

综合

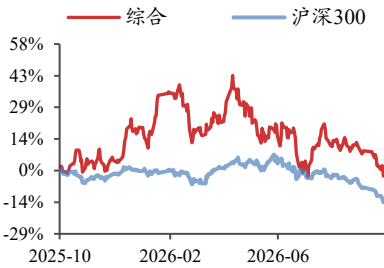
2026 年 10 月 25 日

海外科技: Agent 能力向长程专业任务延伸, SpaceX 与 Meta 业绩均超预期

投资评级: 看好 (维持)

——行业周报

行业走势图



数据来源: 聚源

相关研究报告

《海外科技: AMD 推出 Helion 机架级平台, ASML 评估进一步扩产——行业周报》-2026.10.18

《海外科技: AMD 布局世界模型, 模型能力升级与成本优化同步推进——行业周报》-2026.10.11

《海外科技: AI Agent 打开 CPU 空间, 海外 X86 服务器供应链受益——行业周报》-2026.9.27

初敏 (分析师)

chumin@kysec.cn

证书编号: S0790522080008

叶彬慧 (分析师)

yebinhui@kysec.cn

证书编号: S0790526070007

李秋莹 (联系人)

liqiuying@kysec.cn

证书编号: S0790126060038

● 大模型: Agent 能力向长程专业任务延伸, 任务成本与工具链竞争升温

Meta 发布 Muse Wonder 系列模型, 引入 Active Memory 和 Active Focus, 强化跨对话信息检索及长周期任务持续执行能力; Kimi 发布 K3.1 Pro, 扩大强化学习规模, 提升代码、3D 创作及金融研究能力, 并推出 Highspeed 高速版本; MiniMax 发布 M3.1 Flash, 线上推理速度提升至上一代的近 3 倍, Agent 类任务综合使用成本较上一代降低 70%; DeepSeek 将 V4.1 Flash 更新至 1017 版本, DeepSWE v1.1 得分由 74.2 提升至 77.0, 并针对 Agent 框架优化执行效率, 同步开源模型权重、推出自研搜索 API。我们认为, 技术层面, 长期记忆与持续执行能力的强化, 以及强化学习规模的扩大, 有助于提升模型处理多步骤任务的能力, 推动应用向编程、3D 创作及金融研究等专业 workflow 拓展。商业层面, 竞争维度从 Token 单价延伸至任务综合成本与交付效率, 推理与执行优化降低任务成本, 高速版本则提供更多速度与价格选择。生态层面, Agent 框架适配、搜索 API 及托管智能体服务逐步完善, 有助于降低开发与部署门槛, 推动模型能力接入实际业务流程。

● 云计算: SpaceX 超级智能板块首超连接业务, Meta 资本开支下限上修

SpaceX 2026Q3 首次实现单季度盈利, 超级智能 (SI) 业务收入同比增长约 830%, 占总收入比重升至 51.9%, 首次超越连接业务成为第一大收入来源, 算力部署与云服务放量推动业绩增长。Meta 营收及四季度指引均超预期, 但 AI 基础设施投入持续增加, 2026 年资本开支指引下限上调至 1,350 亿美元, 单季度自由现金流首次转负。整体来看, AI 算力商业化持续推进, 但高额资本开支对盈利能力与现金流的影响仍需关注。

● 投资建议:

大模型与云计算: 推荐标的: 智谱 AI, 德信证券; 备选标的: 谷歌、亚马逊、微软、SpaceX、Meta、MiniMax。

风险提示: 模型及产品推出不及预期、AI 算力需求不及预期风险、算力供给超预期释放风险、宏观经济增长放缓。

目 录

1、大模型 Agent 能力向长程专业任务延伸，任务成本与工具链竞争升温	4
1.1、Meta: Muse Wonder 强化长期记忆与持续任务执行能力，推进个人与专业 Agent 应用	4
1.2、Kimi: K3.1 Pro 加大强化学习投入 3D 建模及金融研究能力同步提升	4
1.3、MiniMax: M3.1 Flash 实现 Agent 与编程能力升级，推理速度提升至上一代的近 3 倍，Agent 任务综合成本较上一代下降 70%	6
1.4、DeepSeek: V4.1 Flash 强化“许愿式编程”能力，任务执行效率与成本同步优化	6
2、云计算: SpaceX 超级智能板块首超连接业务，Meta 资本开支下限上修	7
2.1、SpaceX: 首次单季度盈利，超级智能板块（SI）超越连接板块成为第一大收入来源	7
2.2、Meta: 营收与 Q4 指引双超预期，资本开支下限上调至 1,350 亿美元	9
3、港股周度更新	12
4、模型与算力价格数据更新	14
4.1、模型信息更新	14
4.2、算力租赁价格更新	16
5、投资建议	17
6、风险提示	17

图表目录

图 1: Kimi K3.1 Pro 以较低成本取得接近头部模型的 Terminal-Bench 4.0 得分	5
图 2: MiniMax M3.1 Flash 在 Automation-Bench 等基准中得分领先	6
图 3: 故宫 3D 建模展示效果	7
图 4: SI 业务营收同比增长 830%，带动 SpaceX 营收高增（单位：百万美元）	8
图 5: SI 业务营收占比已达到 51.9%，超越连接业务成为第一大收入来源	8
图 6: 经营利润率首次转为正数，SpaceX 的盈利能力逐步提升	8
图 7: SpaceX 资本开支进一步增长（单位：百万美元）	9
图 8: SI 业务的资本开支已经占到 SpaceX 总资本开支的 84%	9
图 9: 根据管理层指引，SpaceX 算力基础设施容量将快速增长（单位：GW）	9
图 10: Meta 2026Q3 营收 644 亿美元，同比增长 26%（单位：百万美元）	10
图 11: Meta 经营利润率与净利率下滑	10
图 12: Meta 自由现金流首次录得负数（单位：百万美元）	11
图 13: 本周恒生指数（2026.10.19-2026.10.23）在全球主要市场中涨跌幅靠后（单位：%）	12
图 14: 本周（2026.10.19-2026.10.23）恒生科技跌幅较大（%）	12
图 15: 本周（2026.10.19-2026.10.23）恒生航运主题、科技跌幅居前（%）	12
图 16: 2026 年 9 月底以来港股通当日买入成交净额有所波动（单位：亿元）	13
图 17: 恒生沪港通 AH 溢价指数或已触底	14
图 18: OpenRouter 的 token 调用量持续提升	15
图 19: 国产 AI 在 ZenMux 的 Token 市场份额占比接近 70%	16
图: B200 租赁价格高位运行（单位：\$/小时）	16
图: H200 租赁价格高位运行（单位：\$/小时）	16
图: H100 租赁价格高位震荡（单位：\$/小时）	17
图: A100 租赁价格有所回升（单位：\$/小时）	17

表 1: 本周（2026.10.19-2026.10.23）腾讯控股、智谱与 MINIMAX-W 流入资金居前	13
---	----

表 2: 9 月以来新上线模型	14
-----------------------	----

1、大模型 Agent 能力向长程专业任务延伸，任务成本与工具链竞争升温

1.1、Meta: Muse Wonder 强化长期记忆与持续任务执行能力，推进个人与专业 Agent 应用

Muse Wonder 系列在个人及专业 Agent 任务中实现能力升级，覆盖知识工作、业务流程和复杂编程场景。10月12日，Meta发布 Muse Wonder 及 Muse Wonder Full。根据 Meta 官方评测，Muse Wonder 系列在 Personal Superintelligence Index、AutomationBench-AA、MemoryBench、SWE-Marathon 及 DeepSWE 等个人及专业 Agent、长期记忆与复杂编程基准中取得领先表现。

Active Memory 强化跨对话上下文管理，Active Focus 提升复杂任务的持续执行能力。Meta 将 Active Memory 定义为跨对话保留重要信息，并在相关任务中主动检索历史上下文的能力；Active Focus 则强调模型面对复杂目标时持续推进工作，通过尝试不同方法、处理阻碍及适应新信息完成任务，同时在必要时寻求澄清或重要操作确认。

专业领域知识与 Agent 编程能力同步提升，覆盖金融、法律、健康及复杂软件工程任务。Muse Wonder 系列能够将专业知识应用于实际问题，覆盖租赁合同理解、财务及健康信息解读等场景。Meta 采用 Harvey LAB-AA 和 Vals Finance Agent V2 等基准评测模型的专业工作能力。编程方面，Meta 通过 DeepSWE 及 SWE-Marathon 评估复杂软件工程表现。

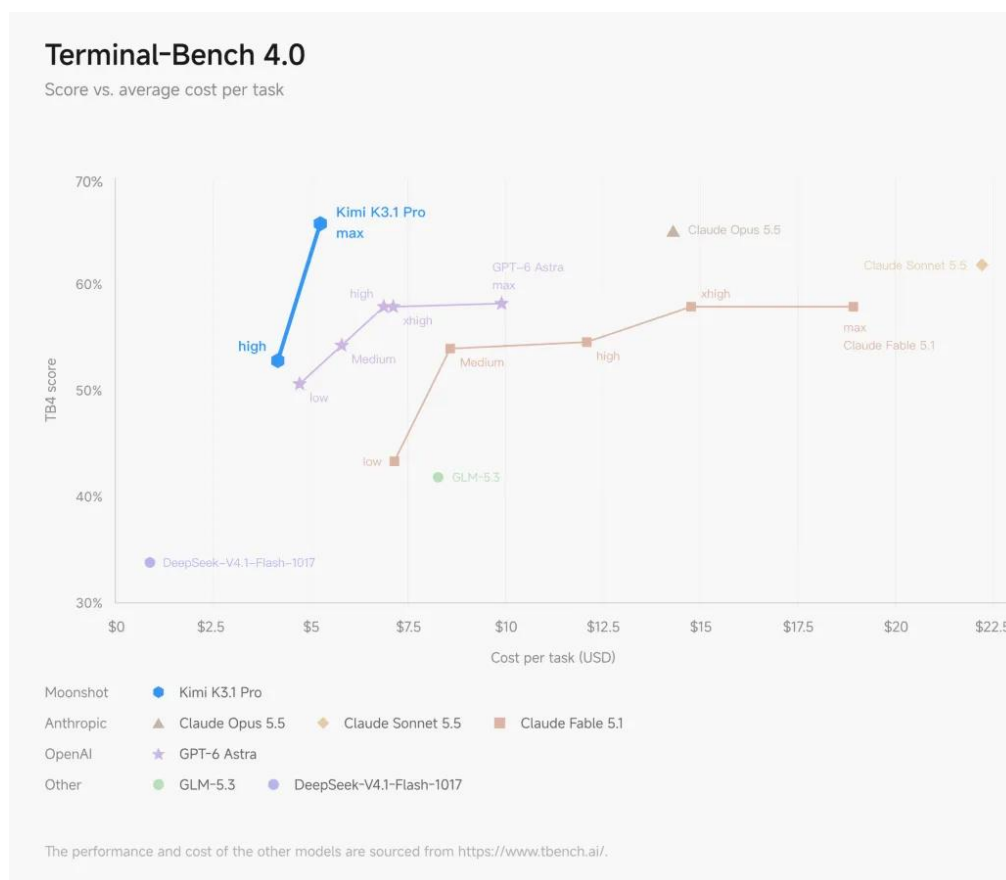
更强基础模型推动能力向后训练阶段未针对性训练的任务泛化，提升复杂场景的适应能力。结合其在后训练目标任务上的得分，Muse Wonder Full 在非目标任务上的表现高于 Muse Spark 1.3 趋势所预测的水平。Meta 将这一表现归因于更大、更强的基础模型，使后训练阶段获得的能力提升能够更有效地迁移至非目标任务。

产品部署覆盖开发者 API 及编程工具。Muse Wonder 已通过 Meta Model API 和 Muse Code 开放，Muse Wonder Full 面向 Muse Code 订阅用户提供。安全方面，Meta 重点强化隐私信息保护、抵御恶意指令等能力。

1.2、Kimi: K3.1 Pro 加大强化学习投入，代码、3D 创作及金融研究能力同步提升

Kimi 发布 K3.1 Pro，扩大强化学习规模并提升复杂 Agent 任务能力。Kimi K3.1 Pro 于 10月24日发布，是 K3.1 系列首款模型，也是 Kimi 目前能力最强的 3 万亿级参数模型，支持 100 万 Token 上下文。新模型基于 K3 进行后训练，并对 RL 框架、长上下文推理、训练效率、沙箱及调度系统进行全栈优化。官方披露，在 Max 推理强度下，K3.1 Pro 在 Terminal-Bench 4.0 中取得 65.7% 的成绩，高于 GPT-6 Astra 的 58.2%；内部 Kimi Code Bench 3.0 Long-horizon 得分为 61.2%，高于 K3 的 55.95%。

图1: Kimi K3.1 Pro 以较低成本取得接近头部模型的 Terminal-Bench 4.0 得分



资料来源：月之暗面 Kimi 微信公众号

复杂编程与 3D 创作能力提升，支持多工具协作及端到端项目开发。K3.1 Pro 能够调用 Unity、Unreal Engine、Godot 等游戏引擎，并结合网络搜索完成游戏、网页及交互场景开发。3D 建模方面，模型在 BenchCAD 基准中借助工具取得 92.7% 的成绩，接近 Claude Opus 5.5 的 96.2% 和 GPT-6 Astra 的 95.9%。模型支持根据文字、草图及参考图重建 3D 模型，无需 MCP 连接器即可直接编写和运行 Blender、CAD 等专业工具脚本。Kimi 网页端与 App 亦同步上线 3D Make 功能，支持 3D 模型在线预览、编辑及导出。

金融研究与财务建模能力增强，覆盖跨文档分析、投资研究及完整建模 workflow。K3.1 Pro 在 OfficeQA Pro v2 基准中取得 64.4 分，该基准主要考察金融文档检索、跨文档核对及多步骤数值推理。根据官方内部评测，相比 K3，新模型在 Financial Modeling 和 Financial Research 基准中的表现分别提升 14% 和 20%，Business Operation 得分约为前代的 4.2 倍。事实准确性方面，在聚焦高难度易错场景的内部 Kimi Hallu Bench Hard 压力测试中，单项事实陈述错误率由 23.4% 降至 8.3%，该压力测试结果不代表日常使用中的整体错误率。

推出 Highspeed 高速版本，API 价格及模型开放计划同步公布。K3.1 Pro 已上线 Kimi 网页端、App、Kimi Work、Kimi Code 及 API，完整模型权重计划于 11 月 4 日前发布。标准版每百万 Token 未命中缓存的输入价格为 13 元，命中缓存的输入价格为 1.3 元，输出价格为 68 元；Highspeed 版本在同等智能水平下，输出速度约为标准版的 2 至 3 倍，未命中缓存的输入价格和输出价格均为标准版的 2 倍，命中缓存的输入价格与标准版相同。此外，Kimi 提供 Hosted Agents 托管智能体服务，该服务

提供 7×24 小时的云端运行环境，支持长程任务执行及多智能体协作。

1.3、MiniMax: M3.1 Flash 实现 Agent 与编程能力升级，推理速度提升至上一代的近 3 倍，Agent 任务综合成本较上一代下降 70%

MiniMax 发布 M3.1 Flash，复杂 Agent、编程及计算机操作能力进一步提升。10 月 19 日，MiniMax 正式发布 M3.1 Flash 模型。此前以匿名形式运行的 Space Bunny 为该模型的早期版本，累计处理超过 200T Token。模型在 AutomationBench-AA 评测中位居前列，可处理跨 SaaS 系统的销售、营销、运营及财务等企业工作流；在 BankerToolBench 中获得 76.2 分，展现出多步骤规划、跨文档一致性、金融合规及链式工具调用能力。

图2：MiniMax M3.1 Flash 在 Automation-Bench 等基准中得分领先

		MiniMax M3.1 Flash	DeepSeek -V4.1- Flash-1017	DeepSeek- V4.1-Flash	Kimi-K3	GLM-5.3	GPT-5.6 Sol	Opus 5
CoWork	Automation-Bench v1.0.6	62.8	60.5	54.8	46.7	48.8	45.8	50.3
	BankerToolBench	76.2	-	73.7	69.1	74.5	74.6	76.2
	GDP.pdf AA	20.8	-	12.8	22.0	11.2	27.2	21.6
	Terminal Bench 4.0	38.1	42.3	31.2	12.6	37.9	39.9	51.8
	LOCA-Bench 256K	60.3	-	-	49.7	50.3	60.0	53.7
Computer Use	Agents' Last Exam	29.6	-	-	28.3	-	30.6	32.2
	OSWorld 2.0	61.7	-	54.3	58.3	-	65.7	70.6
Coding	PostTrain Bench v1.1	39.3	-	30.5	36.6	37.5	36.2	40.4
	DeepSWE v1.1	73.7	77.0	74.2	67.5	66.9	73.0	74.0
	NL2Repo-Bench	61.2	68.2	65.4	58.0	58.0	56.8	75.3
	SWE Refactor Bench	23.0	-	11.8	19.5	10.4	28.5	31.0
网络安全	CyberGym	82.2	-	-	80.0	84.5	83.6	-
通用能力	BabyVision (w/tools)	91.0	92.5	89.6	85.7	-	88.9	94.1
	HLE (w/o tools)	43.4	44.0	39.2	43.5	42.3	49.5	55.6
	CharXiv RQ	82.5	-	69.0	84.8	-	85.2	83.7

资料来源：MiniMax 官网

复杂编程及计算机操作能力提升，覆盖多类长链路任务。M3.1 Flash 在 DeepSWE v1.1 软件工程评测中取得 73.7% 的成绩，该评测重点考察代码理解、功能开发、问题修复及测试验证等复杂任务。计算机操作方面，模型在 OSWorld V2 和 Agents' Last Exam (ALE) 中分别取得 61.7 分和 29.6 分。

推理效率提升，线上速度提升至上一代的近 3 倍，Agent 任务成本下降 70%。模型通过优化推理效率及采用更激进的稀疏注意力架构，日常使用场景下输出速度可达 200+TPS，中位数稳定达到 170+TPS，线上推理速度提升至上一代的近 3 倍。Agent 类任务综合使用成本较上一代降低 70%。定价方面，M3.1 Flash 每百万 Token 输入、输出及缓存读取价格分别为 1 元、4 元和 0.1 元。

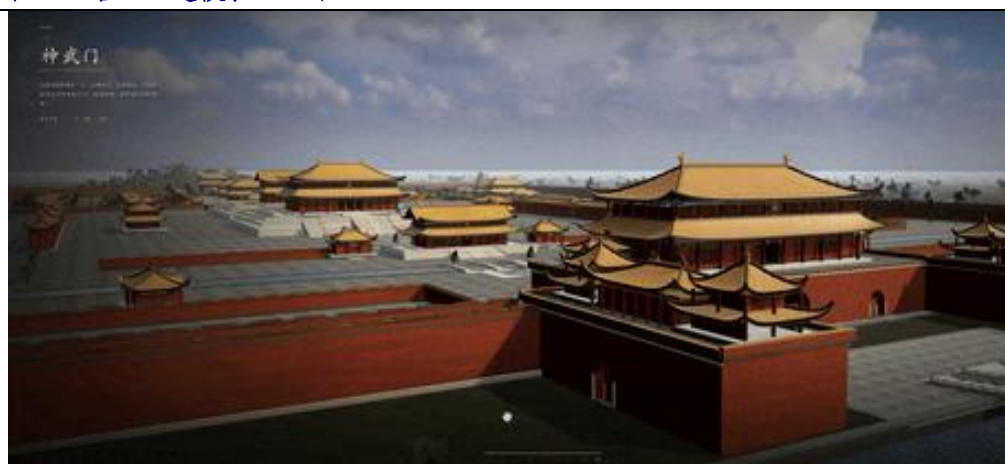
模型已覆盖 API 及编程产品，后续计划推出 Pro 版本。M3.1 Flash 现已上线 MiniMax API、M Plan 及 MiniMax Code。MiniMax 表示，将继续推出侧重更高智能能力的 Pro 版本，进一步拓展模型能力上限。

1.4、DeepSeek: V4.1 Flash 强化“许愿式编程”能力，任务执行效率与成本同步优化

DeepSeek V4.1 Flash 推出 1017 版本，编程能力进一步提升。10 月 17 日，DeepSeek 将 V4.1 Flash 更新至 V4.1-Flash-1017 版本，官方网页端、App 及 API 已同步升级。DeepSWE v1.1 得分由 74.2 提升至 77.0，反映模型在复杂软件工程任务中的执行能力进一步增强，有望提升代码开发、问题修复及自动化编程等场景的应用表现。

自反馈强化学习提升“许愿式编程”能力，支持从零构建完整项目。DeepSeek V4.1-Flash-1017 通过基于自反馈的强化学习训练，提升模型在游戏开发、3D 建模等场景中从零构建完整项目的能力。官方展示 3D 建模、网页版赛车游戏、《传送门 2》游戏复刻及 3D 动态场景生成等案例。

图3: 故宫 3D 建模展示效果



资料来源：DeepSeek 官网

针对 Agent 框架优化模型执行表现，兼顾任务质量、运行时间及 Token 消耗。本次更新重点针对 DeepSeek Harness、腾讯 CodeBuddy 及 WorkBuddy 等 Agent 框架进行优化，通过缩短任务运行时间与输出长度，改善不同框架下的性能与费用表现。

模型权重同步开源，自研搜索 API 正式开放。DeepSeek 已在 Hugging Face 及 ModelScope 开放 V4.1-Flash-1017 模型权重，同时面向公众推出自研 Search 与 Fetch API，Search 与 Fetch API 均按成功请求次数计费，价格为每千次成功请求 5 元。在 DeepSeek Harness、Claude Code 及 Codex 中使用 DeepSeek API 时，经服务端工具调用产生的 Search/Fetch 请求暂不计费，自北京时间 2026 年 11 月 1 日 00:00 起按次收费。

2、云计算：SpaceX 超级智能板块首超连接业务，Meta 资本开支下限上修

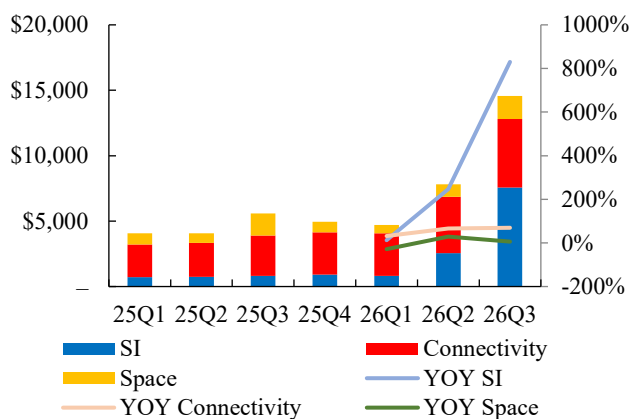
本周 SpaceX 与 Meta 相继披露 2026Q3 业绩，营收与净利润均实现较大增长，反映出两家公司的算力投入与模型变现同时进入验证期，且资本开支均继续增长。

2.1、SpaceX：首次单季度盈利，超级智能板块（SI）超越连接板块成为第一大收入来源

SpaceX 2026Q3 营收高增，首次实现单季度盈利，超级智能（SI）是本轮增长

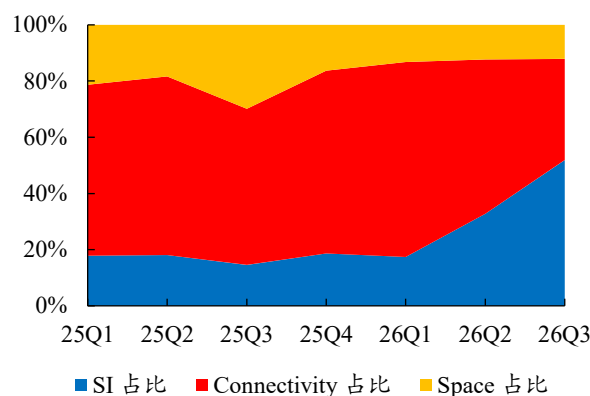
的核心驱动力。公司实现收入 145.75 亿美元，同比增长 161%，其中航天业务收入 17.67 亿美元，同比增长 6%；连接业务收入 52.48 亿美元，同比增长 69%；SI 收入 75.60 亿美元，同比增长约 830%，SI 收入占比已升至 51.9%，超过连接业务成为第一大收入来源。公司实现净利润 44 亿美元，2025 年同期为净亏损 11 亿美元，实现单季度首次盈利；调整后 EBITDA 达 88.09 亿美元，同比增长 306%。

图4: SI 业务营收同比增长 830%，带动 SpaceX 营收高增（单位：百万美元）



数据来源：SpaceX 财报、开源证券研究所

图5: SI 业务营收占比以达到 51.9%，超越连接业务成为第一大收入来源



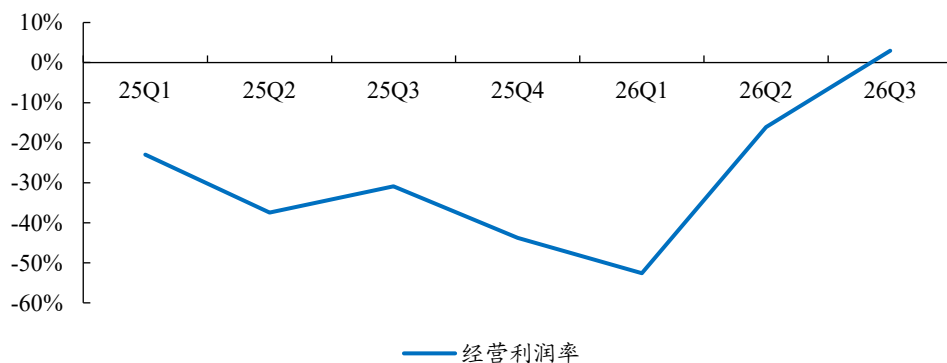
数据来源：SpaceX 财报、开源证券研究所

重点关注 SI 板块营收与利润状况。2026Q3 公司 SI 业务收入 75.60 亿美元，同比增长约 830%，主要系新部署算力变现，云服务合同放量，以及 Cursor 并表贡献（6.08 亿美元来自 Cursor）。SI 业务经营利润为 4.31 亿美元，经营利润率为 3%，首次转为正数，反映出公司的盈利能力逐步提升，调整后 EBITDA 为 49.53 亿美元。SI 业务增长。

（1）云计算：需求高景气，在当前价格水平下，新算力部署的投资回收期不到一年，算力定价目前略高于每瓦 30 至 50 美元区间上限。

（2）模型及应用：公司在加快 Grok 模型的发布节奏，近期预计还将实现性能的显著提升。Grok bot 目前每周付费用户已达 80 万，较上个月增长一倍以上，而 B2B 是增长最快的类别，越来越多的公司将 Grok bot 用于销售、营销、IT 和工程等工作。

图6: 经营利润率首次转为正数，SpaceX 的盈利能力逐步提升

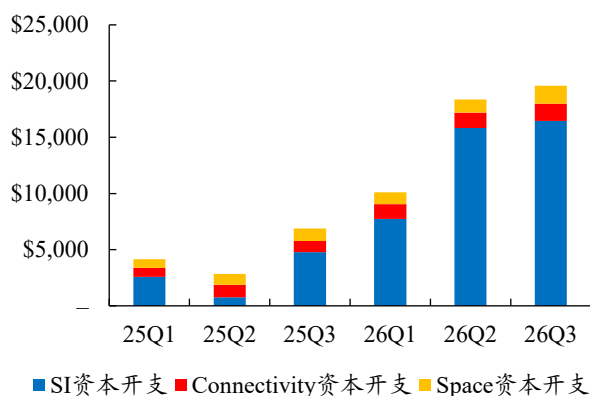


数据来源：SpaceX 财报、开源证券研究所

与此同时，公司资本开支进入加速投入期，并拥有大量在手订单。2026Q3 公司

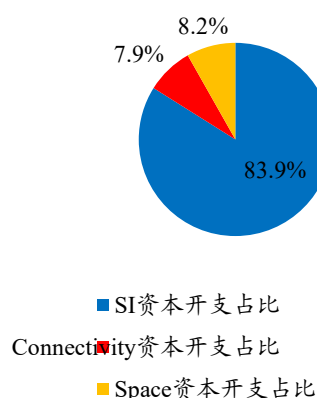
资本开支达 195.93 亿美元，其中 SI 业务当季资本开支高达 164.44 亿美元，占公司总资本开支约 84%，前三季度累计资本开支 480.69 亿美元，公司预计 Q4 资本开支进一步升至 350 亿—400 亿美元。而截至 9 月末，公司在手订单达 630 亿美元，其中约 420 亿美元预计未来一年确认收入，这反映算力基础设施正在快速转化为收入增长。

图7: SpaceX 资本开支进一步增长 (单位: 百万美元)



数据来源: SpaceX 财报、开源证券研究所

图8: SI 业务的资本开支已经占到 SpaceX 总资本开支的 84%



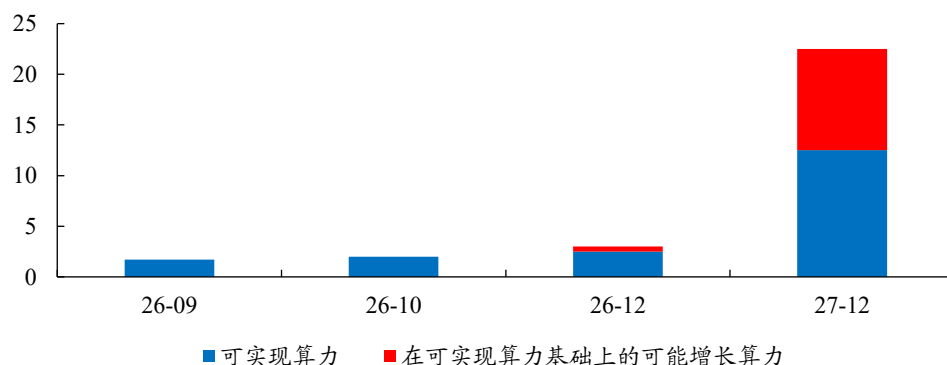
数据来源: SpaceX 2026Q3 财报、开源证券研究所

在未来指引方面，公司管理层对未来发展前景乐观，给出了较为激进的业绩与算力基础设施建设指引：

(1) 收入指引：目标 2026 年 12 月 ARR 1000 亿美元，2027Q1 收入指引 330 亿—370 亿美元，2027 年底 ARR 愿景为 5000 亿至 1 万亿美元。

(2) 算力基础设施：2026Q3 算力容量 1.7GW，预计 10 月底接近 2GW，年末将达到 2.5-3GW，2027 年新增约 10GW，小概率达到新增 20GW。

图9: 根据管理层指引，SpaceX 算力基础设施容量将快速增长 (单位: GW)



数据来源: SpaceX 2026Q3 业绩说明会、开源证券研究所

2.2、Meta: 营收与 Q4 指引双超预期，资本开支下限上调至 1,350 亿美元

收入端：公司 2026Q3 营业收入 646.0 亿美元，同比+26%，其中广告收入 630.6

亿美元、同比+26%，展示量+10%×平均单价+14%，量价齐升。Q4 收入指引 730-760 亿美元、中值 745 亿美元。

利润端：受一次性项目扭曲，GAAP 摊薄 EPS 3.83 美元，其中包含 98.9 亿美元法律和和解费用，剔除后 EPS 为 7.40 美元；GAAP 净利润 98.60 亿美元、同比+264%。总成本费用同比+568%至美元。营业利润 129.9 亿美元、同比-37%，营业利润率由 2025 年同期 40%降至 20%；剔除法律费用后营业利润 228.8 亿美元、利润率为 35.4%。我们认为，剔除一次性项目后仍有约 4.6pct 的利润率下滑，对应 AI 基础设施折旧与运营成本的侵蚀。

模型与应用端：Muse 是 Meta 首个前沿模型与个人 AI 智能体的结合体，上线约六周，美国与加拿大已有超过 1000 万人设置账号。交易场景已初步验证：超过四分之一的周活跃用户使用购物功能，平均购买金额超过 200 美元，相关交易包含旅行消费。我们认为，Muse 的定价逻辑不同于按 token 计费的 API 模式，而是把模型能力嵌入交易链路、按成交抽佣，使收入与 GMV 而非调用量挂钩，单位经济上限更高。约束同样明确：收入确认滞后于算力投入，且抽佣比例受商家议价能力约束，管理层坦言 Muse 计算密集度高，效率与容量是扩大覆盖范围的主要限制。

图10: Meta 2026Q3 营收 644 亿美元,同比增长 26%(单位: 百万美元)

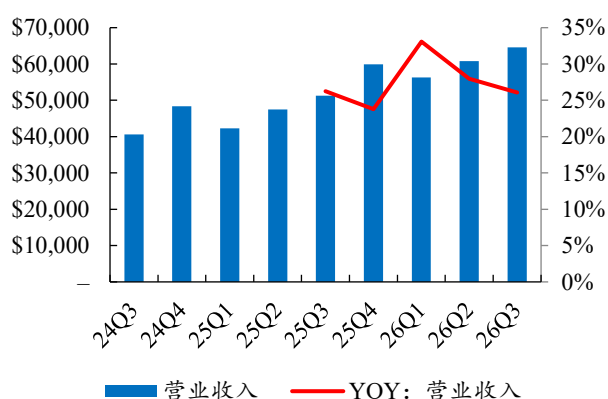
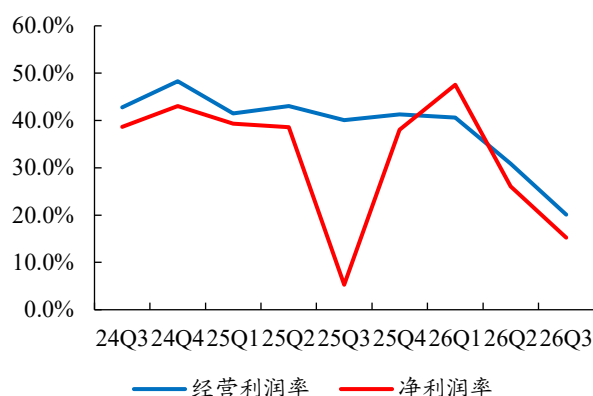


图11: Meta 经营利润率与净利率下滑

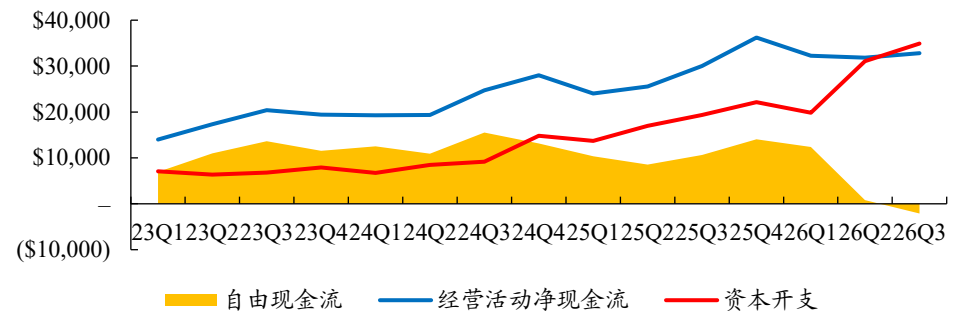


数据来源: Wind、开源证券研究所

数据来源: Wind、开源证券研究所 (注: 若剔除一次性法律费用, 2026Q3 营业利润率为 35.4%)

自由现金流与未来资本开支指引是 Meta 本季度最值得警惕的指标: 2026Q3 公司资本开支 (含融资租赁本金) 349.1 亿美元, 超过 327.9 亿美元的经营活动现金流, 导致自由现金流为 -21.2 亿美元, 首次转负。公司同时将全年资本开支下限由 1300 亿美元上调至 1350 亿美元、上限维持 1450 亿美元。

图12: Meta 自由现金流首次录得负数 (单位: 百万美元)

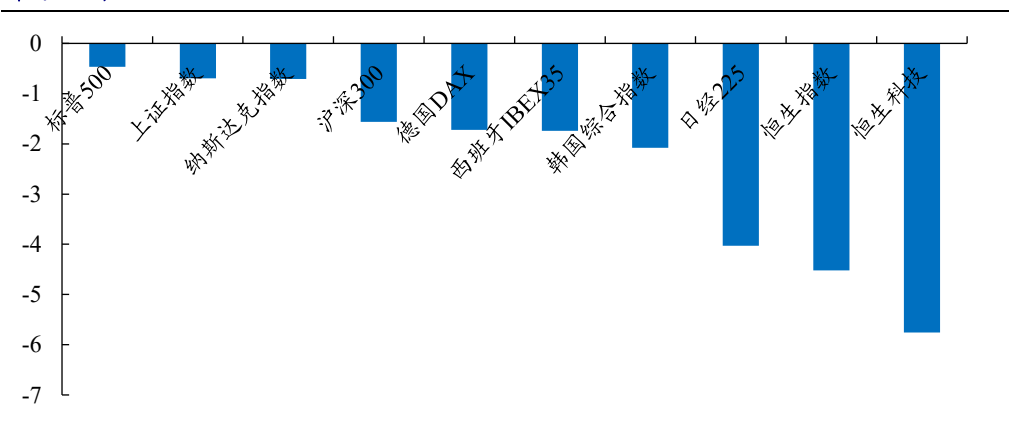


数据来源: Meta 财报、开源证券研究所

3、港股周度更新

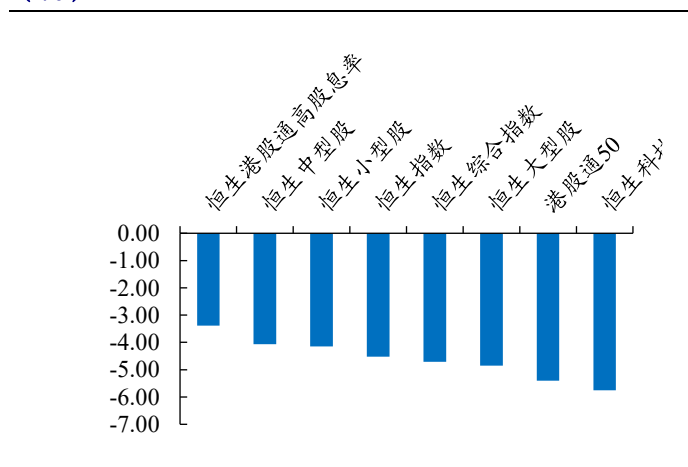
本周恒生指数下跌 4.52%，在全球主要市场中涨跌幅靠后，从成分上看，恒生航运主题、传媒跌幅居前，恒生科技本周下跌 5.76%。

图13：本周恒生指数（2026.10.19-2026.10.23）在全球主要市场中涨跌幅靠后（单位：%）



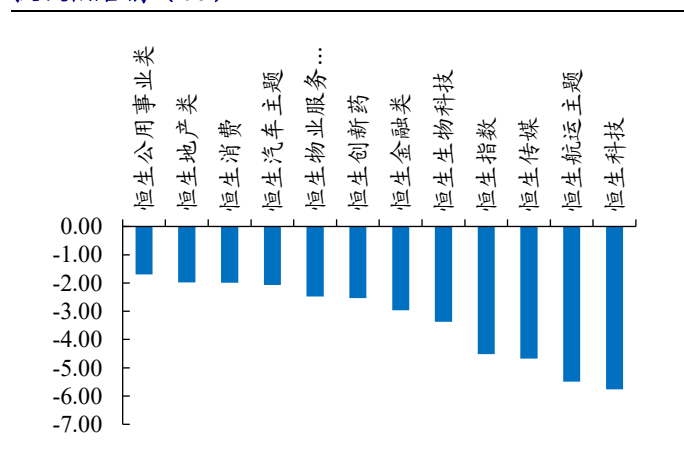
数据来源：Wind、开源证券研究所

图14：本周（2026.10.19-2026.10.23）恒生科技跌幅较大（%）



数据来源：Wind、开源证券研究所

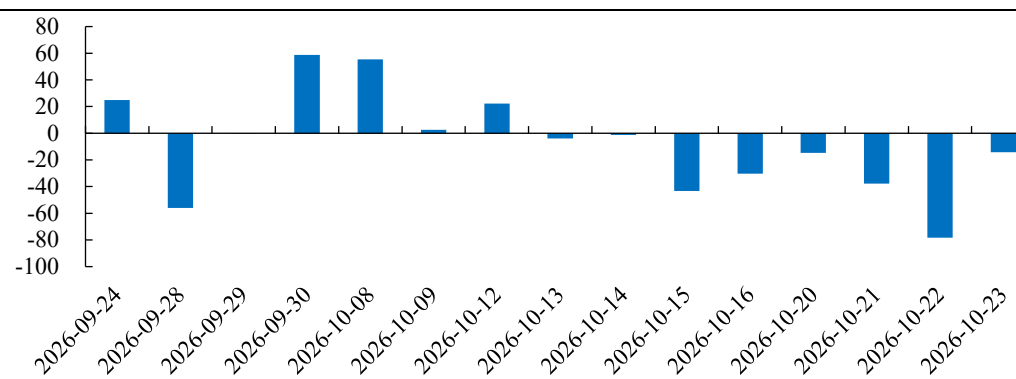
图15：本周（2026.10.19-2026.10.23）恒生航运主题、科技跌幅居前（%）



数据来源：Wind、开源证券研究所

本周港股通成交净额为流出 144.97 亿元，环比下降，从头部活跃个股交易层面来看，腾讯控股（+12.3 亿港元）、智谱（+12.0 亿港元）、MINIMAX-W（+4.9 亿港元）资金净流入居前。

图16: 2026 年 9 月底以来港股通当日买入成交净额有所波动（单位：亿元）



数据来源：Wind、开源证券研究所

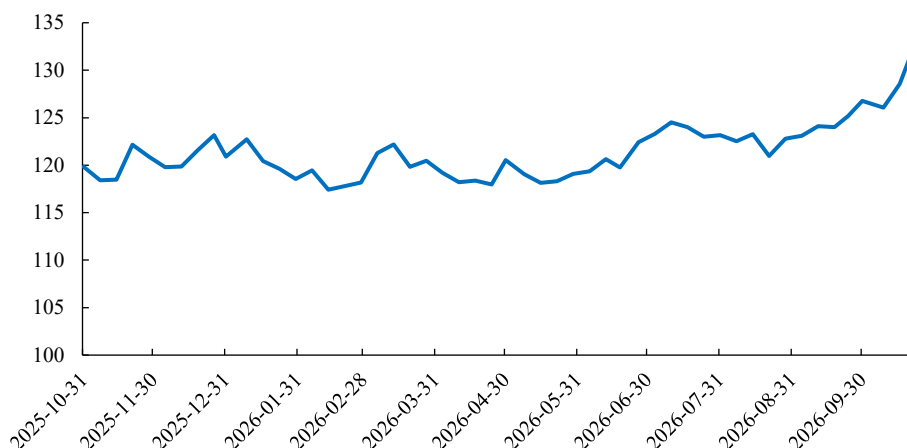
表1: 本周（2026.10.19-2026.10.23）腾讯控股、智谱与 MINIMAX-W 流入资金居前

股票代码	股票名称	成交净买入（亿港元）	港股通持股数占总股本比例
0700.HK	腾讯控股	12.3	12.0%
2513.HK	智谱	12.0	13.8%
0100.HK	MINIMAX-W	4.9	11.2%
1196.HK	伟禄科技股份	2.5	11.7%
3308.HK	中际旭创	1.9	9.2%
6869.HK	长飞光纤光缆	0.7	44.4%
9992.HK	泡泡玛特	0.4	23.7%
6651.HK	五一视界	0.2	5.6%
3690.HK	美团-W	-0.9	18.8%
1888.HK	建滔积层板	-1.3	11.4%
1548.HK	金斯瑞生物科技	-2.0	20.5%
1347.HK	华虹宏力	-3.7	14.7%
0883.HK	中国海洋石油	-5.8	22.5%
2269.HK	药明生物	-7.2	24.0%
0992.HK	联想集团	-7.3	7.1%
0941.HK	中国移动	-10.3	13.2%
1810.HK	小米集团-W	-12.3	19.5%
0981.HK	中芯国际	-29.9	27.1%

数据来源：Wind、开源证券研究所

本周恒生沪港通 AH 溢价指数为 133.03，环比有所上升，AH 股溢价触底回弹，随着后续更多 A 股科技类标的赴港上市，或带动港股板块的“成长性溢价”持续向 A 股靠拢。

图17: 恒生沪港通 AH 溢价指数或已触底



数据来源: Wind、开源证券研究所

4、模型与算力价格数据更新

4.1、模型信息更新

9 月以来全球大模型行业进入新品密集发布期,海内外头部厂商先后推出十余款新模型,产品定位与定价梯队分化明显。本周(2026.10.19-2026.10.23)新上线模型均为开源模型,包括 MiniMax M3.1 Flash、Kimi K3.1 Pro 与 Kimi K3.1 Pro Highspeed。

表2: 9 月以来新上线模型

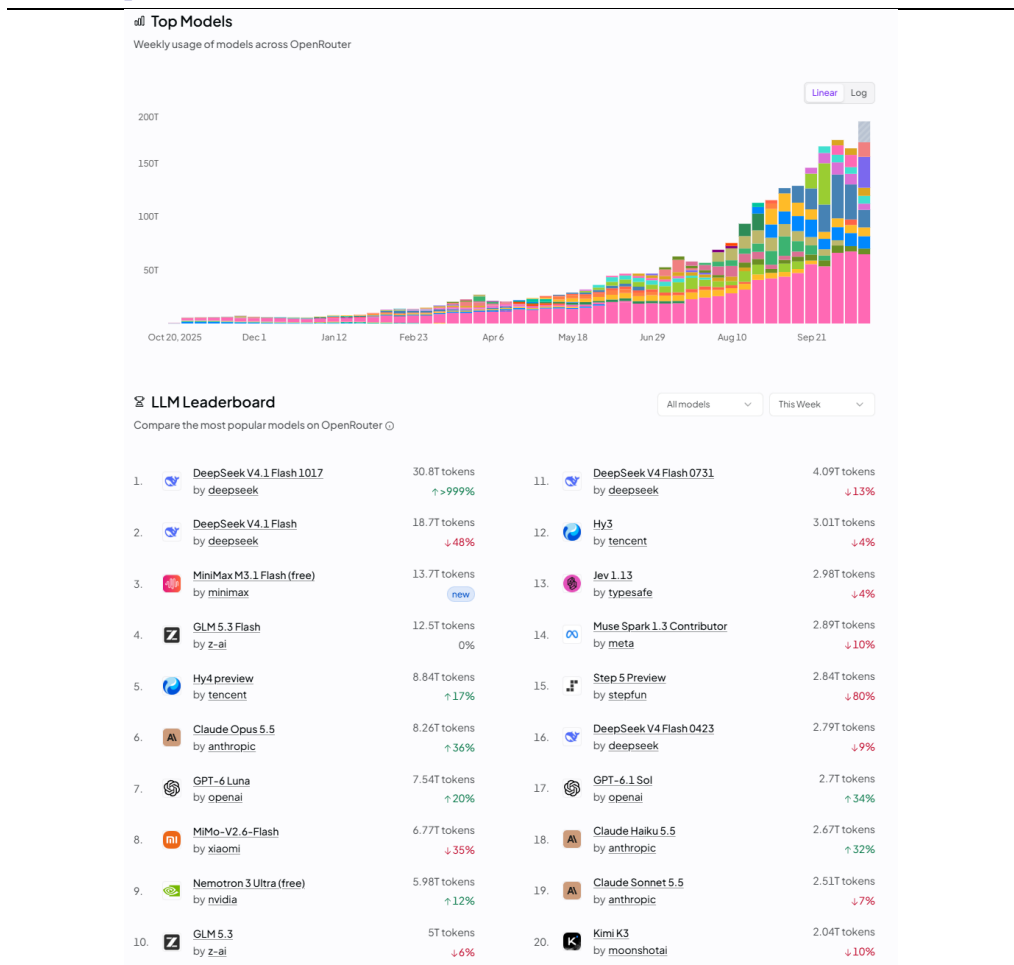
已发布模型	发布时间	厂商	输入价格 (/M)	缓存命中输入价格 (/M)	输出价格 (/M)	上下文窗口
Claude Fable 5.1	2026-09-01	Anthropic	\$10.00	\$0.25	\$50.00	1M
Claude Mythos 5.1	2026-09-01	Anthropic	\$10.00	\$0.25	\$50.00	1M
Gemini 3.8 Flash	2026-09-02	Google	\$0.75	\$0.08	\$3.75	1M
Muse Spark 1.3	2026-09-02	Meta	\$1.25	\$0.15	\$4.25	1M
GPT-6 Astra	2026-09-03	OpenAI	\$10.00	\$1.00	\$50.00	1.05M
Kimi K2.8 Preview	2026-09-11	月之暗面	目前只在 Kimi Code 与 Kimi Work 上线			1M
GLM-5.3-FlashX	2026-09-18	智谱	¥2.00	¥0.57	¥7.00	1M
Step 5 Preview	2026-09-20	阶跃星辰	¥7.00	¥0.35	¥20.00	1M
MiMo-V2.6-Pro	2026/9/22	小米	¥3.00	¥0.03	¥6.00	1M
MiMo-V2.6-Flash	2026/9/22	小米	¥1.00	¥0.02	¥2.00	1M
Grok 4.7	2026/9/22	SpaceXAI	\$2.00	\$0.50	\$6.00	500K
GPT-6 Sol	2026/9/23	OpenAI	\$2.00	\$0.20	\$10.00	1.05M
GPT-6 Luna	2026/9/23	OpenAI	\$0.10	\$0.01	\$0.50	1.05M
Claude Opus 5.5	2026/9/23	Anthropic	\$4.00	\$0.20	\$20.00	1M
Claude Sonnet 5.5	2026/9/28	Anthropic	\$2.00	\$0.20	\$10.00	1M
GPT-6.1 Sol	2026/9/29	OpenAI	\$2.00	\$0.10	\$10.00	1.05M
Gemini 4 Argon	2026/10/01	Google	\$2.00	\$0.10	\$10.00	1M
Beam	2026/10/05	Reflection AI		暂无		1M
Claude Haiku 5.5	2026/10/8	Anthropic	\$0.10	\$0.01	\$0.50	1M

已发布模型	发布时间	厂商	输入价格 (/M)	缓存命中输入价格 (/M)	输出价格 (/M)	上下文窗口
Muse Wonder Full	2026/10/12	Meta	目前只在 Muse Code 订阅中开放			1M
Muse Wonder	2026/10/12	Meta	\$4.00	\$0.20	\$20.00	1M
Muse Spark Open	2026/10/15	Meta	\$1.25	\$0.15	\$4.25	1M
DeepSeek-V4.1-Flash-1017(高峰)	2026/10/17	DeepSeek	\$0.30	\$0.01	\$1.20	1M
DeepSeek-V4.1-Flash-1017(空闲)	2026/10/17	DeepSeek	\$0.15	\$0.00	\$0.60	1M
MiniMax M3.1 Flash	0/19	MiniMax	\$0.15	\$0.02	\$0.60	1M
Kimi K3.1 Pro Highspeed	2026/10/24	月之暗面	\$4.00	\$0.20	\$20.00	1M
Kimi K3.1 Pro	2026/10/24	月之暗面	\$2.00	\$0.20	\$10.00	1M

资料来源：各模型厂商官网、IT之家、开源证券研究所

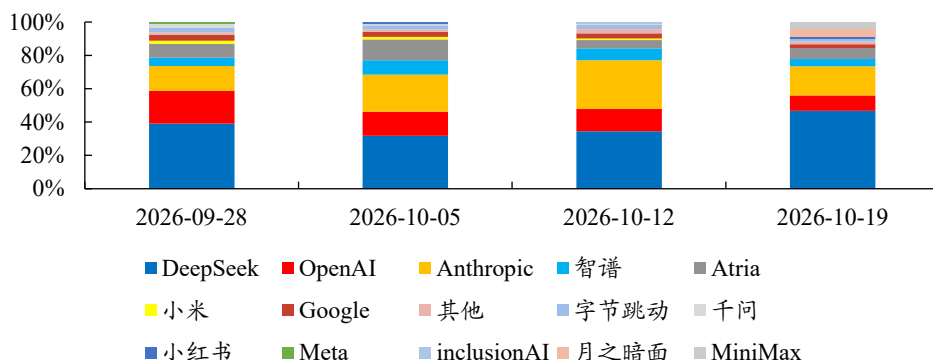
本周（2026.10.19-2026.10.23）OpenRouter 的 token 调用量持续提升，其中国产模型 DeepSeek V4.1 Flash 1017、MiniMax M3 Flash 与 GLM 5.3 Flash 位居前列。截至 10 月 19 日，在 ZenMux 统计的 Token 市场份额中，国产 AI 占比接近 70%，DeepSeek 市场份额保持第一，占比为 46.73%。依托开源生态及出色的性价比优势，国产模型持续扩大市场影响力。

图18: OpenRouter 的 token 调用量持续提升



资料来源：OpenRouter 官网

图19: 国产 AI 在 ZenMux 的 Token 市场份额占比接近 70%



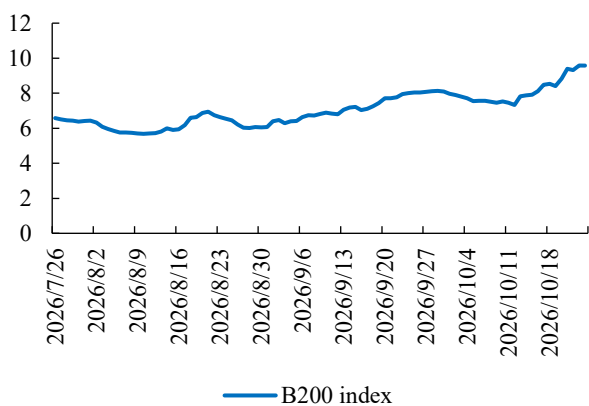
数据来源: ZenMux、开源证券研究所

4.2、算力租赁价格更新

GPU 租赁价格整体高位运行，不同性能 GPU 价格走势有所分化。根据 Ornn 统计的基于 neocloud 市场的租赁交易价格，GPU 租赁价格整体保持高位运行，不过不同性能 GPU 价格走势有所分化，高性能型号 B200 与 H200 租赁价格走势坚挺，而 H100 与 A100 租赁价格走势偏弱。

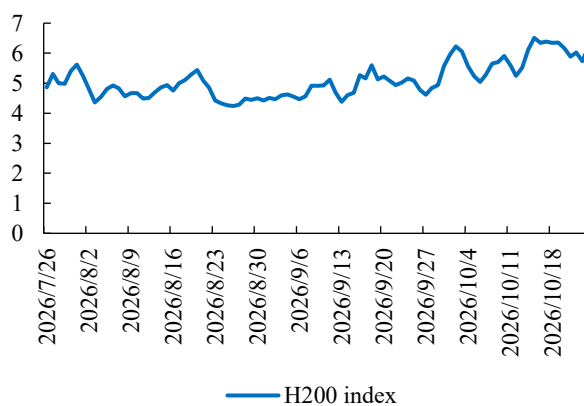
截至 10 月 24 日，B200 租赁价格为 9.57 \$/小时；H200 租赁价格为 6.20\$/小时；H100 租赁价格为 3.02\$/小时；A100 租赁价格为 1.08\$/小时。

图: B200 租赁价格高位运行（单位: \$/小时）



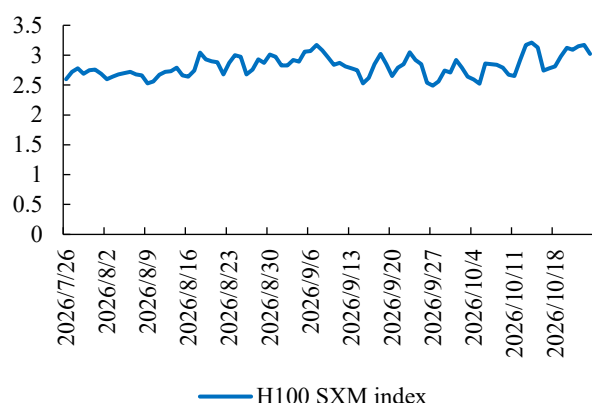
数据来源: Ornn、开源证券研究所

图: H200 租赁价格高位运行（单位: \$/小时）



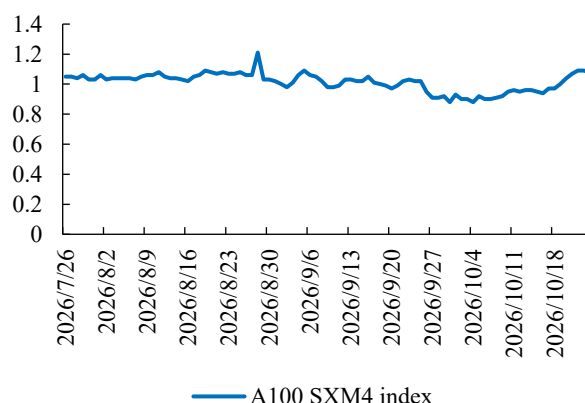
数据来源: Ornn、开源证券研究所

图：H100 租赁价格高位震荡（单位：\$/小时）



数据来源：Ornn、开源证券研究所

图：A100 租赁价格有所回升（单位：\$/小时）



数据来源：Ornn、开源证券研究所

5、投资建议

大模型：Haiku 5.5 与 GPT-6.1 Sol 提升知识工作、编程及 Agent 能力并降低实际调用成本，改善复杂工作流的应用经济性，有望带动 Token 调用和推理算力需求增长。大模型端，国产模型流量增长与商业化推进相互促进，建议关注具备模型迭代、产品落地和客户拓展能力的厂商，推荐标的：智谱、德适-B，受益标的 MiniMax。

云计算：推理需求和企业 AI 应用支撑算力建设，但高额资本开支的回报仍需跟踪，收入兑现与利润率压力可能带来短期估值波动。受益标的：谷歌、亚马逊、微软、Nebius、CoreWeave。

6、风险提示

模型及产品推出不及预期：大模型行业更新较快，市场竞争激烈，若产品推出不及时或不及预期，或影响企业未来竞争力和创收情况。

AI 算力需求不及预期风险：若下游模型迭代放缓、客户融资或算力利用率低于预期，可能导致新增订单放缓、RPO 转化周期拉长或数据中心产能交付延后。

算力供给超预期释放风险：本轮按需算力价格上涨建立在供需缺口之上，若 GPU 与内存供给快速释放，或新云厂商产能集中投放，按需与长约价格可能回落，算力租赁的盈利弹性随之收窄。

宏观经济增长放缓：若宏观经济增长放缓影响终端消费需求、企业 IT 支出等，将影响产业链相关企业的创收能力。

特别声明

《证券期货投资者适当性管理办法》、《证券经营机构投资者适当性管理实施指引（试行）》已于2017年7月1日起正式实施。根据上述规定，开源证券评定此研报的风险等级为R4（中高风险），因此通过公共平台推送的研报其适用的投资者类别仅限定为境内专业投资者及风险承受能力为C4、C5的普通投资者。若您并非境内专业投资者及风险承受能力为C4、C5的普通投资者，请取消阅读，请勿收藏、接收或使用本研报中的任何信息。

因此受限于访问权限的设置，若给您造成不便，烦请见谅！感谢您给予的理解与配合。

分析师承诺

本研究报告的署名人员具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，以勤勉的职业态度，独立、客观地出具本报告，并对内容和观点负责。本报告清晰准确地反映了署名人员的研究观点，所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。本报告采用的各种估值方法及模型均有其局限性，估值结果不保证所涉及证券能够在该价格交易。本报告署名人员保证他们报酬的任何一部分不曾与，不与，也将不会与本报告中具体的推荐意见或观点有直接或间接的联系。

股票投资评级说明

	评级	说明
证券评级	买入（Buy）	预计相对强于市场表现 20%以上；
	增持（outperform）	预计相对强于市场表现 5%~20%；
	中性（Neutral）	预计相对市场表现在 -5%~+5%之间波动；
	减持（underperform）	预计相对弱于市场表现 5%以下。
行业评级	看好（overweight）	预计行业超越整体市场表现；
	中性（Neutral）	预计行业与整体市场表现基本持平；
	看淡（underperform）	预计行业弱于整体市场表现。

备注：评级标准为以报告日后的 6~12 个月内，证券相对于市场基准指数的涨跌幅表现，其中 A 股基准指数为沪深 300 指数、港股基准指数为恒生指数、新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）、美股基准指数为标普 500 或纳斯达克综合指数。我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议；投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。

分析、估值方法的局限性说明

本报告所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。本报告采用的各种估值方法及模型均有其局限性，估值结果不保证所涉及证券能够在该价格交易。

法律声明

开源证券股份有限公司是经中国证监会批准设立的证券经营机构，已具备证券投资咨询业务资格。

本报告仅供开源证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。本报告是发送给开源证券客户的，属于商业秘密材料，只有开源证券客户才能参考或使用。

本报告是基于本公司认为可靠的已公开信息，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他金融工具的邀请或向人做出邀请。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动，过往的业绩表现不应作为其日后表现的预示。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突，不应视本报告为做出投资决策的唯一因素。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。本公司未确保本报告充分考虑到个别客户特殊的投资目标、财务状况或需要。本公司建议客户应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。若本报告的接收人非本公司的客户，应在基于本报告做出任何投资决定或就本报告要求任何解释前咨询独立投资顾问。投资者应自主作出投资决策并自行承担投资风险，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。

开源证券在法律允许的情况下可参与、投资或持有本报告涉及的证券或进行证券交易，或向本报告涉及的公司提供或争取提供包括投资银行业务在内的服务或业务支持。开源证券可能与本报告涉及的公司之间存在业务关系，并无需事先或在获得业务关系后通知客户。

本报告的版权归本公司所有。本公司对本报告保留一切权利。未经本公司事先书面授权，本报告的任何部分均不得以任何方式制作任何形式的拷贝、复印件或复制品，或再次分发给任何其他人，或以任何侵犯本公司版权的其他方式使用。所有本报告中使用的商标、服务标记及标记均为本公司的商标、服务标记及标记。

开源证券研究所

上海

地址：上海市浦东新区世纪大道1788号陆家嘴金控广场1号楼3层
邮编：200120
邮箱：research@kysec.cn

北京

地址：北京市西城区西直门外大街18号金贸大厦C2座9层
邮编：100044
邮箱：research@kysec.cn

深圳

地址：深圳市福田区金田路2030号卓越世纪中心1号楼45层
邮编：518000
邮箱：research@kysec.cn

西安

地址：西安市高新区锦业路1号都市之门B座5层
邮编：710065
邮箱：research@kysec.cn