海外算力：北美云商 ROI 自证充分，AI Capex 叙事强劲

20260907

海外大模型密集发布期临近，算力 ROI 与范式演进驱动商业化拐点

摘要

- 9-10 月海外大模型进入密集发布期，Anthropic 与 OpenAI 将分别发布 Claude 5.1/5.2 及 Astra/GPT-6 Soul 版本，驱动 ARR 冲刺。
- Astra 参数规模已超 10T，未来头部厂商目标指向 50T；闭源模型通过“防蒸馏”特性维持定价权，确保智能化水平领先开源模型。
- AI Self-Improvement（AISI）预示 2027 年进入自学习范式；Physics AI（物理 AI）与世界模型构建将成为实现 AGI 及机器人领域突破的中期关键。
- 商业化路径从 Coding 向 AI for Science 及企业级 AI 拓展；交互形态由文本转向持续性 Voice Chat，Agent 从个人助理向协同工作者演进。
- 预计 2026 年底 OpenAI 与 Anthropic 算力支出达 2000 亿美元，若 ARR 同步达 2000 亿则实现收支平衡，但非头部模型公司回本难度极大。
- AI 云租赁业务 ROI 明确，Blackwell 平台算力集群回本周期不足 3 年；云厂商现金流阶段性转负系投入产出时间差及模型研发开销所致，预计 4-6 季度后改善。
- 市场分歧点在于 AI 能否创造新需求，即便仅替代人力，10%-20%的 Capex 年增长率仍具支撑；硬件股呈阶梯式波动，建议在悲观情绪中寻找布局机会。

Q&A

从近期发布的模型（如 Astra）来看，当前 AI 模型迭代呈现出哪些新的趋势和特点，尤其是在用户体验和商业化潜力方面？

近期发布的 Astra 模型在用户体验上实现了显著进步，核心优势在于提升了复杂任务的可用性并降低了使用门槛。其拟人化程度持续提升，支持边做边聊、可随时打断的持续性交互。在实际应用中，Astra 能够处理更进阶的任务，例如在十几分钟内生成完成度相当高的游戏，这表明其处理长序列、持续性要求的能力相

较于之前的模型有显著提升，且耗时更短。同时，Astra 的 token 消耗量也显著高于早期模型，这反映了其具备较强的商业化变现潜力，可能驱动用户升级付费方案，从而拉动未来 ARR 的增长。

接下来一至两个月，海外 AI 大模型领域预计会有哪些密集的发布活动？这些活动对行业，特别是对 Anthropic 和 OpenAI 等公司的发展节奏可能产生什么影响？

未来一到两个月，海外大模型将进入一个密集的发布期。继 Anthropic 发布 Claude 5.1 和 OpenAI 发布 Astra 之后，市场预期还将出现 Grok 4.7、Meta 的 Watermelon 模型。此外，在 Anthropic 正式上市前，可能会发布 Claude 5.2。OpenAI 也将在 9 月 29 日的开发者大会上备受关注，市场期待其可能发布 GPT-6 的 Soul 版本。这一系列密集的发布活动表明，从 9 月到 10 月，行业可能会迎来一波 ARR 的冲刺，特别是即将上市的 Anthropic 和行业领先者 OpenAI，都有动力在此期间加速产品迭代和市场推广。

从中期维度看，维持海外闭源大模型领先优势的关键因素有哪些？模型参数规模的演进趋势是怎样的？

维持海外闭源大模型领先优势的中期关键因素主要有三点。第一，模型智能化水平需持续领先于开源模型。第二，有效防止“蒸馏”，即防止开源模型通过模仿闭源模型来缩小能力差距。近期发布的 Claude 5.1 和 Astra 都已开始强调其“防蒸馏”特性，以维持其定价能力和技术壁垒，尽管其有效性仍有待验证。第三，持续扩大产能与模型参数规模。目前 Astra 的参数规模已超过 10 万亿（10T），下一步美国头部厂商可能追求接近 50 万亿（50T）参数的模型，而中国字节跳动等公司也计划将参数规模提升至 10 万亿级别，以推动模型向 AGI 方向发展。

AI Self-Improvement 作为一种新的产业趋势，它对 AI 模型未来的迭代范式将产生怎样的里程碑式影响？

AISI 被视为 AI 发展中一个里程碑式的变化，它预示着模型迭代将进入自学习、自回归的新范式。理想状态下，模型将能够自主完成合成数据、自我训练、自我试错的完整循环，摆脱过去依赖人类反复试错的迭代模式。这标志着机器自身可以驱动整个迭代过程。如同 2024 年出现的 Reasoning Model 打开了新的 Scaling Law 曲线，并催生了后续从 Post-training、Fine-tuning 到代码生成等一系列应用创新，AISI 的实现也将成为通往 AGI 道路上的一个关键里程碑。因此，在 2027 年，AISI 的进展将是观察 AI 范式迭代的重点。

在 AI 商业化进程中，除了模型智能化提升，还有哪些关键的场景拓展和产品形态演进方向？

AI 商业化的推进不仅依赖模型智能化，还体现在场景拓展和产品形态的演进上。场景方面，继 2026 年开启的第一个商业化场景 Coding 持续渗透后，新的方向正

向 AI for Science 和 Enterprise AI（企业级 AI）拓展，覆盖金融、法律等多个垂直领域。产品形态方面，交互方式正从文本输入演变为 Voice Chat 和 Voice Command 等持续对话模式。同时，AI Agent 正从个人助理（personal agent）向协同工作者转变，后者具备自发寻找并执行任务的能力，不再仅仅被动接收指令。这些围绕 Agent 展开的应用范式变化，是一个持续深化的过程。

从更长远的中期维度来看，AI 领域下一个可能带来颠覆性变革的范式创新是什么？它对生产力的提升意味着什么？

从中期维度看，下一个更大的范式创新在于 Physics AI（物理 AI），其核心是世界模型的构建。如果 AI 能够成功跨越到 Physics AI 的维度，其应用空间将被极大地打开，并延伸至机器人、仿生人等领域。这将是实现社会生产力发生质的飞跃的关键条件。AI 的发展呈现出阶梯式进步的特点，即通过阶段性的重大突破实现跃升，而 Physics AI 的实现将是推动生产力实现下一次跨越式发展的关键一步。

如何看待当前 AI 硬件股票的市场表现，以及资本市场对 AI 发展前景存在哪些分歧？

AI 硬件相关股票的市场表现呈现阶梯式变化，其股价并非持续线性上涨，而是经历大幅上涨后可能出现大幅下跌的阶段性波动。因此，在整个资本市场对 AI 持乐观态度时，应相对谨慎，因为正向反馈可能不会持续，中间或将出现分歧。反之，当市场分歧极大、情绪悲观时，则应更为乐观，因为产业层面可能即将释放更多的进步与变化。 目前资本市场对 AI 的看法存在分化。一部分观点认为 AI 的进步具有局限性，主要体现在替代部分人类功能，而未能创造新的需求。另一部分观点则更为坚定，认为 AI 作为生产力工具的跃升，最终将创造出新的需求。即便仅考虑替代人类工作的价值，也足以支撑当前 AI 领域的资本支出维持在中低双位数的年增长率，例如 10%至 20%的水平。若未来能进一步证明其创造新需求的能力，增速则有望维持在更高水平。

AI 模型公司的投资回报情况如何？资本市场在评估其 ROI 时存在哪些普遍的误解？

AI 模型公司的回报主要通过其自身收入与算力租用成本进行衡量。以头部模型公司 Anthropic 和 OpenAI 为例，预计到 2026 年年底，两家公司将接近使用 10GW 的算力。按照目前每 GW 算力每年约 200 亿美元的常规租金计算，其年度算力总支出将达到 2000 亿美元。如果届时两家公司的年度经常性收入也能达到 2000 亿美元，则基本可以实现收支平衡。然而，除头部公司外，其他模型公司租用算力后可能难以取得足够的回报，普遍面临无法回本的挑战。 资本市场，[illegible] A 股市场，在评估 ROI 时存在一个普遍的误解，即将模型公司的显性收入[illegible]云服务提供商（Cloud Service Provider）的总体资本支出（Total Cap[illegible]行对比。例如，将两家头部模型公司 2000 亿美元的变现收入与 2026 年云

人在草木间

8,000 亿至 9,000 亿美元的资本支出直接对应，这在会计逻辑上是错误的。这种对比忽略了企业级算力消耗的复杂性，一部分算力支出实际上是替代了传统业务中的人力薪酬成本，这部分价值转换无法在财务报表中直接体现，导致外界难以准确计算其回报。

云服务提供商（Cloud Service Provider）的财务状况和 AI 相关业务的投资回报周期是怎样的？

2026 年，美国四大云服务提供商（Big Four）的收入体量预计在 2 万亿至 2.5 万亿美元之间，其对应的资本支出约为 6,000 亿至 7,000 亿美元。云厂商的算力投资本身就是对一种消耗品的投资，芯片的经济寿命通常为 3 至 5 年，实际使用寿命约 5 至 7 年，需要持续投入。 专注于提供 AI 云租赁服务的业务，其投资回报是明确且可观的。以一个 GPU 集群（GPU Cloud）为例，若投资一个 GW 的算力，采用英伟达 Blackwell 平台的建设成本不到 500 亿美元。最初，该算力的年租金可达 200 亿美元，且近半年到一年内，优质算力的租赁价格持续上涨。按此计算，整个项目的回本周期不足 4 年。考虑到项目中超过 20%的 Capex 是用于土建和楼宇等长生命周期资产（折旧年限可达 20-25 年），硬件部分的实际回本周期不足 3 年。

既然 AI 云业务回报周期短，为何部分云服务提供商的现金流会出现阶段性转负的情况？

云服务提供商现金流阶段性转负主要由几个因素导致。首先，从投资到产生回报存在时间差，在投资加速阶段，现金流自然会下滑。其次，云厂商自身也是模型公司，例如谷歌需要投入大量资金研发和维护其 Gemini 模型，这部分用于模型训练和推理的投入会产生额外开销。相比之下，模型业务投入较少的微软，其现金流表现则相对稳健。 以 AWS 为例，其现金流本身就因其零售业务背景而具有较大波动性。尽管其第一季度现金流为负，但第二季度的负值已有所收窄，显示出 AI 云业务的回报已开始体现在财务数据中。总体而言，只要专注于云租赁业务，其盈利模式是确定的。当前现金流的压力主要源于加速投入与回报周期之间的时间差，预计在 4 到 6 个季度后，随着 AI 云业务收益的持续兑现，现金流状况将得到显著改善。