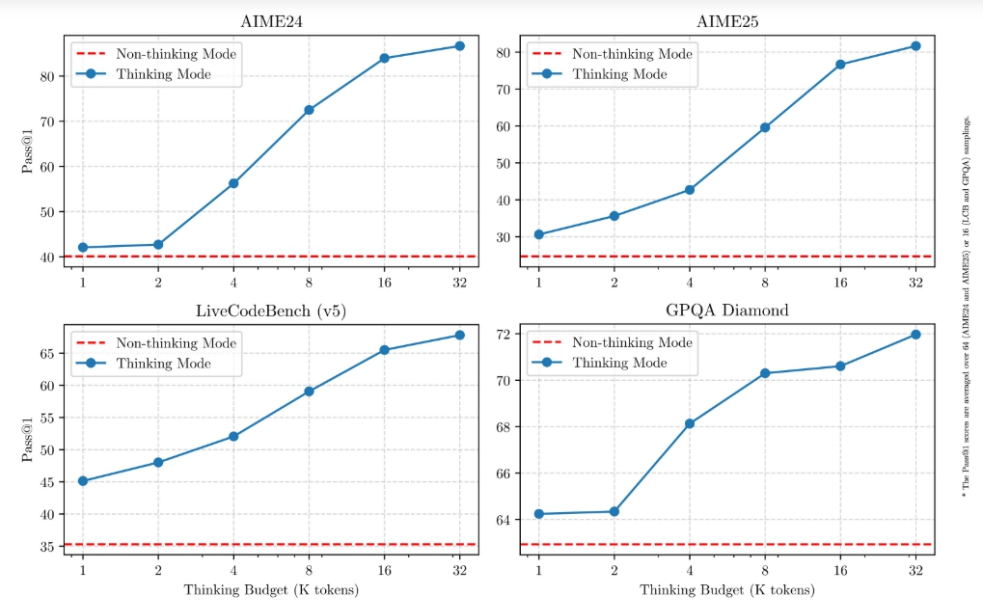
图12：多维度评测

	Qwen2.5-72B Base	Qwen2.5-Plus Base	LLaMA-4-Maverick Base	DeepSeek-V3 Base	Qwen3-235B-A22B Base
# Architecture	Dense	MoE	MoE	MoE	MoE
# Total Params	72B	271B	402B	671B	235B
# Activated Params	72B	37B	17B	37B	22B
General Tasks					
MMLU	86.06	85.02	85.16	87.19	**87.81**
MMLU-Redux	83.91	82.69	84.05	86.14	**87.40**
MMLU-Pro	58.07	63.52	63.91	59.84	**68.18**
SuperGPQA	36.20	37.18	40.85	41.53	**44.06**
BBH	86.30	85.60	83.62	86.22	**88.87**
Mathematics & Science Tasks					
GPQA	45.88	41.92	43.94	41.92	**47.47**
GSM8K	91.50	91.89	87.72	87.57	**94.39**
MATH	62.12	62.78	63.32	62.62	**71.84**
Multilingual tasks					
MGSM	82.40	82.21	79.69	82.68	**83.53**
MMMLU	84.40	83.49	83.09	85.88	**86.70**
INCLUDE	69.05	66.97	73.47	**75.17**	73.46
Code tasks					
EvalPlus	65.93	61.43	68.38	63.75	**77.60**
MultiPL-E	58.70	62.16	57.28	62.26	**65.94**
MBPP	76.00	74.60	75.40	74.20	**81.40**
CRUX-O	66.20	68.50	77.00	76.60	**79.00**

资料来源：Qwen，国信证券经济研究所整理

Qwen3 具有 3 个核心亮点，具备多种思考模式、多语言以及增强的 Agent 能力。作为国内首个实现“混合推理”的模型，Qwen3 深度融合了人类直觉思维与逻辑推演机制，模型支持两种思考模式：1）思考模式：在这种模式下，模型会逐步推理，经过深思熟虑后给出最终答案。这种方法非常适合需要深入思考的复杂问题。2）非思考模式：在此模式中，模型提供快速、近乎即时的响应，适用于那些对速度要求高于深度的简单问题。在强大数据量的支持下，Qwen3 模型支持 119 种语言和方言。同时优化了 Qwen3 模型的 Agent 和代码能力，也加强了对 MCP 的支持。

图13：Qwen3 两种思考模式



资料来源：Qwen，国信证券经济研究所整理

Qwen3 模型广泛覆盖多元应用场景，为各类需求提供精准适配方案。对于本地测试及科研范畴，Qwen3 - 0.6B 与 Qwen3 - 1.7B 凭借较低硬件要求，为快速实验搭建便利平台；在手机端侧应用场景内，Qwen3 - 4B 有效平衡性能与效率，为移动端部署提供理想选择；针对电脑或汽车端的对话系统、语音助手等应用情境，Qwen3 - 8B 可充分满足其功能需求；企业落地场景，面对复杂任务挑战，Qwen3 - 14B 与 Qwen3 - 32B 以卓越性能实力从容应对；而在云端高效部署方面，MoE 架构的 Qwen3 - 30B - A3B 速度出众，Qwen3 - 235B - A22B 则凭借强劲性能与低显存占用优势，成为该场景的优选方案。

表1：Qwen 稠密模型应用场景

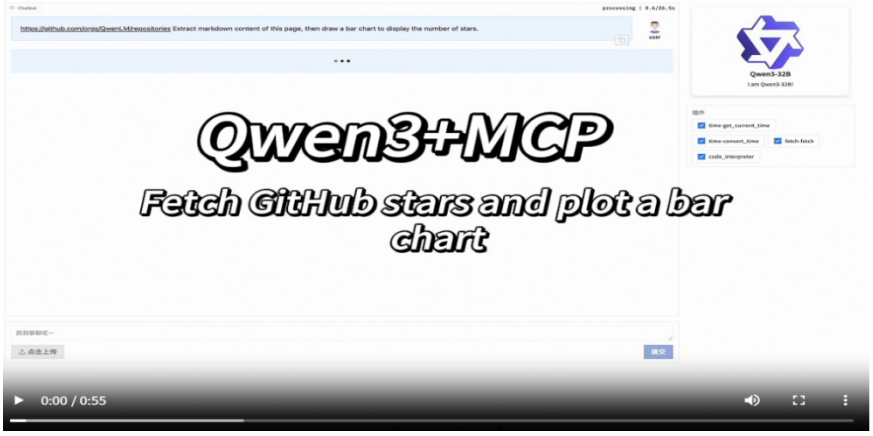
参数规模	适用场景	上下文长度	性能对标
0.6B	科研实验 / 边缘计算	32K	Qwen2.5-1.2B-Base
1.7B	算法验证	32K	Qwen2.5-3B-Base
4B	手机端侧部署	32K	Qwen2.5-7.2B-Instruct
8B	车载 / IoT 设备	128K	Qwen2.5-14B-Base
14B	中型应用	128K	Qwen2.5-32B-Base
32B	企业级部署	128K	跨级超越 Qwen2.5-72B

资料来源：IT 之家，国信证券经济研究所整理

Qwen3 原生支持 MCP，能够更加准确的识别外部函数和进行多工具的串联和并联调用，具备高效的 Agent 开发性能。根据 Qwen 官方的示例，在思考模式下，让 Qwen3 旗舰版调用 MCP 工具，统计并绘制某个 Github 项目的历史新增增长图，Qwen3 能自主围绕复杂任务进行思考和拆解，并围绕 5 项外部工具进行任务规划和工具调用，示例中 Qwen3 表现出的 Agent 性能非常卓越。Qwen-Agent 内部封装了工具调用的模板和工具调用解析器，大大降低了代码复杂性。在 Agent 测评中，Qwen3 创造了 70.8 分的 BFCL 评测新纪录，超越谷歌 Gemini、OpenAI。根据 Qwen

团队表示，Qwen 正从专注于训练模型的时代过渡到以训练 Agent 为中心的时代。Agent 的全面广泛应用，将是 AI 产业发展下一驱动力。

图14：Qwen3+MCP



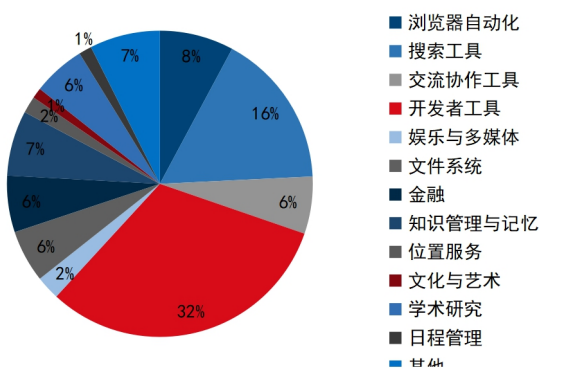
资料来源：Qwen，国信证券经济研究所整理

阿里前瞻成立魔搭社区，MCP 应用快速增长。魔搭社区成立于 2022 年 6 月，是一个模型开源社区及创新平台，由阿里巴巴通义实验室联合 CCF 开源发展委员会，共同作为项目发起方。目前，魔搭社区模型总量已超过 5 万个，涵盖 LLM、对话、语音、文生图、图生视频等多个领域，已服务超过 1300 万开发者。2025 年 4 月 15 日，魔搭社区推出全新 MCP 广场，上架千余款 MCP 服务，包括支付宝、MiniMax 等全新 MCP 服务在魔搭独家首发。截止同年 6 月 7 日，MCP 服务数量已提升至 3524 个，Agent 应用呈现快速膨胀之势。目前根据社区对 MCP 服务分类，开发者工具、搜索工具、浏览器自动化等应用占比较高。热度最高的 MCP 服务为 Fetch 网页内容抓取、高德地图、12306-MCP 车票查询工具、支付宝 MCP 等。

图15：社区上线大量 MCP 服务

资料来源：魔搭社区，国信证券经济研究所整理

图16：各 MCP 服务占比（截止 2025 年 6 月 7 日）



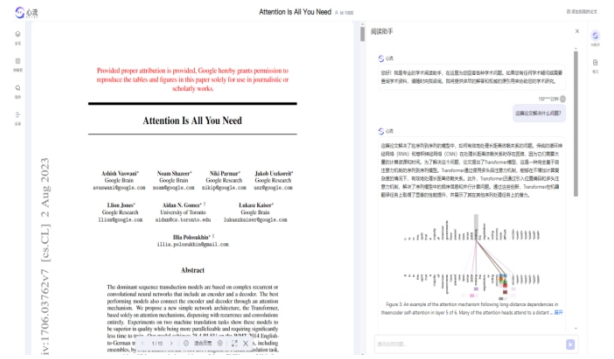
资料来源：魔搭社区，国信证券经济研究所整理

阿里 C 端 AI 进展：心流 Agent 主打科研，新夸克升级成为 C 端流量入口。阿里在 2024 年推出 AI 搜索助手“心流”，定位于科研人员、高校师生、职场人员使用，核心功能包括智能搜索问答、长文本分析及辅助创作，支持学术论文的检索、总结等。心流 AI 背后由淘宝星辰大模型支持，据统计，心流 AI 的长文本阅读准

确率超过 99%，搜索问答能力超越 GPT-4，构建了千亿级别的专业知识图谱，语义理解能力领先市场。

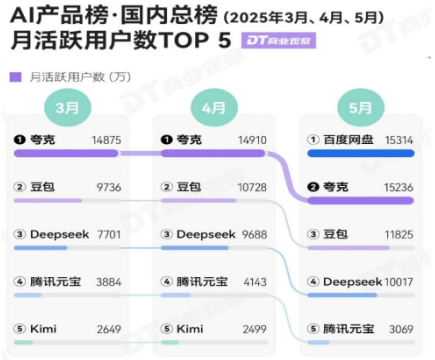
阿里将夸克全面升级为“无边界的 AI 超级框”。2025 年 3 月，阿里将夸克升级为“AI to C”的核心产品，也首次设立专职 CEO，在阿里 AI 战略生态中地位显著提升。新夸克基于阿里通义的推理及多模态大模型，5 月再次升级“深度思考”功能。根据 AI 产品榜，今年 3 月、4 月，夸克以接近 1.5 亿的月活跃用户，超过豆包登顶第一；5 月新纳入榜单的百度网盘排名第一，夸克第二，但差距较小，目前月活超过 1.5 亿。最初夸克仅是一款轻量级浏览器产品，随着不断和 AI 技术的融合，在搜索、网盘、文档等领域均获得用户认可。新夸克已不再是一款单一工具，而是 AI Agent 应用代表，其可自动识别用户意图，调用不同模型和工具来完成任务。自 2023 年阿里将夸克列为四大战略级创新业务以来，夸克的用户规模以每年翻倍的速度增长，00 后用户占比超过 50%，在 2024 年中国市场移动端 AI 应用中占比最高。

图17：心流对长文本论文泛读



资料来源：心流官网，国信证券经济研究所整理

图18：25 年 3-5 年 AI 产品榜



资料来源：AI 产品榜，国信证券经济研究所整理

阿里 B 端 AI 进展：大模型和 Agent 本地化落地，合作伙伴至关重要，亚信科技成为阿里云 AI 闭环关键一环。随着 Qwen3 模型尺寸变小，且性能更强，国内环境下 AI 本地化部署仍是 AI 产业发展的重要领地。在面对千行百业的个性化需求时，阿里需要合作伙伴帮其做用户需求梳理和定制化开发，目前亚信科技是其核心合作伙伴，双方在软件和硬件生态均进行了合作。近期，亚信科技和阿里云联合推出“算力+平台+应用+服务”四位一体的大模型一体机解决方案，面向各行业提供从需求梳理、规划设计到私有化部署的“开箱即用”的大模型软硬一体全流程产品服务。其中阿里基于自身软硬件优势，一体机具备极高的性价比，支持单机 16 卡轻量部署，支持全精度 16/8/4-bit 下高并发满血版 DeepSeek-R1 671b；生态友好，适配多种主流框架。亚信科技提供覆盖需求梳理、方案设计、数据赋能、模型优选、系统集成、效果验证等的全流程服务，解决 Qwen 和 Agent 落地企业最后一公里。目前双方在政务、电力、制造、石化等多个行业形成典型案例，例如在制造企业，双方共同赋能客户的 AI 文件解析、智慧客服、BOM 物料搜索、智能招商采购，大幅提升运营效率。根据亚信科技披露数据，25Q1 大模型交付业务订单大幅增长，进一步验证 AI 落地产业趋势。

图19：阿里云大模型和亚信科技联合解决方案

资料来源：亚信科技官网，国信证券经济研究所整理

字节：发布多模态 Agent，Coze 空间开始商业化

字节跳动推出的 UI-TARS-1.5，一款基于视觉-语言模型构建的开源多模态智能体，它能够在虚拟世界中高效执行各类任务。该版本不仅在 7 个典型的 GUI 图形用户界面评测基准中取得 SOTA 表现，而且首次展现了其在游戏中的长时推理能力和在开放空间中的交互能力。UI-TARS-1.5 基于此前提出的原生智能体方案 UI-TARS，通过强化学习进一步增强了模型的高阶推理能力，使模型能够在“行动”前先进行“思考”，显著提升了模型在面对未知环境和任务时的泛化能力。

图20：UI-TARS-1.5 基准测试表现

Benchmark Type	Benchmark	UI-TARS-1.5	OpenAI CUA	Claude 3.7	Previous SOTA
Computer Use	OSworld (100 steps)	**42.5**	36.4	28	38.1 (200 step)
	Windows Agent Arena (50 steps)	**42.1**	-	-	29.8
Browser Use	WebVoyager	84.8	**87**	84.1	87
	Online-Mind2web	**75.8**	71	62.9	71
Phone Use	Android World	**64.2**	-	-	59.5

资料来源：字节跳动 Seed，国信证券经济研究所整理

UI-TARS-1.5 的核心功能包括增强视觉感知、System 2 推理机制、统一动作建模以及可自我演化的训练方法。这些技术支撑使得 UI-TARS-1.5 能够实现精准的 GUI 操作，并且在执行复杂任务时展现出优秀的多步规划与决策能力。此外，UI-TARS-1.5 还展示了以游戏为载体来增强基础模型推理能力的新愿景，通过“思考-再行动”机制可以像人类一样“打游戏”，并且在 Minecraft 这样的开放环境中表现出色。这种结合了视觉、推理、记忆与操作的一体化架构，让 UI-TARS-1.5 成为当前最具代表性的开源智能体框架之一。

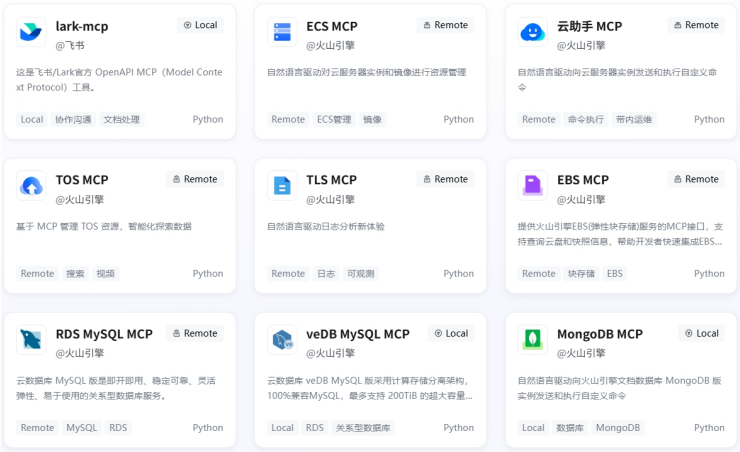
图21：UI-TARS-1.5 在 MineRL 中的测评表现

Task Type	Task Name	VPT	DreamerV3	Previous SOTA	UI-TARS-1.5 w/ o Thought	UI-TARS-1.5 w/ Thought
Mine Blocks		0.8	1.0	1.0	1.0	**1.0**
		0.0	0.0	0.0	0.2	**0.3**
		0.0	0.0	0.1	0.4	**0.6**
	200 Tasks Average	0.06	0.03	0.32	0.35	**0.42**
Kill Mobs		0.0	0.0	0.1	0.3	**0.4**
		0.4	0.1	0.6	0.7	**0.9**
		0.1	0.0	0.4	0.5	**0.6**
	100 Tasks Average	0.04	0.03	0.18	0.25	**0.31**

资料来源：字节跳动 Seed，国信证券经济研究所整理

火山引擎借力 AI 保持高速增长，全面支持 MCP 生态。2024 年火山引擎实现收入超 120 亿元，凭借 AI 领域的增长驱动力，2025 年营收目标超过 250 亿，未来目标将达到千亿收入规模。2024 年，中国公有云大模型调用量达 114.2 万亿 Tokens，其中火山引擎份额第一，高达 46.4%。2025 年 3 月，豆包大模型日均 tokens 调用量 12.7 万亿，同比增长近 100 倍。在此基础上，火山引擎于 2025 年 5 月发布大模型生态广场 MCP Servers，集成丰富工具实现全链路开发闭环，极大地简化了开发流程，开发者可高效构建 AI 应用。通过“MCP Market（工具广场）+ 火山方舟（大模型服务）+Trae（应用开发环境）”深度协同，AI 应用繁荣将进一步拉动云业务增长。目前 MCP Servers 已有 178 种服务，主要为开发者工具较多，火山引擎相关服务为主，也有飞常准（航班数据）、汉得精准营销（用户行为分析）等优质三方生态工具。

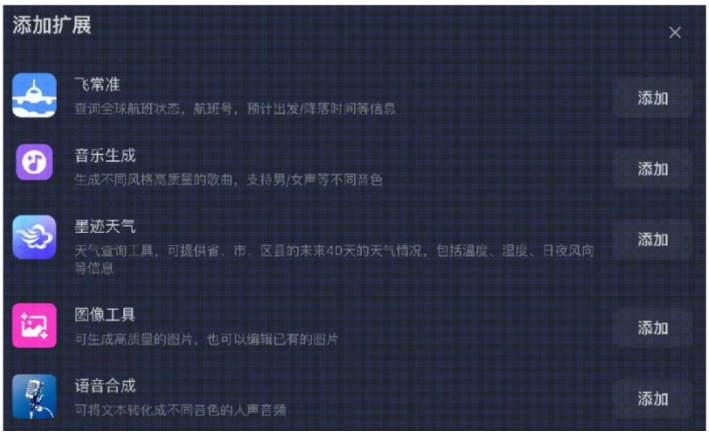
图22：火山引擎大模型生态广场 MCP Servers



资料来源：火山引擎官网，国信证券经济研究所整理

扣子空间开启内测，重点突破复杂任务 Agent。2025 年 4 月，字节跳动 Agent 产品扣子空间（Coze Space）正式开启内测，旨在“成为用户与 AI Agent 协同办公的最佳场所”。扣子空间能够自动分析用户需求并拆解为多个子任务，并调用工具（如浏览器、代码编辑器等）执行任务。扣子空间引入专家 Agent 体系，也提供通用 Agent 入口，并支持 MCP 生态。例如，在 Coze agent 中可以添加 MCP 扩展，如“水滴信用”，可以快速在任务中查询企业的工商信息、股权结构等信息。通过 MCP 的扩展，Coze 解决复杂任务能力大幅提升。

图23：Coze 添加 MCP 扩展



资料来源：Coze，国信证券经济研究所整理

Coze 商业化开始推进，AI Agent 将成为变现利器。随着 AI Agent 成为 AI 应用逐步落地的范本，海量企业和开发者将基于 Coze 等平台来进行自己 Agent 的搭建，Coze 商业化也迎来新的发展窗口。目前 Coze 推出个人免费版、个人进阶版、团队版以及企业版订阅套餐，各套餐的权益范围不同，采用包年包月+按量付费的混合计费模式。相较于个人版，Coze 团队版和企业版套餐提供更强大的功能支持，例如更高的资源额度、企业级安全特性等。

图24：Coze 会员价格

资料来源：Coze，国信证券经济研究所整理

大量创新 agent 开始涌现

Manus 向全球用户开放注册，推出“文生视频”功能。Manus 于 2025 年 3 月发布，是全球首个通用 Agent 产品。Manus 在 GAIA 基准测试中取得了 SOTA 的成绩，显示其性能超越 OpenAI 的同层次大模型，因此一经发布便一码难求。2025 年 5 月 12 日，Manus 宣布面向全球用户开放注册，无需等待名单，所有用户每天可免费执行一项任务（300 积分），所有用户一次性获得 1000 积分奖励。Manus 提供三档付费订阅方案，价格分别为每月 19 美元、39 美元和 199 美元。开放注册后，Manus 也由原来“有限内测”转为“广泛应用”阶段。2025 年 6 月，Manus 进一步推出原生的“文生视频”功能，可以在几分钟内将文本命令转换为井然有序的视频，目前已经面向付费会员用户开放抢先体验。

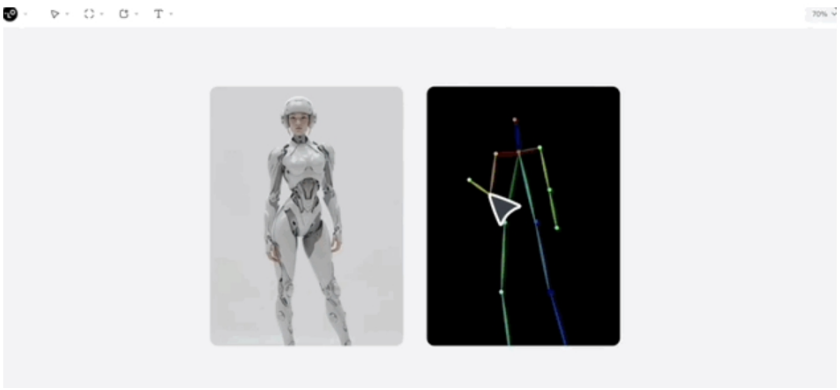
图25：Manus 定价

资料来源：Manus，国信证券经济研究所整理

Lovart 是全球首个专注于设计领域的 AI agent，与可灵 AI 达成合作。Lovart 由国内 AI 创作平台 Liblib AI 的海外子公司推出，一经发布，其全链路设计能力和无缝创意体验立刻受到全球设计师的追捧。Lovart 单次能生成多达 40 张设计成品图，相比于单纯的文生图，其更类似与专业设计师，协助完成整个设计任务。

Lovart 的核心是建立在最先进的思维链(MCoT)之上，通过自然语言交互，形成初步的设计方案，大幅降低了设计门槛。Lovart 平台还无缝集成了 AI 和非 AI 工具的完整矩阵--GPT4o、Stable Difusion、Flux、Triple、Ps、Figma 等等，允许设计师和创作者以完美的兼容性导入和导出任何设计格式。同时，Lovart 保持着对话式命令、预测性建议、增强型画布界面三层交互系统，使用户更方便的将灵感转化为视觉设计。近期，Lovart 与可灵 AI 达成深度合作，Lovart 可以调用可灵 API，将其作为核心引擎嵌入视频生成流程。

图26：Lovart 在视频、3D 领域创作表现



资料来源：量子位，国信证券经济研究所整理

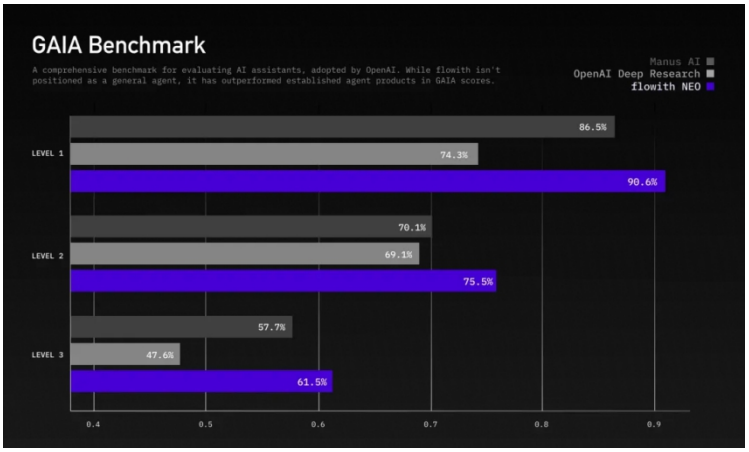
目标颠覆 Chatbot 模式，Flowith 打造画布式的 AI 创作工具。Flowith 于 2024 年 4 月由国内团队发布，已为数十万全球用户提供服务，根据 2025 年 5 月数据，其流量环比 +240%、MAU 环比+228%，用户保持快速增长。flowith 定位为画布式 AI 创作工具，主打多线程、非线性交互。用户可在同一界面同时与多个 AI 模型协作，支持长内容生成、多结果对比、Prompt 调试等复杂任务。Flowith 支持多个主流模型 GPT-4、Claude 3.5、Gemini Pro、DeepSeek R1 等，将用户各种杂乱知识进行“画布式”可视化整理，形成自己的知识花园。Flowith 应用场景主要在内容创作、产品设计、教育培训等方向，已推出多种收费版本。

图27：Flowith 画布式交互

资料来源：Flowith 官网，国信证券经济研究所整理

Flowith 发布 Agent NEO，全球首款无限流 AI 工具。NEO 于 2025 年 5 月 19 日正式发布，其是全球首个支持无限步骤、无限上下文、无限工具的 AI Agent。NEO 一经推出就受到了广泛的追捧，公司宣称在 0 营销投入的情况下，已实现了 130 万美元的 ARR 收入。在 GAIA 智能体评测标准的三个等级中，Neo 不仅超越了 Manus，还创下了新的评分记录。其三大特性来看，“无限步骤”：赋予了 Neo 支持超过 1000 个推理步骤，实现任务进程无限延长，完成长达数小时甚至数月的复杂任务，特别适用于超长周期项目。“无限工具”：通过 Oracle 框架，Neo 可根据需求动态调用无限数量的工具，涵盖图像生成、联网搜索、提示词优化等，满足多样化的任务需求。“无限上下文”：可支持长达 10M token 的上下文窗口，具备处理超大规模数据的能力，能够生成诸如长篇小说、复杂代码库等完整输出。NEO 的无限特性让其在多种场景与竞品拉开优势，但其对云端算力和网络依赖较大，尤其超长任务会出现成本上升。

图28：Flowith NEO 在 GAIA 智能体评测



资料来源：虎嗅，国信证券经济研究所整理

在大模型能力稳步迭代下，Agent 称为 AI 落地确定产业趋势，各类应用有望百花齐放。除了互联网巨头基于自身模型推出 Agent 产品之外，众多创新 Agent 持续涌现，尤其在 MCP、A2A 等协议开源后，通用和垂直 Agent 应用将进一步普及，开始在各类工作流中崭露头角。基于三个热门创新 Agent，腾讯科技进行了多个场景的测评，三大产品在定位和场景上均有明显差异。首先，Manus 和 Flowith 为通用 Agent，Manus 侧重交付完成数字化工作结果，通过分解任务，并调用一系列工具将结果落地；Flowith 也能完成通用任务，但是其强调“可视化协作”，侧重知识库的建设，并可通过无限步骤完成超大任务。而 Lovart 深度垂直于设计场景，拆解用户需求，并采用多模态能力完成设计工作。其次，在应用场景上，Manus 擅长知识类工作，如市场研究报告，法律文件阅读分析；Flowith 擅长信息量巨大且需要多人迭代的创作场景，例如通过大量文献，多人来完成产品研发；Lovart 的重点则在品牌视觉与内容营销上，例如生成海报和视频广告。另一方面，各 AI Agent 产品也有自身的局限性，仍需持续迭代，但其价值和商业模式已经逐步形成，产业趋势已确立。

表2：三大创新 Agent 对比

产品	Manus	Flowith [Agent NEO]	Lovart
一句话总结	通用型多代理(multi-agent)系统	面向“深度工作”的无限画布+"Infinite Agent"	首个专业 AI 设计 Agent
功能定位	可在云端异步执行长链任务(招聘筛简历、深度股研、自动建站等),并提供“Manus's computer”侧边栏实时回放。	大于 1000 推理步骤、大于 1000 万 token 上下文、可 24×7 后台跑多代理并并行分工。	文本→专业级视觉/视频/音乐“三模”一条龙；内置图层管理、局部重绘、文本分离等设计师向工具。
最强项	透明可回放的多代理架构所有浏览、填写表单、调用工具的步骤都能事后复现，方便审计与调优。	超长上下文+持续自治可把整本书或公司知识库喂给 Agent,让其数小时/数日连续推理并周期性回报。	全链路多模态调度一句话即可产出品牌 VI、动效片头、BGM,且能无缝导入 Figma/Photoshop。
典型应用场景	复杂研究报告与数据分析 旅行/活动全流程筹划 一键生成简历网站、营销落地页等	长篇内容生成与版本对比 超大规模文献整理 3D 互动网页/游戏快速搭建	从零到一产出视觉包：Logo+海报+短视频社媒快闪海报/meme 套图
并发与效率	人工用时 2h 的深度研究任务，可以在 8min 内交付	云架构支持 10 M-token 上下文且大于 1000 推理步骤； 生成全功能 3D 网页用时小于 5min	实测 5 min 内可以生成 4-6 张同风格海报；官方 Beta 测速多尺寸稿件“几分钟”交付，效率提升 5×以上。
引擎栈（底层模型）	Anthropic Claude 3.5 Sonnet、阿里 Qwen 等云端 LLM 组合	GPT-4o/o1-preview、Claude 3.5 、Stable Diffusion、DALL ·E3 等	GPT-4o、Stable Diffusion、Flux、Kling、Tripo、Suno +本地 PS/Figma API;自动模型选型保证风格一致性
能力弱项	长链规划几乎没有逐步日志，出错时用户难以定位问题； 在研究类任务中仍会出现工具中断、引用混淆等不稳定现象； 多模态能力和设计能力较弱。	大画布并行节点过多时常出现内存飙升、移动端卡顿与同步丢失； 交互逻辑比较复杂，初学者要花时间适应； 任务需在云端执行，本地离线模式尚未开放； 对要求数据完全封闭在内网的企业来说，Flowith 当前的云依赖比较难解。	生成的视频“一次成片”,无法对中间镜头做细粒度调整，过场效果常被吐槽“幻灯片感”； 只输出RGB数字稿,印刷前仍需人工做CMYK转换与色彩校正，否则品牌色易失真。

资料来源：腾讯科技，国信证券经济研究所整理

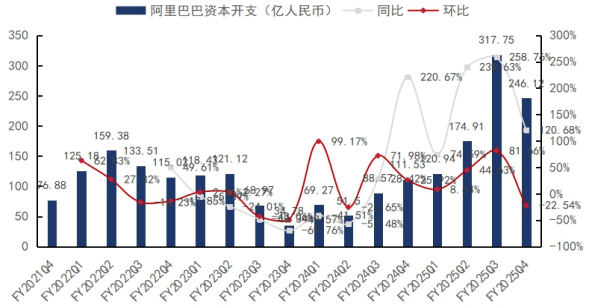
算力需求依然景气，租赁订单持续落地

互联网巨头保持高投入，自建和租赁算力并行

FY2025 阿里巴巴资本开支同比高增。阿里巴巴 FY2025 资本开支为 859.72 亿元人民币，同比+167.93%，资本开支快速扩张，用于 AI 相关基础设施建设（例如 AI 服务器、IDC 等),进一步夯实公司在 AI 云领域的优势地位;分季度来看,FY2025Q4 资本开支为 246.12 亿元，同比+120.68%、环比-22.54%，受贸易摩擦和供货节奏波动影响，环比下滑。

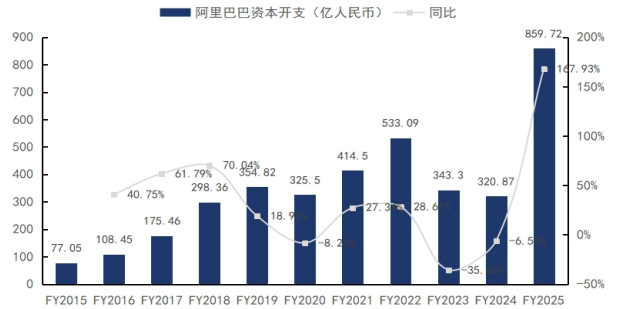
阿里巴巴未来 3 年资本开支指引乐观。2025 年 2 月，阿里巴巴集团 CEO 吴泳铭宣布，未来三年，阿里巴巴将投入超过 3800 亿元，用于建设云和 AI 硬件基础设施，总额超过去十年总和；因此，我们预计 2025 年阿里巴巴资本开支将持续扩张，AI 相关基础设施（包括算力租赁）等需求旺盛。

图29：阿里巴巴分季度资本开支（亿人民币）



资料来源：公司财报，国信证券经济研究所整理

图30：阿里巴巴年度资本开支（亿人民币）

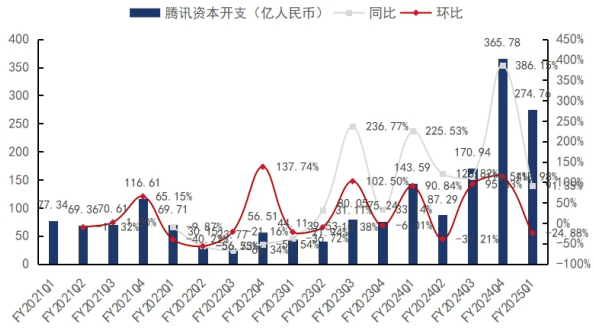


资料来源：公司财报，国信证券经济研究所整理

FY2025Q1 腾讯资本开支同比高增。腾讯 FY2025Q1 资本开支为 274.76 亿元人民币，同比+91.35%、环比-24.88%，同比高增，受贸易摩擦和供货节奏波动影响，环比下滑。从公司财年来看，FY2024 合计资本开支为 767.6 亿人民币，同比+221%，资本开支快速扩张。

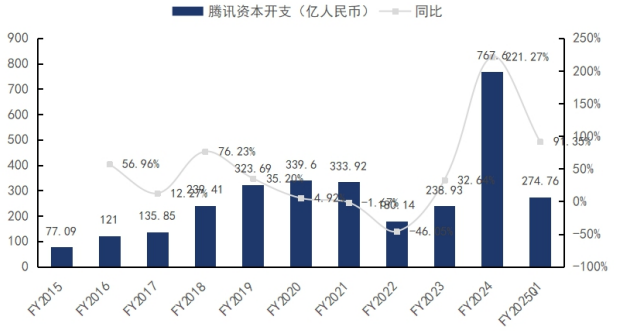
腾讯 2025 年资本开支持续上行。根据腾讯 2024 年财报披露数据，预计 2025 年资本开支持续上行，主要满足公司 AI 相关领域需求，包括：1）内部业务需求（例如广告、内容推荐、游戏等）；2）训练基础模型；3）为 AI 应用（元宝、微信 AI 等）提供推理支持；4）为外部客户提供云服务。综上，随着 2025 年腾讯资本开支上行，公司对 AI 相关基础设施建设需求持续提升。

图31：腾讯分季度资本开支（亿人民币）



资料来源：公司财报，国信证券经济研究所整理

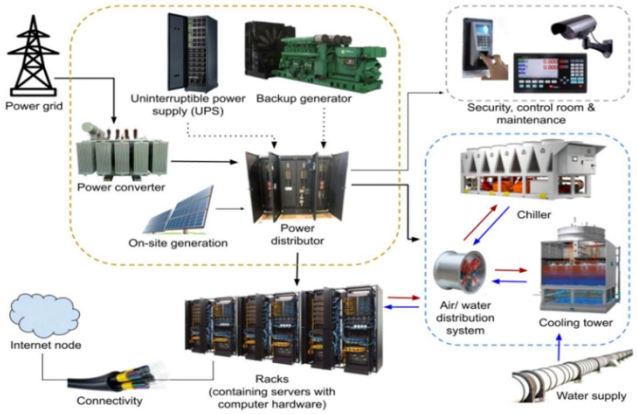
图32：腾讯年度资本开支（亿人民币）



资料来源：公司财报，国信证券经济研究所整理

互联网大厂对 AI 基础设施建设的需求快速提升，自建和租赁均积极投入。自建算力基础设施方面，互联网大厂需自己完成相关服务器、AI 芯片的采购，以及建设供电系统、冷却系统、安全及监控系统、网络系统等。租赁算力基础设施方面，互联网大厂通常会与被租赁公司签订算力服务合同，约定服务年限及合同金额，相关服务器、AI 芯片采购以及 IDC 等基建由被租赁公司完成。

图33：自建算力基础设施



资料来源：Konstantin 等著-《Compute at Scale - A broad investigation into the data center industry》-ArXiV（2023）-P6，国信证券经济研究所整理

各厂商积极响应，算力租赁订单持续落地

海南华铁：累计签订算力订单 66.7 亿元。公司 2024 年 5 月首次开展智算业务，拟投资 10 亿元，为客户提供 2400p 算力服务；2025 年 3 月 5 日，公司披露《算力服务协议》，新获算力订单 36.9 亿元；根据公司《海南华铁：浙江海控南科华铁数智科技股份有限公司投资者关系活动记录表 20250428》披露数据，2025 年一季度公司算力业务进展显著，新签算力订单 41.95 亿元，累计签订算力订单达 66.7 亿元，新交付算力资产 4.88 亿元，累计交付智算设备资产达 11.59 亿元。

表3：海南华铁算力租赁相关公告梳理

日期	公告	主要内容	订单额	投资额	算力规模	合同期限	实施主体
2024 年 5 月 7 日	《海南华铁：浙江华铁应急设备科技股份有限公司关于投资智算中心建设的公告》	公司拟投资 10 亿元开展智算业务，通过向客户提供 GPU 级的高端算力资源租赁及增值技术服务取得收益		10 亿元	2400p	3 年	上海科思翰智算智能技术
2024 年 12 月 5 日	《海南华铁：浙江华铁应急设备科技股份有限公司投资者关系活动记录表 20241205》	已累计签约算力服务金额已达 24.75 亿元，公司将根据合同约定向合作方提供为期 3-5 年的长期算力服务，合计交付智算设备超 6 亿元					
2025 年 3 月 5 日	《海南华铁：浙江海控南科华铁数智科技股份有限公司关于子公司签署《算力服务协议》的公告》		36.9 亿元	超 20 亿元		5 年	华铁大黄蜂
2025 年 4 月 22 日	《海南华铁：浙江海控南科华铁数智科技股份有限公司 2024 年年度报告》	累计签订算力服务金额 24.75 亿元（2025 年 3 月末已达 66.7 亿元）					
2025 年 4 月 28 日	《海南华铁：浙江海控南科华铁数智科技股份有限公司投资者关系活动记录表 20250428》	2025 年一季度公司算力业务进展显著，新签定算力订单 41.95 亿元，累计签订算力订单达 66.7 亿元，新交付算力资产 4.88 亿元，累计交付智算设备资产达 11.59 亿元					

资料来源：公司公告，国信证券经济研究所整理

有方科技：子公司同航锦科技深度合作。公司 2023 年 11 月中标中国电信宁夏公司算力服务项目，订单额 0.95 亿元；2024 年 7 月 26 日，公司控股子公司有方数据与航锦科技签署《战略合作协议》，双方拟统筹推进智算中心、数据存储、数据灾备、融合计算等业务的合作落地，在智算中心业务拓展方面，航锦科技在智算算力集群租赁及服务的市场拓展、建设、运营和维护等方面逐渐积累了技术和经验，有方数据在高性能存储服务器、存储软件等云基础设施方面逐渐积累了产品和技术，双方将在智算和存储市场进行协同合作，共同拓展市场，共同推广双方优势产品和服务。

表4：有方科技算力租赁相关公告梳理

日期	公告	主要内容	订单额	实施主体
2023 年 11 月 1 日	《有方科技：关于自愿披露项目中标的公告》	公司此前参与了中国电信宁夏公司 2023 年算力服务集中采购项目的公开招标，确认公司中标项目标段 2（招标编号：08-11-04A-2023-D-A20771），不含税预估金额为人民币 9,500 万元，后续招标人将按照招标文件和投标文件与公司签订合同	0.95 亿	有方科技
2024 年 7 月 26 日	《有方科技：关于子公司签订《战略合作协议》的自愿性披露公告》	控股子公司有方数据近日与航锦科技签署了《战略合作协议》，双方拟共同推进智算中心、数据存储、数据灾备、融合计算等业务的合作落地		有方数据

资料来源：公司公告，国信证券经济研究所整理

智微智能：成立控股子公司，深入 AIGC 基础设施建设。根据公司公告披露数据，24 年 1 月公司出资设立子公司南宁市腾云智算，注册资本 2000 万元，公司以货

币资金出资 1020 万元，持有腾云智算 51%的股权。腾云智算定位为 AIGC 基础设施全生命周期服务商，围绕 AI 算力规划与设计、设备交付、运维调优、算力租赁、算力调度管理、设备维保及置换等提供端到端的智算中心全流程综合服务，目前主要服务于互联网大厂及运营商等对训练算力要求较高的客户。根据 2024 年报披露数据，24 年子公司腾云智算实现营业收入 3.0 亿元，净利润 1.74 亿元，净利率为 58%，盈利能力优秀。

表5：腾云智算 2024 年财务情况

子公司名称	流动资产	非流动资产	资产合计	流动负债	非流动负债	负债合计
	4.27	0.13	4.4	2.37	0.09	2.46
腾云智算	营业收入	净利润	综合收益总额		经营活动现金流	
	3	1.74	1.74		10.27	

资料来源：公司财报，国信证券经济研究所整理

润建股份：绑定阿里大客户。公司 2023 年 7 月，在五象云谷云计算中心基础上，拟投入资金 2 亿元打造智能算力中心，提供最高可达 2533Pops（Int8）定点算力或 43Pflops(FP32)单精度浮点算力及配套云存储，服务于人工智能大模型、行业模型等；2023 年 11 月，公司发布公告，公司控股子公司五象云谷与阿里云就算力服务和数字化云签署了《合作协议》，由润建股份投资 2500P 算力服务器部署在五象云谷，由五象云谷提供算力基础底座所需的数据中心电力、制冷等基础环境，由阿里云参与建设“中国-东盟智算云”统一平台，对算力底座统一管理、统一运维和统一运营。

表6：润建股份算力租赁相关公告梳理

日期	公告	主要内容	投资额	算力规模	合同期限	实施主体
2023 年 7 月 24 日	《润建股份：智能算力中心可行性分析报告》	本项目拟建设智能算力中心，提供最高可达 2533Pops（Int8）定点算力或 43Pflops(FP32)单精度浮点算力及配套云存储。建成后主要提供 AI 大模型训练、推理算力、图形渲染算力服务，为人工智能大模型、行业模型等，公司将按客户需求进行调整具体投入。	2 亿	2533Pops（Int8）或 Pflops(FP32)		润建股份
2023 年 11 月 22 日	《润建股份：关于与阿里云计算有限公司签订合作协议的公告》	润建股份控股子公司五象云谷与阿里云就算力服务和数字化云签署了《合作协议》，由润建股份投资 2500P 算力服务器部署在五象云谷，由五象云谷提供算力基础底座所需的数据中心电力、制冷等基础环境，由阿里云参与建设“中国-东盟智算云”统一平台，对算力底座统一管理、统一运维和统一运营		2500P	三年	五象云谷

资料来源：公司公告，国信证券经济研究所整理

协创数据：算力租赁先行者，发布算力采购大单。根据 2024 年 3 月 29 日公司披露的《协创数据：2024 年 3 月 29 日投资者关系活动记录表（2023 年度网上业绩说明会）》，公司已经开始布局算力租赁业务。2024 年 10 月，公司间接控股子公司广州奥佳向上海域允采购 GPU 服务器，预计采购金额不超过 9 亿元；2025 年 3 月，公司拟向多家供应商采购服务器，采购合同总金额预计不超过人民币 30 亿元，主要用于为客户提供算力租赁。

表7：协创数据算力租赁相关公告梳理

日期	公告	主要内容	投资额	实施主体
2024 年 3 月 29 日	《协创数据：2024 年 3 月 29 日投资者关系活动记录表（2023 年度网上业绩说明会）》	公司已布局服务器租赁业务		
2024 年 8 月 23 日	《协创数据：关于控股子公司签署云算力服务框架协议的自愿性披露公告》	协创数据控股子公司麦塔倍斯于近日与头部互联网公司签订了《云算力服务框架协议》，麦塔倍斯按双方约定为客户 A 的应用程序提供云算力服务		麦塔倍斯
2024 年 10 月 11 日	《协创数据：关于公司算力业务进展情况的自愿性披露公告》	2024 年 6 月，公司与优必达、优崴签订《云业务合作协议》的三方协议，合作领域包含以高端算力为基础的云算力租赁、云安防和大模型的合作、面向跨境电商的 AIGC 等业务领域。 2024 年 6 月，公司与中移国际签订《云业务合作协议》，双方将整合公司在海外区域的算力资源、云业务平台、高可用 IT 服务等业务能力，以及中移国际在全球 DICT 方面的技术、资源和运营的综合服务优势，共同打造高效云服务体系		协创数据
2024 年 10 月 24 日	《协创数据：关于公司算力业务进展情况的自愿性披露公告》	2024 年 10 月，因规划建设具备大模型训练和推理能力的大型算力服务集群（万卡级），公司间接控股的子公司广州奥佳与上海域允签署《采购框架合同》，向上海域允采购 GPU 服务器，包括 H20 NVLINK 型 AI GPU 服务器，预计采购金额不超过 90,000 万元	9 亿	广州奥佳
2024 年 10 月 24 日	《协创数据：关于与合肥综合性国家科学中心人工智能研究院签署战略框架协议的自愿性披露公告》	协创数据于近日与合肥综合性国家科学中心人工智能研究院签订了《战略合作框架协议》，由协创数据提供经费和服务器等算力基础设施并搭建算力平台，乙方提供包括 AI 技术、AGI 技术在内的智能算力技术资源		协创数据
2024 年 12 月 27 日	《协创数据：关于签署战略合作框架协议暨关联交易的自愿性披露公告》	协创数据及控股子公司奥佳软件，近日与张江集团、关联方奥飞数据签订了《战略合作框架协议书》，协创数据、奥飞数据、奥佳软件，三方发挥在 AI 智能体、训推一体、专有云建设及运营方面的丰富经验，为张江集团园区企业提供计算（CPU、 GPU 等）、存储、网络、数据库、安全等各类聚焦自主可控、丰富多元的云产品及 AI 智能体训推一体资源池，提供数据中心机柜租用、场地定制化服务、高电高密解决方案、基础和增值运维服务、各类合规审计服务等，打造安全合规、低碳绿色、智能高效的算力底座		协创数据、奥佳软件
2025 年 3 月 8 日	《协创数据：关于公司购买资产的公告》	协创数据根据经营发展需要，拟向多家供应商（以下合并简称“X”）采购服务器，并签署相关采购合同，采购合同总金额预计不超过人30 亿民币 30 亿元，公司购买服务器主要用于为客户提供算力租赁服务		协创数据

资料来源：公司公告，国信证券经济研究所整理

投资建议

AI 产业持续迭代，Agent 成为当下应用最确定性方向。在基础大模型持续迭代的背景下，AI Agent 应用逐步水到渠成。当前各互联网厂商以及创新公司持续推出 Agent 相关产品，Agent 已开始逐步进入各个场景的工作流中，成为人机协同新范式。根据 MarketsandMarkets 最新发布的《AI Agents Market Report 2025》，全球 AI Agent（含自主智能体软件与服务）市场规模预计在 2025 年达到 7.9 亿美元，并将在 2030 年增至 526 亿美元，复合年增长率约 46%。

互联网巨头和创新公司持续推出 Agent 产品，重点关注国内推出 AI Agent 产品公司。海外来看，谷歌 AI 重回市场中心，Gemini 2.5 Pro 表现优异，开源 A2A 协议，叠加 MCP 后将进一步推动 Agent 应用繁荣。国内来看，阿里 Qwen3 性价比再大幅提升，BC 两端 Agent 生态加速；字节发布多模态 Agent；创新产品 Manus、Flowith、Lovart 等均表现不俗。尤其是众多产品已经完成了商业化探索，定价和云端运营模式清晰，部分产品已经形成可观收入，未来有望进一步加快。国内重点关注在 AI 应用和 Agent 持续布局的厂商，如金山办公、合合信息、用友网络、税友股份、亚信科技、新大陆等。

互联网巨头持续加大 AI 基础设施投资，算力租赁厂商受益明显。阿里巴巴预计未来三年，将投入超过 3800 亿元，用于建设云和 AI 硬件基础设施，总额超过去十年总和。据腾讯 2024 年财报披露数据，预计 2025 年资本开支持续上行，主要满足公司 AI 相关领域需求。目前已经有众多上市公司积极在算力租赁布局，部分公司已经披露相关订单，如海南华铁、有方科技、智微智能、协创数据、润建股份等。重点关注 AI 算力租赁产业。

风险提示

宏观经济低迷影响下游各行业 IT 支出。

AI 终端产品性能、功能等表现不及预期，市场接受度有限。

产品及服务同质化，导致行业竞争加剧。

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