

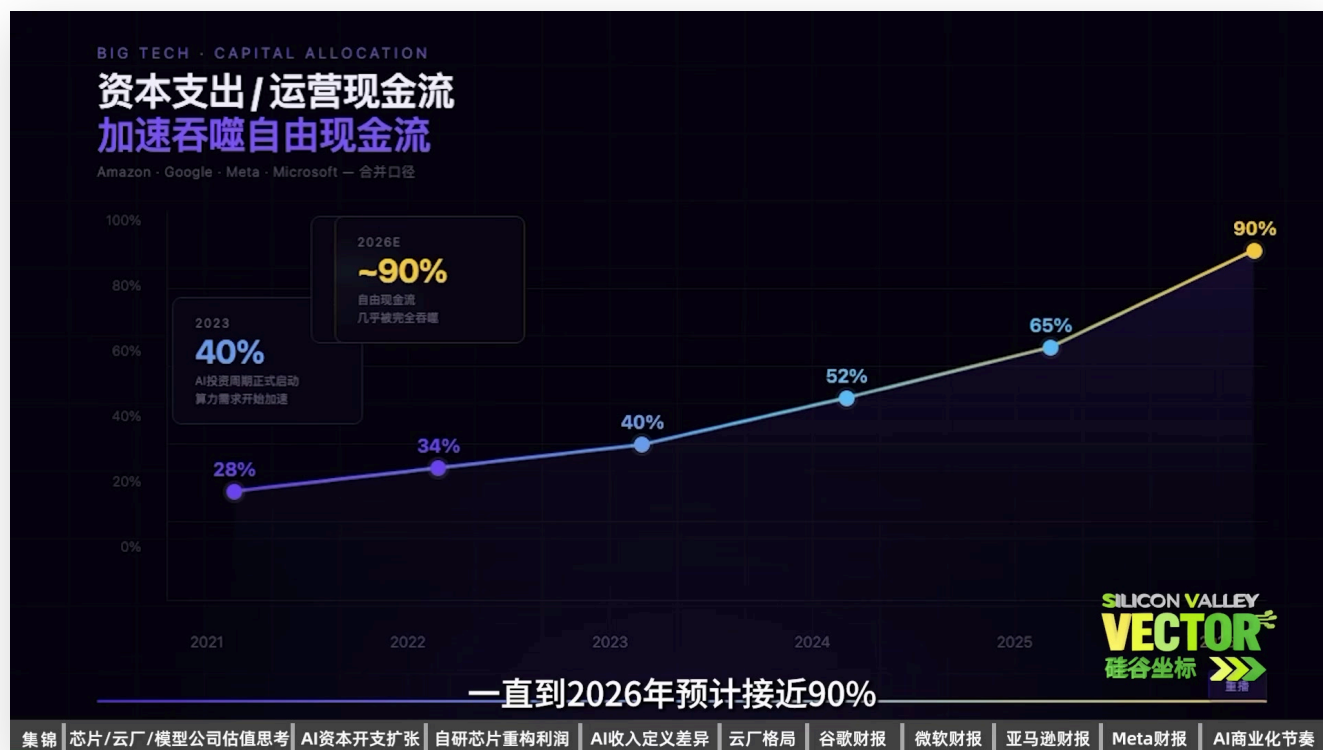
硅谷坐标 x FundaAI周默:四大科技公司财报后的AI产业深度观察 – 图文解说

一张关键画面对应一段完整解说，按原视频时间推进。

文章摘要

本期分析 Amazon、Google、Meta、Microsoft 的 AI 资本开支、自由现金流、云收入和自研芯片。周默将财报里的 CapEx、GPU/TPU/MTIA、Copilot 与 AI 收入放在同一张产业链地图上，讨论大厂为何必须继续投入、怎样收回投资，以及模型公司、云厂商和芯片公司的利润如何重新分配。核心结论是：AI 基建投入仍在扩张，但不能把资本开支直接等同于确定的收入兑现，真正的观察重点是利用率、产品化和持续现金流。

01. 开场：科技巨头资本开支吞噬现金流（00:01-01:32）



场景 1

本节主旨 开场：科技巨头资本开支吞噬现金流

完整内容

曹卿云：2026年，亚马逊、谷歌、Meta 和微软这四家科技巨头合计资本支出超过6300亿美元。这个数字占 GDP 的比例，接近20世纪所有重大基础设施项目的总和。更直观的是另一组数字：这四家公司的资本开支占运营现金流的比，从2023年的40%上升到2025年的65%，一直到2026年，预计接近90%。

说话人2：Every sign that we're seeing in our own work and across the industry gives us confidence in this investment.

曹卿云：亚马逊最极端，过去12个月的自由现金流已经从380亿跌到12亿，资本开支把自由现金流几乎全吃掉了。这笔钱到底怎么赚回来？

说话人2: **We are compute constrained in the near term.** As an example, our cloud revenue would have been higher if we were able to meet the demand.

曹卿云: 今天我们请到的嘉宾是 FundaAI 的创始人**周默**，一起聊聊如何思考科技大厂资本开支和 **AI 收入**。**周默**，欢迎来到《硅谷坐标》。四月底，几家硅谷科技大厂刚刚发布了新一季度的财报，那么几家大厂的资本开支和 **AI 收入**，也仍然是所有人最关注的焦点。

画面说明

最终关键帧显示一张“资本支出/**运营现金流**”折线图，图中可见四家公司合并口径从**2023年的40%、2024年的52%、2025年的65%升至2026E的90%**；另一张关键帧显示带有 **Amazon** 标志的**数据中心**建筑，字幕写着“过去12个月**自由现金流**已经从**380亿跌到12亿**”。

02. 股价分化与模型公司的估值（01:33-06:12）



场景 2

本节主旨 股价分化与模型公司的估值

完整内容

曹卿云：那么今天在聊财报之前，我想先跟你聊一聊几个大厂过去的股价表现。我们知道，在过去两年，可能英伟达都已经涨了几倍，但是大型科技公司的表现就远远没有那么亮眼。同样都是在做 AI，市场为什么定价得这么不一样？

周默：对，这是个很好的问题。其实从去年的7月份之前，这几个大的科技公司还是能够跟上半导体的涨幅的，或者至少从趋势上看，是很一致的。但是从7月份之后，其实是一个明显的分界线，就是大家要开始算2026年的 CAPEX。然后你算上2026年的 CAPEX 以后，就发现到了2026年，这些公司的现金流已经几乎没有了。现金流没有，其实很影响对于这些公司长期现金流的估值，所以你对它的收入要求就更高，你需要看到明显的收入加速。

但是如果你看典型的公有云，像微软还有 AWS，它们去年下半年的加速其实不是很明显。**AWS 到今年一季度才是比较明显的加速，所以就跑不赢其他半导体公司。**然后另外一点是，这一部分的 alpha 主要都在各种各样半导体的上下游公司，但是整体上，以前投资人对半导体所分配的份额都还比较小，大部分份额都分配给了这些大科技公司。那半导体的比重提高以后，就要减少大科技公司的配置。

所以两方面，一方面是基本面的原因，另一方面是配置的原因，就造成大科技公司跑得比半导体差。

曹卿云：我们从产业经济学的角度来看，利润率最高的地方就是资源最稀缺的地方。我们看到英伟达的毛利率达到**70%多**，然后云计算中心的经营利润率大概在**35%到45%**之间，**模型公司**的毛利率也是从去年的**三四十**，可能今年会有一些上升。那我们看到最稀缺的是芯片层，但估值最高的是**模型公司**，这个矛盾应该怎么理解比较好？

周默：如果你是用 PE 的角度去看这些模型公司，这些模型公司都很贵。但如果你从远期的角度来看，没有那么贵。比如我们举个例子，像 Anthropic 可能到**4月底**已经接近**450亿美金**的 ARR 了，到年底就是一个**1000亿美金** ARR 的公司。昨天刚传出来**9000亿美金**的估值，差不多到年底就是**9倍**的 ARR。如果你假设它以后能够做到**20%到30%**的利润率，差不多也就是**30倍** PE。

你像这些大科技公司，在过去很长时间也差不多是这个 PE。软件你去对标微软，微软常年一般都是在**35倍**估值左右。所以你往远期来看，没有那么贵。

然后另外一点，什么样的公司有这样的估值，也看它的 bargaining power、它在整个产业里面的稀缺程度。毫无疑问，从上云的 SaaS 时代，也就是从**2015年到2021年**，公有云公司都享有很高的估值，而且大家都喜欢这种轻资产的运营模式，可以躺着赚钱。

公有云公司有很高估值的时候，它的特点是什么？就是它在分发所有的 SaaS 流量。**虽然我卖 CPU 租赁，可能是一个40多毛利率的生意，但是我还卖软件、卖数据库，所以合起来我变成了一个60多毛利率的生意，而且所有的软件要通过我这里分发。**所以我其实是整个软件行业里面的入口。

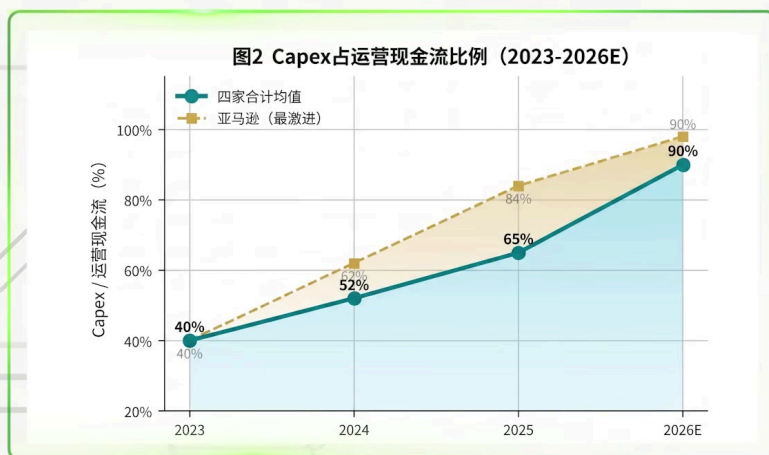
但是你到现在，这个入口的关系有点变了，模型公司有点像之前公有云的角色。如果你通过模型公司、通过 Anthropic 去买 API，或者你没有直接通过 Bedrock、Azure AI Foundry 这类服务去买 API 的时候，你其实不太关心这个 API 背后到底跑在哪个云上。**所以云从原来直接对接客户终端、bargaining power 很强的角色，现在变成了模型公司的下游。**模型公司以后有分配权，bargaining power 就会更强，也就变得更稀缺。

曹卿云：这个现象在今年越来越明显了。你怎么理解这一轮资本开支持续增加的逻辑？它是真实需求在拉动，还是更接近没有人敢停下来，大家都是军备竞赛的囚徒困境？

画面说明

最终关键帧为周默坐在书架前的访谈画面，画面下方字幕可见“你像这些大科技公司在过去的很长时间也差不多是这个 P/E”；画面没有额外图表或姓名条。

03. 资本开支扩张：需求、供给与租赁价格（06:13–10:59）



SILICON VALLEY
VECTOR
硅谷坐标

但 OpenAI 和 Gemini 它们还有季度环比接近 50% 的增长

集锦 | 芯片/云厂/模型公司估值思考 | AI资本开支扩张 | 自研芯片重构利润 | AI收入定义差异 | 云厂格局 | 谷歌财报 | 微软财报 | 亚马逊财报 | Meta财报 | AI商业化节奏

场景 3

本节主旨 资本开支扩张：需求、供给与租赁价格

完整内容

周默：其实资本支出这个故事，从2024年底到现在已经经过了好几轮变化。你像在2024年底的时候，大家担心的是数据不够，担心模型不能被泛化，这在当时确实也是真的，所以就出现了一月份的 DeepSeek moment。然后到了去年二三月份的时候，看到了 CoT 的量迅速起来，就不太担心短期的商业化了。

再往后，跑通了强化学习，跑通了 mid-training，跑通了 post-training，也跑通了前面的标注体系。这一套标注体系成为了一个完整的生态，所以强化学习不缺数据了。到了去年年终，它可以分配一半以上的算力到强化学习上面，所以强化学习成了新的增长的算力需求点。

这个趋势一直延续到现在，只不过这一轮有一个比较大的变化。如果我们看最近三个月，之前大家可能会担心**模型公司**的体量上去以后，ARR 会降速。但是过去三个月最大的差别是，当这些 frontier lab 的收入体量上去以后，增速不光没有降速，反而在加速。以增长最快的 Anthropic 为例，过去几个月基本上是每个月环比**50%的** ARR 增长。你可以说 OpenAI 和 Gemini 稍微慢一点，但 OpenAI 和 Gemini 也还有季度环比接近**50%的增长**。他们的体量已经比去年大很多了，所以变成了体量越大，增速越快，在第一季度有很明显的反应，那你的需求就越强。

然后到了衡量供需关系的时候，我们要看算力，或者反映到云的需求来看，到底是供过于求还是供不应求。我们有很多前置指标可以看这个问题。你可以看 **GPU** 租赁 contract 的 lead time，这个 lead time 是变长的，指的是你下合同到能够用上 **GPU**，大概要多长时间。或者你可以看 GPU 的单价，来反映供需的变化。

GPU 的单价分成好几种：长约、短约和现货。长约在这个季度之前，过去一年大概每个月都会降 0.05到**0.1美元**。但这个趋势从去年**12月**就出现了反转，今年无论是现货价格还是 GPU contract 价格，都在往上涨。价格往上涨，很明显就是需求很多。

往上涨的原因其实也很简单。你可以想象 GPU 租赁是一个大的池子。在之前的情况下，租长约的人最多可能占整个池子的**90%**，在**2024年**的时候。随着 GPU 的量越来越多，长约池会慢慢下降，从**90%**可能下降到**80%**，但还是比租 CPU 的时候好。租 CPU 的时候，长约可能占六成，短约占四成。

但这个趋势反转以后，大的 frontier lab 下了更多订单，锁了更多长约。因为强化学习、CoT 的需求，它们下了更多长约。长约比重从**80%**渐渐往上涨，留给短约的比重就会变小。短约是很有弹性的，也是最反映短期供需的，就像五一出去住酒店，酒店价格会涨很多。短约有点像酒店的价格，所以短约价格反弹非常明显，造成一季度整体 GPU 价格往上涨，供不应求的现象特别明显。

再加上**存储**涨价。**存储**涨价为什么会对这个季度的 CAPEX 影响非常明显？因为这批公有云普遍都是从**3月中旬**到**4月下旬**，去和存储公司谈长协。长协就是 long-term contract。谈完长协以后，就要锁定后面几个季度、后面几年的底价。锁了底价以后，这次存储上涨就不是一次性的事情，而是会反映到后面几个季度的变化。

所以**存储底价也涨了以后，你得一次性调整后面几个季度存储计入 CAPEX 的影响**。而这个事情可能还没调完，因为现在锁的是底价，没有锁住它继续涨价的趋势。

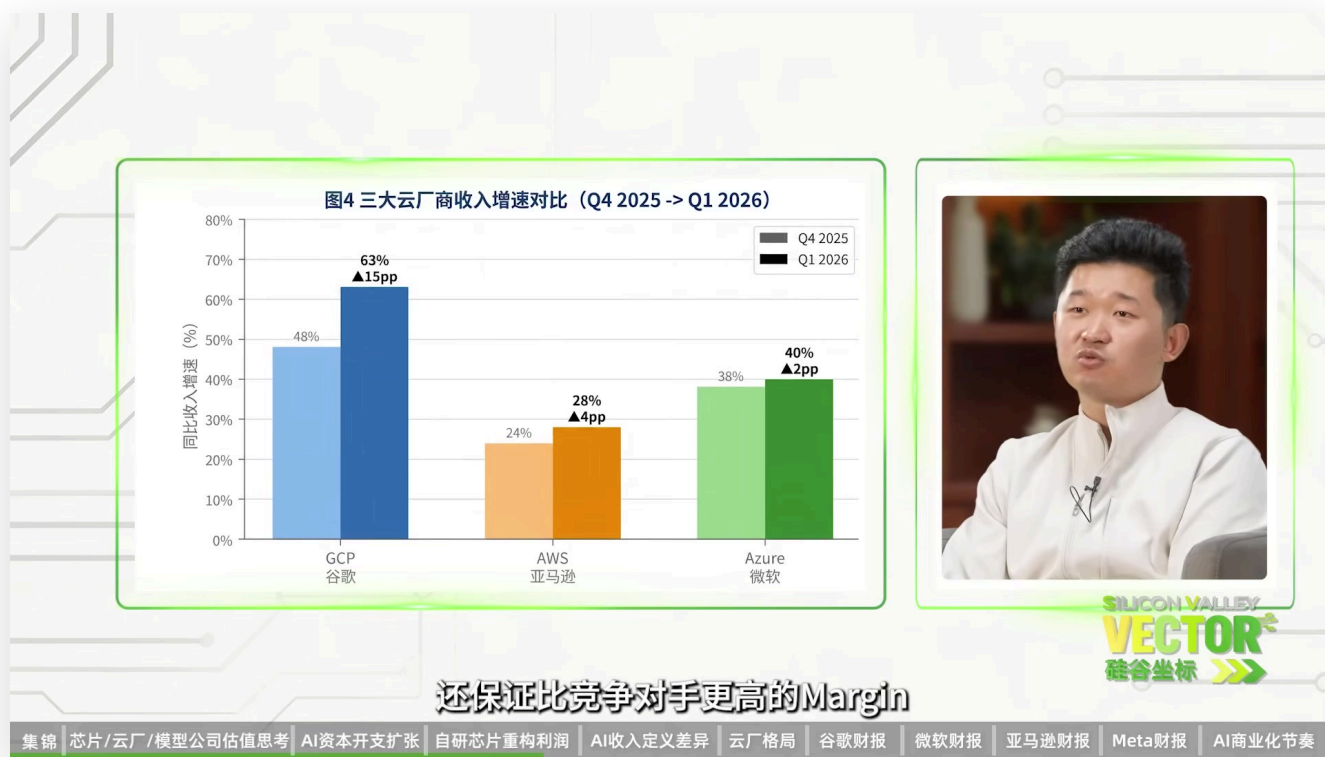
曹卿云：等到后面继续涨，它还得继续提。现在 GPU 还有**内存**价格的上涨，占到整个 CAPEX 上涨的百分之多少？未来这个趋势还会越来越大吗？

周默：两边都有，这个季度最直接的是存储涨价。

画面说明

最终关键帧显示周默在书架前的访谈画面，字幕可见“每一个月的生产力衡量的标准都是不一样的”之外，没有图表、产品界面或姓名条。

04. 硬件涨价与自研芯片的利润分配 (11:00-16:42)



场景 4

本节主旨 硬件涨价与自研芯片的利润分配

完整内容

周默：因为这个季度是大家统一签长协，后面几个季度都要调整底价。但是往后，其实 frontier lab 的需求现在也是供不应求的。比如你可以看现在 Anthropic 的需求：如果你直接去跑 Anthropic API，它特别慢；如果你通过 AWS Bedrock 去调 Anthropic API，可能速度快一点，但是也可能经常出现各种各样的问题。

你很明显能够看出来，现在 Anthropic 是严重的 supply constrained。如果你看 OpenAI 明年更激进的 Data Center 计划，明年的 CAPEX 肯定也要提。

曹卿云：如果把 CAPEX 上涨中间的这个蛋糕分一下，你会怎么分？有多少来自通胀、内存价格上涨，有多少来自需求扩张？

周默：如果看这个季度，绝大部分来自通胀；越往后，需求的影响越大。

曹卿云：现在我们知道这四家公司同时都在做自己的**自研芯片**，包括亚马逊的 Trainium、谷歌的 TPU、微软的 Maia 和 Meta 的 MTIA。它们的目标都是一样的，想把**推理**负载从英伟达迁移到**自研芯片**，把原来付给英伟达的钱留在自己的利润里面。亚马逊的 CEO **Andy** 这次在财报里也说了一句话，我印象很深刻：训练每年能为亚马逊节省数百亿美元的 CAPEX，同时也能够提供数百个基点的运营利润率优势。**那这对整个 AI 产业链的利润分配意味着什么？**以后谁的蛋糕会变小，谁的蛋糕会变大？

周默：这是一个很好的问题。但是亚马逊在这个季度有一点特殊，因为 Trainium 这张芯片，去年其实给很多客户打的折扣比较深。Trainium 最主要的客户就是亚马逊自己和 AWS，但是今年一季度是一个很特别的季度：所有 hyperscaler 都处在过去一年 bargaining power 最强、最供不应求的季度。所以 Trainium 这个业务，原来可能需要对外部客户打低价，现在到了 Anthropic 那里也供不应求，AWS 下面都是部署的。于是 Trainium 从原来一个不太产生利润的业务，到这个季度产生了很多增量利润。

但如果从整个云的体系来讲，今年想做一个很好的云业绩，有 Cloud API。无论是 Claude Code 还是 OpenAI Codex，有 API 就能带来收入加速。但如果有 ASIC，就能保证更高的利润率。同样以谷歌为例，它租 TPU 的毛利率比租 GPU 的毛利率高。所以从企业角度，ASIC 可以保证毛利率，也可以保证对客户和对芯片的 bargaining power。

从另一个角度讲，现在好的云公司，硬件公司的属性也很强。**Andy** 特别喜欢讲 Graviton。**Graviton** 在整个云行业是一个很重要的划时代事件，也是不断强化 AWS 过去几年生态位的事件。

把时间往前调，大概在**2021年** Graviton2 大规模推出的时候，很多 SaaS 公司纷纷调整收入预期和成本预期，就是因为 AWS 大幅使用 Graviton2。Graviton2 对亚马逊的帮助非常大：如果你有**30%**的性价比提升，传导**20%**给客户，那客户就会想用 Graviton2，而不是 x86。客户享受**20%**的性价比提升以后，中间就有**10%**的差价。这**10%**的差价可以变成 margin，也可以变成竞争优势。

如果变成竞争优势，就可以给 Graviton 降价。因为公有云的商业模型其实是一个通缩模型。在过去 CPU 时代，每年都要对 CPU 和所有产品降价，降价是因为摩尔定律。但如果有自研芯片，就能在降价的同时保证比竞争对手更高的 margin。所以对公有云来说，做自研芯片在历史上一直都非常有价值。AWS 过去就是靠单点最优、集中做硬件，在有市场竞争力的同时保持 market share。

所以到了这个时代，大家都想做 ASIC。但做 ASIC 肯定没有那么容易：有软件问题、CUDA 问题，还有非常多适配问题。这里面真正跑得通的是 TPU。TPU 的很多适配问题，已经和 Gemini 一起协作时跑通了，所以 Anthropic、**Meta** 使用 TPU 时，就没有遇到那么多问题。Trainium 和 Anthropic 之间的沟通，肯定不如 TPU 和 Gemini 团队自己的沟通，但对于所有云厂商来说，这都是 must do。跑得最通的是 TPU。

所以如果看切蛋糕的问题，TPU 产业链很多都跟光相关。我记得曹老师之前也做过一个光的访谈。如果看那个访谈，大部分 OCS 供应商其实在 TPU 里面都拿到了新的蛋糕。

我们说完支出端，现在也来说 AI 的收入端。

画面说明

最终关键帧包括周默的访谈近景，以及一张“三大云厂商收入增速对比”柱状图：GCP 从**48%**升至**63%**、AWS 从**24%**升至**28%**、Azure 从**38%**升至**40%**；图右侧为周默画面，底部可见节目章节条。

05. AI 收入口径、商业化路径与云的议价权 (16:42–33:40)



场景 5

本节主旨 AI 收入口径、商业化路径与云的议价权

完整内容

曹卿云：那在聊具体数字之前，可不可以请你先帮我们理清 **AI 收入** 这个概念？因为 **AI 收入** 这个词，每家公司说的其实是完全不同的东西，也是一个比较有争议的话题。

周默：对，每一家的 AI 收入概念非常不一样。我们以微软为例，微软之前会披露一个数值，就是 AI 占到 Azure 增长的多少个点。这里面包括很多场景，包括 **GPU 租赁** 的收入，包括帮 OpenAI 卖 API 的收入，也包括与 AI 相关的各种各样产品的收入。

比如，如果它定义某一项数据库产品 Fabric，或者甚至是 Databricks 的第三方产品，因为 AI 产生了新的增量，它也会把它定义到 AI revenue 里面。

如果看亚马逊，亚马逊主要定义的就是 AWS 的 Bedrock 收入。你看这个季度 AWS 的 Bedrock 环比增长了**170%**。Bedrock 里面的收入，其实很多是 Anthropic 给它的收入分成，因为 Anthropic 在 AWS 上卖的 API，大部分跑在 Bedrock 上面。除了 Bedrock，它还有一些芯片的 AI 收入，以及其他跟 AI 直接相关的 service 收入。

然后看谷歌，谷歌没有特别披露一个 AI 收入口径，但是谷歌的体系其实也非常清晰，包括 GPU 租赁收入、TPU 租赁收入、TPU 直接销售收入。还有 Gemini 的收入，Gemini 不是100%都在 GCP 体系里面，大概有70%多会进入 GCP 体系；再加上其他 AI 产品，所以谷歌里面的 AI 占比肯定是最高的，也反映在收入加速上面是最高的。

Meta 从来没有披露过 AI 占收入的口径，因为**推荐算法**对整个收入的贡献不太好界定。很多是算力提高、让**推荐算法**的模型复杂化带来的收入增量，也有 **Meta** 模型在 decoding、encoding 端带来的增量，或者是理解端更多的增量。所以很多时候只能说，今年可能有几个点的影响、带来了几个点的增量，但很难量化到底有多少收入是 AI 带来的。

每个公司的口径确实有不小的差异，但是今年有一个特点：从一季度开始，几个公有云的 AI 收入占比都开始加速了。

曹卿云：你刚刚说到这三种 AI 收入的路径，你觉得哪一种最可持续、天花板最高？

周默：我理解这里面有几种路径：一种是直接卖算力；一种是在算力以外再卖一些 AI 产品，比如 Office Copilot、Copilot Studio、Dynamics Copilot；或者是卖跟推荐算法有关的**增量收入**。

如果看收入天花板，毫无疑问公有云收入最高。因为从**模型公司**的角度讲，**模型公司**现在基本上已经不做什么 1P Data Center 了。OpenAI 之前有一个很大的 1P Data Center 计划，但是发现做 1P Data Center 非常 challenging，所以大部分 Data Center 也转成 3P 租赁需求了。Anthropic 虽然有 1P Data Center，但它的 1P Data Center 都是 GCP 把整个 TPU、机柜加 OCS 整包放到它的 Data Center 里面的。可能唯一自己做 1P Data Center 的现在只有 xAI。

所以大部分 Data Center 的量都在公有云和 neocloud 里面。如果看整个生态，模型公司不跟公有云抢 Data Center 的量，那公有云收入肯定最高。

但公有云的问题是，模型公司以后的 bargaining power 越来越强。比如 Anthropic 现在 bargaining power 就非常强，它可以租你的 TPU，也可以通过你去卖 API。API 这个事情很典型：一季度它可能有**80%到90%**通过云卖的 API 都走 AWS Bedrock。

从去年四季度开始，它也通过 GCP 的 Vertex AI 卖 API；到去年底，它和 Azure 也部署好了；从今年一季度开始，它也通过 Azure AI Foundry 卖 Cloud API。你想，一季度可能只有几个点的 API 跟 Azure 有关，但到二季度这个比重可能迅速提高到十几个点、二十个点。**所以你会发现，Anthropic 也会想通过不同的云渠道卖 API，越分散，bargaining power 就越强。**

包括 Anthropic 现在去接 Data Center 的条件也越来越苛刻。你是去和云签，还是和 neocloud 签？**因为 Anthropic 已经成了世界上最大的算力需求方，所以后面的条件也会越来越苛刻。**只不过今年一季度、二季度供需特别不平衡，所以对云的 bargaining power 会非常强；到了以后，其实不确定。

到了卖 Copilot、卖 agent 这里，有一个问题是，你很难证明自己做的 agent 产品能比 frontier lab 做得更好。这个问题在微软身上会非常明显。虽然 Azure 占了微软接近四成的收入，但是微软大部分利润还是由软件贡献的。Office 是全世界最大的软件产品，有**4亿用户**，无论从收入体量还是用户体量上都是全世界最大的产品。

但是现在的趋势是，有 Claude Code，有 Gemini Enterprise，大家通过 computer use 以后也可以调用 Office，甚至多模态 PPT 输出上面，Gemini Enterprise 的效果还可能更好。**所以从去年下半年开始，大家就在质疑：微软除了是一个云公司以外，还是全世界最大的软件公司，是最 software 的 megacap。**它的 agent、Copilot 这套产品到底能不能跑赢这些大厂的 LLM？

大公司的工作方法和创业公司在现在的 AI 线上完全不一样。原来产品迭代可能两三个月一次大的迭代，季度维度的迭代符合 Office 的开发思路。但是从去年**12月底**到现在，Claude 4.5 出来了，Claude 4.6 出来了，大家开始做 multi-coding-agent 工作；Claude 4.6 出来以后，task design 能力大大加速了编程速度；到**3月份**，harness 出来了。

以 Copilot 为例，Copilot 之前在调度层用了很多 hard code 的 prompt engineering，也比不过 Anthropic 的产品。所以大的软件公司在 AI 开发上面本身没有那么 native。**很多时候创业公司一周就开始做一个很大的迭代，但是对大公司来说，三周可能连 PPT 到汇报的流程都还没做完。**

所以从市场上来讲，大家没有那么看好 Copilot 或者 agent 这些产品在 megacap 里面真正带来的收入贡献。微软那边 Office 的增速一直是**14%、15%**，可能偶尔有一个点加速，但始终没有云那么强的爆发力。作为云的对照，AWS 的收入从上个季度**24%**变成这个季度**28%**，GCP 从上个季度**48%**变到这个季度**63%**。云的加速比重远远超过 Office 的加速比重，甚至还要担心 Office 可能会有席位减少。

对于这些公司，至少从长期来看，推荐算法端的竞争优势也在变化。**2024年和2025年**，大家最喜欢 AI 对广告的贡献，它对广告公司的增量是全局的贡献，贡献两到三个点，是对总收入两到三个点的贡献。云里面贡献的几个点，是对云收入的贡献，而且很多贡献会带来利润率压力；对于推荐算法公司，贡献反而对利润率有帮助，因为买 GPU 的时候本来就考虑了 ROI，Meta 买 GPU 也会考虑 GPU 和推荐算法 ROI 的关系。

但问题是，推荐算法的贡献在整个业务里面比重会越来越小。一开始有 low-hanging fruit；再往下就要 scale up 推荐算法模型。以前是7到8层、金字塔结构，现在要变成十几二十层，把金字塔变成圆筒形结构。你投入更多，但可能只有一两个点的 ROI 贡献，在整个系统里面就没那么明显了，无论反映到 margin 还是利润端，都没有以前好看。

所以现在三种模式的角色也在互换：以前大家喜欢 AI 对广告的贡献，从今年开始，慢慢开始喜欢 AI 对云收入增速的贡献。**刚刚说到微软的 AI 年化收入有370亿。**

曹卿云：这也是这次财报季最大的 AI 收入数字之一。外界对这个数字也很有争议：到底有多少是 AI 新产生的收入，有多少是原来就会发生、只是换了一个 AI 标签？你会怎么判断 AI 真实的**货币化**进度？

周默：最重要的就是看增速。**Backlog** 反映长期需求，如果有足够的 supply，就能够把收入放出来。我觉得这几家公司的 AI revenue 确定性都很高，收入真实性也很高。

每家云公司的加速频率不一样。比如看 GCP，它的加速频率肯定最高。它有几个特点：Gemini 的 revenue 大部分、**70%以上**都 book 在 GCP revenue 里面。Anthropic 通过 AWS 卖的 API，首先 API 只是 Anthropic ARR 的一部分；卖的 API 里面，可能只能把**40%的** revenue book 到 AWS 的 business model 里。

第二，GCP 的 TPU 数量最大，所以能看到 GCP 的加速非常明显。如果有 API 的 ARR，又有自己的 ASIC 芯片，对收入的帮助就会更大。

微软就不是特别讨巧。**第一个点是微软没有 TPU，也没有 Trainium 这种 ASIC。**它现在有 Maia 200，但 Maia 200 现在只能跑 GPT-4o，也就是稍微小一点的 GPT 模型；跟 coding 相关、跟 Codex 相关、跟 GPT-5.2、5.3、5.4 相关的模型，都没有办法在 Maia 上跑。

第二，OpenAI 的 API 增长以前也不是 OpenAI 主要的增长点，OpenAI 主要靠产品增长带来收入，跟 Anthropic、Gemini 不一样。所以它的 API 体量在 Azure 里的贡献本来就没有那么大，增速也没有那么快，没有 Anthropic 到 AWS、Gemini 到 GCP 那么强的爆发力。其实这跟你的 business

model、卖 API 的角色，以及你怎么看待**自研芯片**，都更相关。

曹卿云：刚刚聊到云业务的持续加速。这个季度谷歌云增速达到**63%**，Azure 是**40%**，AWS 是**28%**。最直观的感受是谷歌云加速最快。但是如果从绝对值来看，AWS 新增的绝对云收入，实际上比谷歌整个季度新增的还要多。亚马逊也提到，AWS 增速上限是 Data Center 建设的速度，而不是客户需求比另外两家云更慢。Bedrock 这一季度处理的 token 数，也超过了**2023年**上线以来的历史总和。我们应该怎么理解现在几家云的格局？更应该看增速百分比，还是看增量绝对值？

周默：肯定看增速百分比更重要，或者看每个季度新增收入量里，**incremental revenue** 里面每一家的占比更重要。这里面明显是 GCP 和 AWS 量更大一点，Azure 的量会更小一点。

云的格局里还有一个领先指标叫 **RPO**，也就是剩余履约义务，是已经签好的合同，但是还没有交付，也没有算净收入。它是一个领先收入6到12个月的指标，比当下增速更能说明未来。

曹卿云：**RPO** 的增速这次非常惊人。对于三家的积压合同，你觉得如何看待它们的意义和质量？

周默：质量都很高，只不过口径上大家有点差别。比如看 backlog 的增长，GCP 的 backlog 增长肯定最快。**GCP backlog 增长里面一个很大的原因，是它开始直销 TPU 了。**

直销 TPU 和租赁 TPU 有很大差别。如果租赁 TPU，生命周期是**5年**，相当于每年确认五分之一的收入；但如果直销 TPU，TPU 一下就卖出去了。当然可能是 net revenue 口径，要扣掉博通的成本，但无论如何，都比租赁确认的 backlog 大很多。所以 GCP 的 backlog 里面有很多是口径原因。其他两家公司跟口径关系没那么大，反映的是长单，也就是长合同的占比提高了，长的单子越来越大了。

曹卿云：微软的 RPO 里面有**45%**来自 OpenAI，你觉得这是比较担心的话题吗？

周默：这一直是大家比较担心的话题，特别是在 Azure 和 AWS 的语境里，因为大家会觉得这不算 high-quality revenue。OpenAI 和 Anthropic 以后都会分散它们的云供应商。比如 OpenAI 在微软财报和谷歌财报前两天，和 **Microsoft** 更新了新的协议。

AWS 一季度的一些 capacity 问题，我理解也是为了等 OpenAI。OpenAI 其实很早就希望能够上 AWS，但是因为和微软的合作关系，一直到了四月份才解决这个问题。**所以一季度很多留给 OpenAI 的量，可能一季度没办法确认到收入里。**大家会担心以后模型公司的份额会变化：今年在你这里，明年会不会就在别的地方？

如果因为接了很多大客户的量，导致毛利率下降，大家会更担心。因为毛利率下降以后，大家会担心你在新的 GPU 场景里面还没有跑稳，不像以前在 CPU 上那么强。再加上绝大部分收入来自单一客户，就会进一步担心 margin。

但是整个市场其实也很宽容。如果你的加速比重很高，这个担心就没那么大。最担心的是加速比重不高：每个季度是平的，或者只加速一个点，但 CAPEX 又涨得很快，大客户比重又很高。那是大家担心的。如果你加速很快，其实也还好。

接下来我们来过一遍这几家大公司。首先我们来看看谷歌。

画面说明

最终关键帧为白底绿框的图表与周默画面组合：左侧可见“图14 Anthropic API渠道分散化：议价权持续上升”，饼图标注 AWS Bedrock、GCP Vertex AI、Azure AI Foundry 及不同季度比例；右侧为周默访谈画面，底部可见节目章节条。

06. 谷歌：TPU 供给、Gemini 与搜索广告变现（33:40–46:52）



场景 6

本节主旨 谷歌：TPU 供给、Gemini 与搜索广告变现

完整内容

曹卿云：谷歌这次的业绩很亮眼，包括云的利润率从大概17%一下跳到30%多，积压合同也翻倍，第一方模型每分钟处理的 token 数增长了很多。对此你怎么看谷歌的护城河，以及 Gemini 对谷歌护城河的贡献？Gemini 的模型训练思路，可能和原来的 OpenAI、Anthropic 稍微有点区别。

周默：Google 很多优化是围绕 benchmark 做的。谷歌在 TPU 上有非常大的优势：同等规模 FLOPs 下，TPU 训练模型的成本可能只有 OpenAI 用 GPU 训练成本的20%到30%。所以谷歌做模型，可以在同一代做出比 OpenAI、Anthropic 大两到三倍的模型。

但现在每一代模型参数大概就是三倍，所以谷歌其实是用下一代的参数规模，去对比这一代的 OpenAI、Anthropic。因此它之前很多 benchmark，以及通用文本、语音场景，都能做得非常好。

但是到了 Claude，其实需要很多强化学习、很多数据标准。谷歌去年在强化学习上的投入没有那么多，今年才做了非常大的投入，所以 coding 可能还需要3到6个月时间去补上。确实，Gemini 的 ARR 在 GCP 上的爆发没有 Anthropic 那么猛。

但对于谷歌的好处是，这些收入基本全是它的收入，放在 GCP 的基数里面就非常明显。GCP 本来就比 AWS、Azure 小，Gemini 又能确认几乎全部收入，所以收入贡献在谷歌增速里面会很明显。Gemini 贡献了这个季度4到5个点的收入加速，还有 TPU 租赁贡献的加速，以及 TPU 直销贡献的加速。

我觉得谷歌的护城河在 BigQuery 里面，肯定比 AWS 和 Azure 要好。首先，因为有 Gemini，所以保留了 API 加速的爆发力，可以保证自己的商业模型和商业化闭环。另外，你有 TPU，就有更强的 bargaining power。大客户可以用你的 TPU，用你的 TPU 以后，margin 会比租 GPU 的 margin 高。所以做模型、向客户提供云租赁、做新一代 AI 租赁，对谷歌不是一个拖住 margin 的业务，但对另外两家云公司，可能会变成打低 margin 的业务。

这有点像当年电动车打油车。油车也想做自己的电动车，但会和原来的油车业务前后打架，所以在电动车投入上不够坚决。现在公有云上也出现类似情况，2024年底摇摆过一次，造成2025年拿 Data Center 的量不够。

AWS 一直非常坚定地投入 Trainium，Trainium 可能比 TPU 稍微落后一点；谷歌则一直坚定投入 TPU、投入 AI 生态化。而且它做 TPU 的 margin 比原来卖 CPU 的 gross margin 还高。你卖 CPU 的 gross margin 也是40%多，但做 TPU 的 gross margin 比 CPU 更高，所以谷歌在 AI 投入上非常坚决，做这件事情的 ROI 回报比 CPU 时代更高，激励机制也完全不一样。

曹卿云：这次劈柴也被问到，谷歌现在算力受限的情况下，TPU 是优先供给搜索，还是供给 Cloud 外部客户？他说谷歌内部有一个 ROI 框架，但没有给过比例。可以说，谷歌内部需求和外部 Cloud 客户在竞争同一批 TPU。这是 Google Cloud 增长的天花板，还是可以用更多 CAPEX 解决？

周默：如果能分给 TPU 更多的 CoWoS，谷歌就能满足需求，也希望有更多客户。Google 去谈这些 capacity 的时候，其实都非常 aggressive，希望要更多 CoWoS。但是到了产能这里，很多 CoWoS 被英伟达锁住了，导致没有办法给谷歌分这么多 CoWoS。谷歌其实想要很多 CoWoS。

TPU 是现在最供不应求的芯片，但问题是沒有足够高的产能。TPU 今年产能不到400万张，明年新分配给客户的产能已经将近1000万张。你想，明年是今年的很多倍，天花板也很高，可以对照英伟达。

而且再往回看，世界上的三个 SOTA 模型，有两个已经基于 TPU 训练了。Gemini 3 的上一代，以及即将推出的 MIXTO，pre-training 也是基于 TPU；只有 OpenAI 的 GPT-5.5 基于 GPU 训练。三个 frontier lab 里面，有两个使用 TPU 训练。**对适配大厂来说，Gemini 已经帮 TPU 跑通了很多适配问题。**

所以 TPU 的需求很强，但没有足够 capacity，内部需求和外部需求的冲突会越来越激烈。这个冲突去年还没有那么激烈，到了今年就非常激烈。今年谷歌要给出90多万张卡，大部分是 TPU v7，也有一部分 TPU v5，大部分都要最新 TPU 芯片。**你要训练下一代模型，谷歌也需要 TPU v7；甚至现在的3.5、4也不一定有足量的 TPU v7。**Anthropic 的 TPU 也完全不够，如果 Anthropic 一季度 TPU 很充分，就不会出现直接调 Anthropic API 很慢的情况。

但这里对谷歌来说是两个部门：一个是 Gemini，一个是 TPU 部门。**虽然两边合作紧密，但是不同部门。**从 TPU 角度，对外卖更多 TPU 是好处：越早进入 frontier lab 的适配流程，以后在市场里的 share 就越大。你现在在 Anthropic 上跑得越快、适配得越快，软件栈问题解决得越快，就越早能在 frontier 模型里面用上。每个公司都有惯性，越早进去，惯性就在你这里，而不在别的地方。

所以 TPU 部门肯定想在外面卖得更多，但这也影响 Gemini 的用量。今年不太好解决，TPU 严重供给不足；明年会稍微好一点，因为明年 TPU 拿到更多 CoWoS，从分配 CoWoS 那端也看到了 TPU 更持续的需求。

曹卿云：除此之外，还有谷歌的搜索业务。去年搜索业务遭到市场很大质疑，但今年搜索业务又加速了，似乎还开拓出了一个新的市场。你怎么看谷歌**广告收入**的加速？

周默：谷歌最近 AI 对它的贡献，确实比 AI 对 **Meta** 的贡献更加明显。AI 对**广告收入**的贡献，从去年二季度开始有明显变化。

谷歌有一个很反常识的地方：它其实有**60%到70%**的搜索 query 没有变现。很多 query 很长或者很短，变现价值都不高。对广告主来说，投关键词最直接的关键词最好，比如“我要买运动鞋”，这个 query 对 Nike 价值最高。

但如果用户发一大串 query，像跟 AI 对话一样：“我有什么需求、脚多少码、我喜欢旅游、现在应该买什么鞋？”这不是广告主最喜欢的，因为它没有直接和广告主的关键词匹配出来。所以很多问句形式、长 query，之前推不出广告。

有了 AI Overview 以及配套投放的 AI Max 产品以后，事情发生了变化。现在理解长 query 的能力比以前更强，而且可以在不同步骤里面理解该推广告的概率。一个人开始和 AI model 互动时，可能先说买鞋需求、喜欢的价位和品牌；互动两三轮以后，发现他已经在挑最想要的品牌、购买意愿很强，这时给他推一个大的 banner，不用太在意占页面多少，因为这是他购买前的最后一步，CPM 就应该卖得非常贵。

通过这种方式，就能解决过去不好变现的长 query，拿到更多增量。

另一方面，谷歌也很聪明。原来的广告都在 PageRank 那套 system 里面，一个页面可以推三个、四个广告。但谷歌每次推广告时，都可以判断：这次刷新后，是推三个传统广告好，还是推两个传统广告加一个 AI 广告好，还是在三个广告上面再加一个 AI 广告？到底哪种组合能带来增量 revenue？

比如去年洛杉矶有很大的山火，山火发生时，跟家庭、防火有关的保险，核心价值就特别高，可以卖特别贵。那段时间，谷歌推广告 query 时，三个传统广告下面还会再给一个 AI Overview 的 ads。所以它可以保证，在任何 PageRank 形式下都找到增量收入，同时满足 AI Overview、AI Max 的渗透率要求。

因为还有大量 query 没有商业化变现，在这些 query 里推 AI，既可以满足 AI Max 渗透率变化，又不用影响传统广告的渗透率。谷歌有很多方法做增量收入。

这件事在去年二季度、三季度、四季度的北美越来越明显。今年欧洲广告主也开始做，因为欧洲广告主一般比北美广告主慢两到三个季度，既有产品上线时间的影响，也有欧洲广告主合规和审核要求更高的原因。

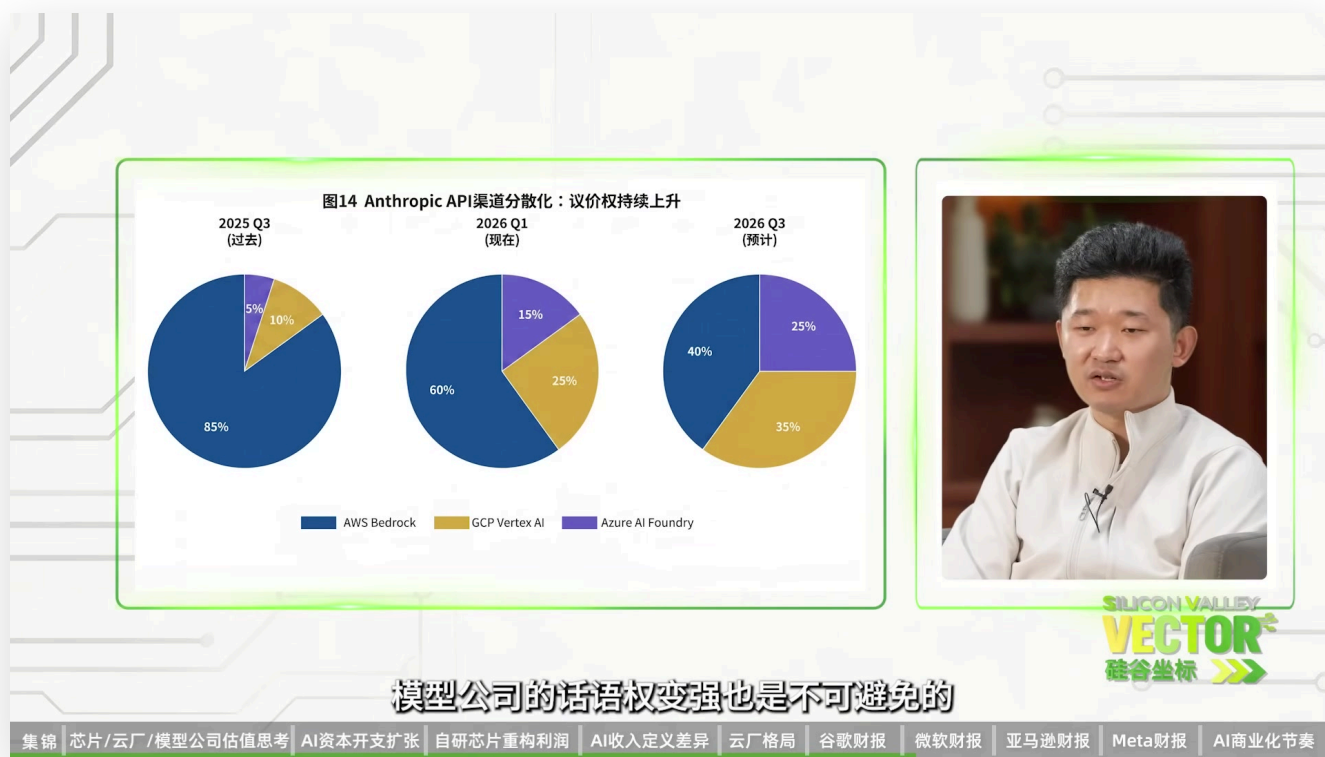
所以今年是欧洲广告主刚刚开始做 AI 的第一年，去年是北美广告主和亚太广告主做 AI 的第一年。

接下来我们讲微软。两年前，OpenAI 是微软的一个专属资产。

画面说明

最终关键帧为周默访谈画面，字幕可见“大部分是 TPU v7”；画面未显示额外图表、产品界面或姓名条。

07. 微软：失去独家合作后的 Copilot 竞争力（46:53–55:32）



场景 7

本节主旨 微软：失去独家合作后的 Copilot 竞争力

完整内容

曹卿云：OpenAI 的独家合作曾经是 Azure 最核心的差异化。但现在微软和 OpenAI 的独家合作消失了，亚马逊也宣布投资 OpenAI 500亿美元，OpenAI 也进驻了 AWS 的 Bedrock。你怎么看这个格局变化，以及未来 Azure 的核心竞争力？

周默：无论对于 Azure 还是 OpenAI，可能都是不可避免的。对双方来说，都可能是一件好事情。

从 OpenAI 的角度讲，OpenAI 之前 Codex 在 Azure 里面卖，其实丧失了最大的开发者市场。最大的开发者市场还是 AWS。如果你卖一个 coding 产品，一定要卖到开发者市场里面，不只是卖到 enterprise 市场，因为开发者市场的爆发力最强。所以 OpenAI 一直希望能够跟 AWS 有更深度的合作。

对于微软也是一样。API 这件事情本身在 OpenAI 收入占比里面不是很高，它也希望能够卖更多 Anthropic API，卖最强模型的 API。所以双方最后都会走到更加分散的场景。

另一个不可避免的趋势是，对于**模型公司**来说，**模型公司**的话语权变强也是不可避免的。Anthropic 原来主要绑定 AWS 做销售，去年四季度开始 GCP 也被接入，今年一季度开始 Azure AI Foundry 也被接入。它开始分散渠道，OpenAI 也开始分散渠道。模型公司的话语权越来越强，不和单一云绑定也是必然趋势。

曹卿云：你觉得 Copilot 应该如何应对像 Claude Code 这样的新竞争对手？

周默：现在竞争肯定越来越激烈。回顾当年 Copilot 出来的时候，Copilot 主打几个点。

第一个点，Copilot 在 Office 体系里面是 native 的，所以可以和 Office 各个产品无缝衔接。**Office 本身是一个 GUI 工具，各个产品背后其实都是 database，所以调用 database 时会更 native。**特别是在之前多模态 API 还没有那么好的时候，这套原生性让很多多模态内容识别得更好。

第二个点，因为和 Office 打通，可以用原来 Office 那套身份管理系统。权限管理在 enterprise 里面非常复杂。个人做一个 Claude Code，接到企业电脑上完成一个场景，可能要做十道授权手续，点来点去很累。**但是原来的 Office 身份管理已经走通了，所以不需要额外授权，对企业来说省了一笔权限管理设置成本。**对于用户来说，微软更懂 Office 的场景，也打通了权限管理系统。

但是它的质量未必比 Office 之外的产品高，这有技术原因，也有微软自身的特点。技术上，Gemini Enterprise 的多模态能力很强。Copilot 用的很多多模态能力，其实是上一代的 computer interpreter，不是这一代的多模态能力，所以在多模态做 PPT 上面不是特别好看。

再到更长任务的处理，Anthropic 的 harness 做得很好，harness 这个体系就是 Anthropic 提出来的。**到了 Office 场景里，哪怕是第一代 Copilot Cowork，它现在的 harness 也没做好。**可能要等它正式发布、打磨一段时间以后，harness 才能做好。

但是技术变化太快了，harness 这个事情还没给你很快的适应周期，就已经要求你的老产品整个 harness 化。**所以 Copilot 又有很高的技术要求，这也反映了 AI-native 公司迭代速度比传统公司快很多。**

站在 Copilot 角度，可能会说 harness 本身还是 engineering 工程。以前可能依赖 prompt engineering，很多场景靠工程师做；以后 harness 也是很多工程化工作，所以最终也能适应上，不一定是明显的算法劣势。**但是如果过几个月它不是 harness，而是下一代技术，你是不是又要适应一**

遍？

另外一点和微软自身问题有关。Office Copilot 是一个非常 enterprise 的产品，微软非常注重合规，所以使用模型时有一套调度层，根据 query 难易分配各种模型。以前最大的问题是没办法分配给 Claude Code，也没办法分配给 Claude 4.6，现在可以了，和 Anthropic 的合作已经深化。

但人为多了一层分配，在降本和满足延迟的同时，一定会牺牲一部分质量。第二，微软还搭了一层 Responsible AI。搭了 Responsible AI 以后，在涉及价值观、合规的 query 里，会人为做合规化处理。有些问题本来可以更有创造性地回答，但因为合规处理，反而降质了。合规、价值观和质量之间要做 balance，所以微软产品没有那么灵活，而且这个问题越来越大、越来越明显。

微软这几个季度没有显现出来，是因为整个大 beta 很大。微软其实有一个产品非常好，叫 Copilot Studio。Copilot Studio 比较早做低代码的企业内部 Agent Builder，可以做客服、HR 场景，现在甚至可以做面向终端客户的场景。这个产品卖得非常好，每个季度都是翻倍增长。

另一个好处是，面向企业内部的 agent，权限管理会更严格。你要打通各种数据产品，打通企业内部各种 SaaS 产品。Office 恰好已经把合规接完了，所以能比别的 Agent Builder 跑得快，销售网络也更成熟。

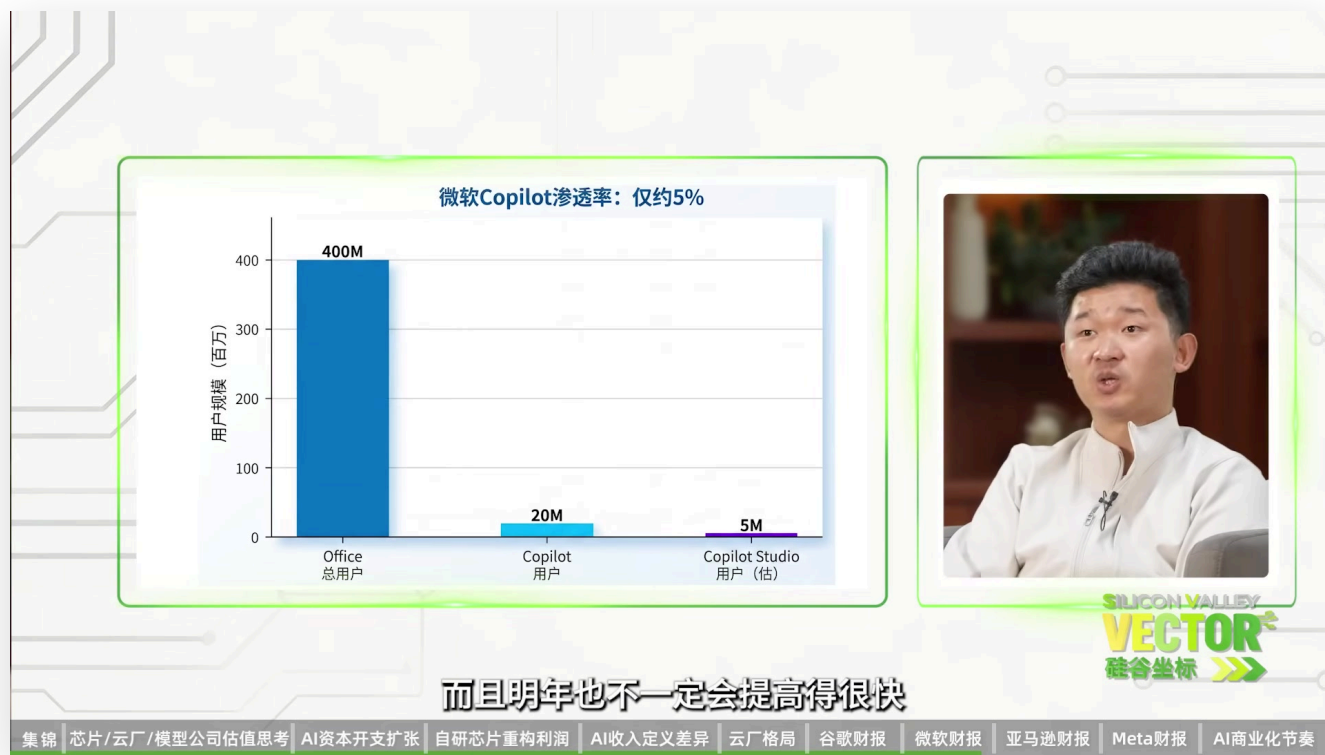
所以微软的挑战是：已经用合规、权限管理、身份网络和销售网络打开了 market share，但是产品能不能跟上这些 AI-native 公司的质量？

曹卿云：能不能保住这部分 share，这是微软以后面临的更重要问题。Satya 这次也说，Copilot 渗透率越来越高，从以前卖座位到现在卖使用量。

画面说明

最终关键帧包括周默访谈画面，以及展示 Anthropic API 渠道分散的饼图：图中可见2025 Q3、2026 Q1、2026 Q3 三个阶段，图例为 AWS Bedrock、GCP Vertex AI、Azure AI Foundry；底部字幕可见“模型公司的话语权变强也是不可避免的”。

08. Copilot: 从席位收费到用量收费 (55:32-58:00)



场景 8

本节主旨 Copilot: 从席位收费到用量收费

完整内容

周默: 那它对微软 AI 收入意味着什么? 其实你去看, Office 的增速还是因为 Copilot 和 Copilot Studio 加速了, 但是加速不是特别明显。因为对 Copilot 来讲, 整体渗透率还没有那么高。

Office 是一个**4亿用户**量的产品, Copilot 是一个**2000万用户**量的产品。如果按照 seat 去卖, 渗透率现在也就只有**5%**, 而且明年也不一定会提高得很快。所以对 Office 来讲, 最大的增长点是怎么把 Copilot 从一个按 seat 卖的商品, 变成一个卖 consumption、卖 token 的商品, 这就是 Copilot Studio 要做的事情。

我们估算 Copilot Studio 的 revenue 已经达到 Office Copilot 的五分之一到四分之一，而且这个比重会越来越快。Copilot Studio 是一个很典型的按 consumption 计费的产品：你已经用 Copilot Studio 搭成了各种内部 agent，build 了各种 agent template。搭好 template 以后，用户用得越多，消耗的 token 越大，consumption 越大。

所以现在一些 Office Copilot Studio 用户，一个月的花费确实达到1万到2万**美元**，而不是原来卖一个 Copilot 席位的**30美元**。现在要提高的是它的用量。

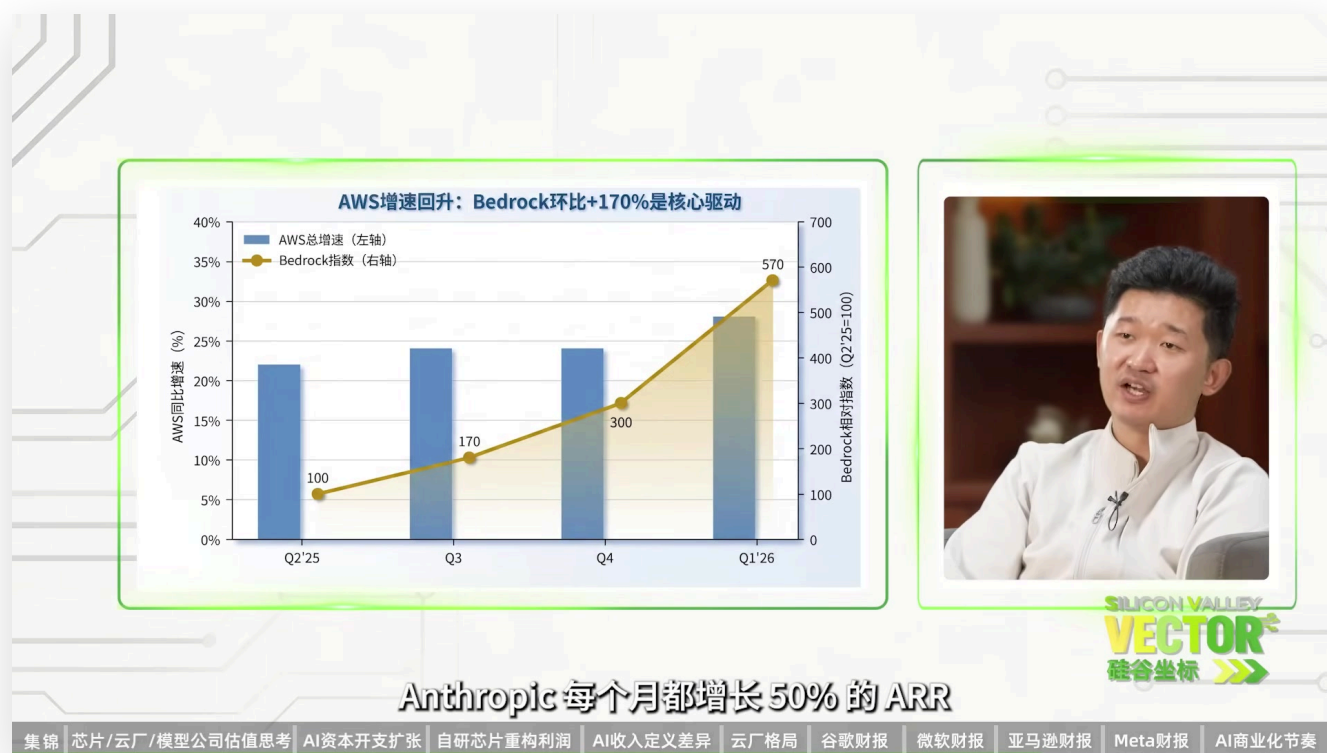
但与此同时，这个场景也是竞争最激烈的。回过头来看，很像当年 Palantir 做 AIP，或者原来做机器学习的场景。Palantir 刚开始做这些场景时，有很多定制化 case、很多 use case，慢慢搭建，最后变成一套标准化产品。

Copilot Studio 也是这个想法，Palantir 是这个想法，Gemini Enterprise 也是这个想法，OpenAI 和 Anthropic 的商业化也是这个想法。所以以后大家都会重叠，都往这个场景做，竞争会非常激烈。现在有先发优势，但不一定以后还能保住。

画面说明

最终关键帧显示“微软 Copilot 渗透率：仅约5%”柱状图，图中可见 Office 总用户400M、Copilot 用户20M、Copilot Studio 用户估计5M；右侧为周默画面，底部字幕可见“因为你已经用 Copilot Studio 搭成了各种各样的内部 Agent”。

09. 亚马逊：AWS 加速背后的收入与利润逻辑（00:58:00–01:05:13）



场景 9

本节主旨 亚马逊：AWS 加速背后的收入与利润逻辑

完整内容

曹卿云：这个先发优势能够保住。接下来我们聊一聊亚马逊。亚马逊今年业绩最亮眼的是什么？因为去年大家觉得亚马逊在 AI 上落后了，而今年 AWS 的增速又回到了**28%**，也是**15个季度**最快的一个；运营利润率**13.1%**，也是历史最高的。亚马逊从不被买账到重新被市场喜欢，这个认知转变的核心是什么？

周默：亚马逊确实经历了几轮叙事变化。你会发现，亚马逊其实在去年年中的时候股价是一个高点，去年底又跌下来，到了今年又涨回去，现在又重新到了新高，很明显经历了几轮 narrative 的变化。

应该这么说，亚马逊在**2024年下半年到2025年上半年**，是上一个 AWS 加速周期，从**16%到17%的增速**，一直到**20%出头**。这是因为当年在卖 GPU，GPU 上量了。到了去年，亚马逊又碰到了很多 capacity 问题。

很多 capacity 问题也是因为亚马逊是一家很像硬件公司的云公司，希望把很多问题在硬件层面解决好。它做 Data Center 时，既希望放 GPU，又希望放 Trainium，所以对 Data Center 的建设要求**非常高**。无论是冷却、电，还是很多资源分配，要求都非常高，就造成里面很多问题总是 delay。亚马逊买了这么多电的 quota，但是放不出收入来。

去年亚马逊永恒的主题就是，大家都觉得它又要加速了，但最后发现加速不是很明显；一问公司，公司说有产能瓶颈。很多产能瓶颈也是来自它希望把硬件环节做得最好，能够 leverage 自己**自研芯片**的优势，实际上这个难度很高。

到了今年，问题解决了，Data Center 的问题已经解决。它又买了最多的电，亚马逊去年买了**4GW**的电，比微软买得多，所以今年不缺电，能放出更多用于租赁的 capacity，于是有**更多容量做加速增长**。

另一方面，今年一季度整个模型的 alpha 肯定在 Anthropic。Anthropic 每个月都增长**50%的 ARR**，大部分 API 的量放在 AWS 的 Bedrock 上，所以造成 AWS Bedrock 增速非常快，公司说 Bedrock 环比增长**170%**。这个量跟 Anthropic 收入环比增速基本完全对齐，按 Anthropic 收入环比增速来算，差不多是**160%到170%**。

ARR 是一个 runway 概念，收入是一个平均概念，所以在加速过程中，ARR 的环比增速一定比收入增速快。

对于 AWS 来说，这个季度确实有一个变化：**Anthropic revenue 对 AWS 是 low-quality revenue，因为 Anthropic bargaining power 很强，所以会影响 margin**。如果 AWS 能证明 margin 没有受到影响，对 AWS 的逻辑帮助会更大。

恰好今年 AWS 的 Trainium **利用率**上去了，Trainium 供不应求，不需要在外面打折，所以 Trainium 帮助 AWS 稳住利润率。AWS 无论从加速还是利润率来讲，都能证明这是 high-quality revenue，也能证明它对收入加速有贡献。

这个季度亚马逊的业绩不是完美的，市场预期 AWS 增速能够到**30%**，但实际是**28%**。但是公司描述得很清楚：**自研芯片同比增长40%**，Bedrock 增长**170%**，说明收入没问题，增长就是这么快；Graviton 和 Trainium 收入也没问题，涨得就是这么快。

我理解一季度其实受到了 Blackwell 放量的影响。Blackwell 放量可能有一部分是 capacity 问题，也可能跟 OpenAI 合作的问题有关，但是下个季度这个问题大概率会解决。

曹卿云：你刚刚讲到，亚马逊已经累计投资 Anthropic 超过330亿，Anthropic 也承诺在 Trainium 上消耗5GW 的算力。同时亚马逊也投资 OpenAI **500亿美元**。AWS 既是 Anthropic 的投资人，也是它的算力供应商，同时还在跟 OpenAI 合作。这个多重角色的战略逻辑是什么？

周默：对于 AWS 和 GCP 来看，你可以看到它们都投资了很多**模型公司**。AWS 之前是 Anthropic 的主要投资人，现在也希望成为 OpenAI 的投资人。

从云的角度讲，它希望锁定两件事。**第一是锁定自研芯片的用量，需要有一个大的 frontier model 客户作为自研芯片的场景，才能跑通各种软件栈和适配问题。**就是我给你钱，让你用我的芯片，帮我跑通这些问题。

第二是希望把**模型公司**锁在自己的云上。除了自研芯片，也给你更多 GPU 用量，让你的 API 放在我这里，让我在你那里 API 的份额更大、在 inference 里面的份额更大。AWS 的逻辑是这样，谷歌的逻辑也是这样。

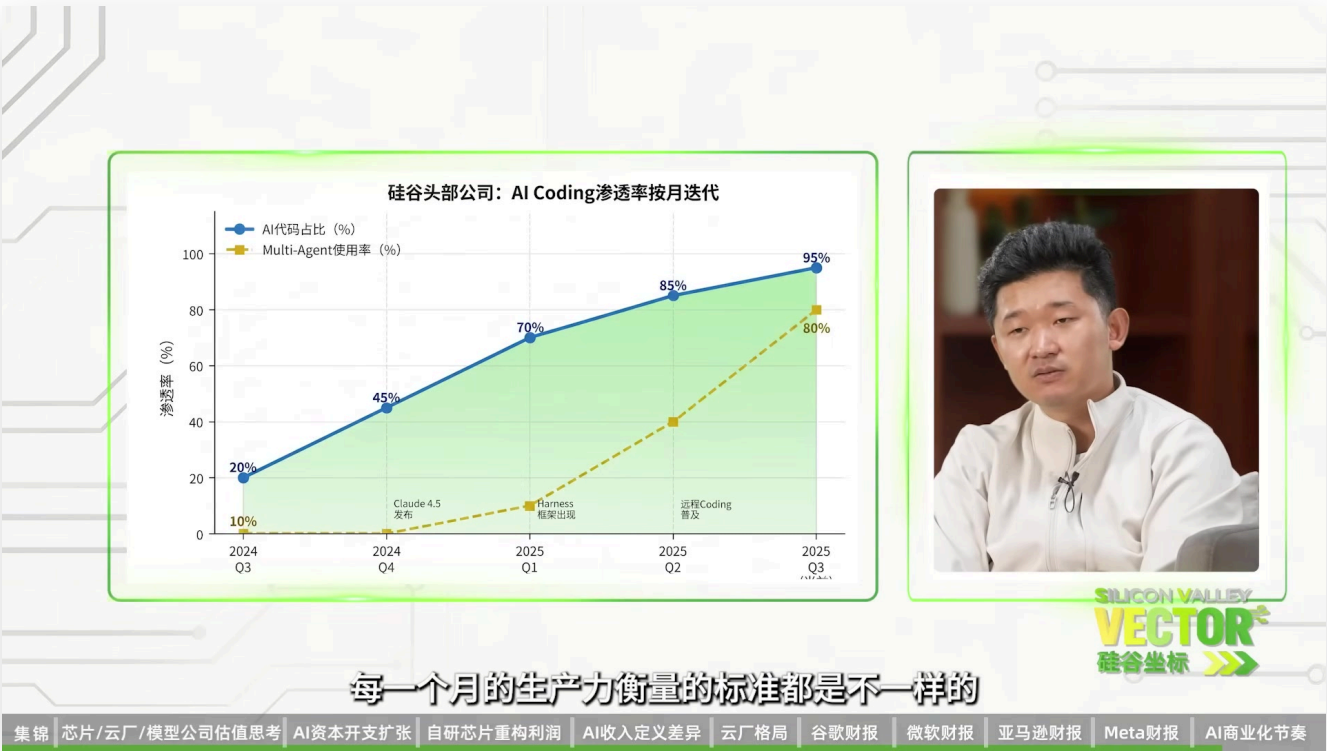
谷歌也是 Anthropic 的投资人，最希望 Anthropic 成为 **TPU** 最大的外部客户，把 **TPU** 场景跑起来。从 TPU 角度讲，它已经完成了一部分任务：**Meta** 模型已经可以在 TPU 上 pre-train 了，但 Trainium 还没有达到这个要求。现在有些模型还没有办法在 Trainium 上 pre-train，不过很多 experiments 已经可以在 Trainium 上做了。

曹卿云：最后讲到 **Meta**，这次广告数据里也有一个挺有意思的现象：广告展示量增长了**19%**，同时

画面说明

最终关键帧显示 AWS 增速与 Bedrock 指数的组合图：标题可见“AWS增速回升：Bedrock环比**+170%**是核心驱动”，横轴为 Q2’25、Q3、Q4、Q1’26，Bedrock 指数从100升至570；右侧为周默访谈画面。

10. Meta：广告提效、模型自用与 AI Coding（01:05:14–01:27:20）



场景 10

本节主旨 Meta：广告提效、模型自用与 AI Coding

完整内容

曹卿云：这个广告单价也涨了12%。我们知道在广告行业，这两件事情通常是矛盾的：增加了广告库存，供给多了，单价应该往下降。但这次两者同步上涨，怎么理解这个增长？

周默：我觉得这个季度 Meta 的增长，同比加速了9个点。其中4个点是汇率影响，两个点是2024年底有美国大选，所以基数稍微低一点。还有三个点的影响，一半是推荐算法的贡献，一半是产品维度的贡献。

对于 **Meta** 来说，很多库存变化来自产品形态的变化。Meta 走的路很像当年**2020年、2021年、2022年**抖音在国内走过的路径：从去年三季度开始，猛推一个广告产品，叫 overlay ads。Overlay ads 就是软广。

广告的视频形式里面会加一个广告。我们在抖音上看过很多，比如一个主播讲了五分钟，突然开始推转转，或者推瓜子二手车。在 **Meta** 里面，这个东西刚刚起步，而且美国很多广告主和 agency 还没有找到加软广的方法。

以前 Meta 对这里限制得很严，会要求广告形式和视频形式匹配。比如你是运动鞋广告，尽量放在运动博主的视频里面，不能在运动博主的视频里推二手车广告。但是这种形式在国内抖音已经很常见了，所以这涉及匹配机制，也涉及生态和广告主怎么学习加软广。

这个过程一旦开始搭生态，库存就会放得非常快。所以最近两个季度，Instagram 的广告库存放得非常快。软广本身又是一个很好的广告形式，可以卖不错的价格。

Overlay ads 从今年一季度开始改名叫 partnership ads。对 **Meta** 来说，它能够在不影响广告单价的情况下，拉动 impression 增长。

广告单价方面也有很多影响：广告正常就会有单价提升，也有通胀影响、各种其他情况，以及广告推荐效率变化的帮助。但是**推荐算法**对于 Meta 这样的广告公司，每年不是5个点以上的东西，而是5个点以内的贡献。单个季度能够有一到两个点的贡献已经很好了。

为什么一季度贡献会很好？因为去年**12月底** Meta 上了一个模型，叫 LLaMA。这个模型在 Meta 财报后发了 paper。对大多数 AI 公司来说，发 paper 并且 paper 被公司采用，是少见的事件。

我们可以有意思地说，Gemini 发的大部分 paper，跟 Gemini 真正的算法迭代都没有强相关性；Gemini 大多数 paper 是一年以前的技术，或者不会用到的技术。**OpenAI 和 Anthropic 也是，大家都不希望技术被竞争对手学到，所以倾向于发过时的 paper，或者没有 significant contribution 的 paper。**

大家可能记得，一两个月前出现过一个谷歌优化 KV cache 的乌龙，后来发现那个 paper 在 Gemini 里根本没用，因为只能用在**10倍**以下的东西。但 Meta 这次不一样：这次 LLaMA 是真的做得好，才发 paper。我觉得 LLaMA 对一季度有比较明显的贡献。

但是这个贡献衡量下来也是小个位数的贡献。大多数情况下，一个季度的贡献可能不到一个点。我们在 AI、frontier model 时代，会觉得每年几个点的贡献习以为常；但在 frontier model 之前，在引入模型、引入 GPU 之前，当年一次大的算法迭代能够有两个点的转化率提高，在整个行业已经是非常震惊的事件。

所以推荐算法的改动不是大几个点的贡献，每年更多是小几个点，更多还是 Meta 自己产品迭代的变化。

Meta 过去几年，无论做 conversation API，还是做 Advantage+，都是投放迈向自动化的过程，只是不一定是大模型语境。

曹卿云：Meta 做 AI，除了对推荐算法有帮助之外，还有什么意义？

周默：它会有短期帮助，也会有长期帮助。

短期帮助里，大部分非推荐算法的变化主要集中在 Meta 的几个产品。比较典型的是 Advantage+，它是一个自动化投放系统；还有 Advantage+ Creative，可以做 AIGC 素材，这两个产品可以联动，可以轮播，自动做素材广告。

你放一个广告样式，就可以让 Advantage+ 调不同背景，做图生图，调不同素材，做到千人千面，提高精准度。很多第三方 AIGC 素材在 Meta 里面用不好，一部分是审核原因，还有很大原因是没法和自动化系统做轮播投放联动。

另一方面，Meta 在很多场景也开始做 agent 体系。Meta 有一个很好的产品，Business Messenger。我们想象一下，有点像在微信朋友圈里碰到一个企业微信广告，点进去以后，由一个企业微信的人接触你；这个人可能是真人，也可能是 agent，你可能不知道。

Business Messenger 也是这样。你跳转以后，它会在 Messenger 里，也可能在 WhatsApp 里。原来这类销售线索产品，Meta 主要在东南亚推得好，因为东南亚人力成本便宜。但如果在美国做，就没那么划算，因为美国人力成本贵。

投 Business Messenger 广告，按单个销售线索转化，也就是 CPA 来算，一半成本是广告成本，另外一半甚至更多是背后人力客服的成本。为什么在东南亚能跑通？因为客服人力成本低，所以可以卖客单价比较低的产品，不一定要卖客单价很高的 B2B 线索、轻奢产品或教育产品。

但在美国，人力成本很高，所以销售线索类产品只能卖单价比较贵、ASP 比较高的项目，很多是 B2B 线索。现在的趋势是，如果把背后的人力成本 AI 化，降掉**80%**，总广告成本是不是会下降**40%**？一半成本降价**80%**，总成本就下降**40%**。这**40%**里，除了一部分变成投放公司的利润，大部分会重新 scale 到广告预算里，产品单价也会往上涨。

它的 business model 没有推荐算法那么简单，但稍微绕一下，可以把这个经济模型算出来。对于广告数的变化，它就可以做更下沉的线索：不光做 B2B 线索，也可以做普通商品；原来只能卖**1000美元**的东西，现在也可以卖**100美元**的东西，量就会涨；不光接大商户，也可以接小商户，投放广告的客户群也会增加。

再往后，Meta 在做一套新产品，叫 Business AI。**里面有一个产品，你给它一个 URL，或者给它一张产品图片，它就帮你做素材。**制作素材的成本可以降得更低。

它不是原来的图生图。比如把官网 URL 给它，它就配好素材、配好文案。**很多个人开发者、小软件公司、小 APP 公司，或者美容院之类的商家，对素材要求没有那么高，这对它们很有帮助，所以有可能扩大更多广告主。**

另一方面，AI 降低了广告投放门槛：做素材的门槛、优化广告的门槛，以及背后人力客服的门槛，都能优化。**它比推荐算法稍微间接一点，但从长期看有很大价值。**

Meta 很多广告品类的比重做不过谷歌。比如车和旅游都是谷歌很大的品类，旅游在谷歌广告预算里差不多接近**10%**，车也是大几个点，但在 Meta 里的比重比较低。你很容易想到，如果要旅游、要买车，是不是先去搜索？如果要买车，已经搜了特斯拉，再给他出特斯拉广告，承担的价值可能性最高，那个关键词就应该卖最高价值。

但对于 Meta 来说，你不会在 Meta 里面搜索买什么车，也不会搜索去哪旅游，所以 Meta 缺少这部分线索和用户数据。现在它接入了 search chatbot，search 背后接自己的模型以后，用户可以在里面搜索“我要买什么车”“我想去哪旅游”。那 Meta 就有用户线索和匹配数据，长期可以提高在旅游、汽车里面的广告预算占比。

这个事情不会很快，但是很重要。以前 Meta 的搜索栏里很多用的是 Gemini；现在 MuseSpark、Alvake、Alcado 上了以后，可以用自己的模型，拿到更多线索数据。**所以长期来说，Meta 可以扩展广告品类。**

曹卿云：说到 MuseSpark，它是 Meta 第一个自建的 frontier model，而且 Meta 没有 cloud business，只能通过 AI 投入把自己的产品变现。如果这些模型让 Meta 建立起**35亿日活用户**的数据飞轮，那对 OpenAI 和谷歌意味着什么？

周默：我觉得对 OpenAI 和谷歌不会有特别大的影响，因为 Meta 的模型还是主要定位于自用，虽然以后有可能变成商业模型。

但如果要变成商用模型，最重要的还是能不能跟上 top one 的三家公司，甚至现在 top one 可能是两家公司。谷歌可能落后0.25或0.5个身位，但因为谷歌算力很大，很快也会追上。如果不能到前三家，很难在商业化上有很大优势：既没有开源模型的价格，也没有**一线模型公司**的能力。

所以对 Meta 来说，能够把自己的场景定位好，在自己的场景里有更高性价比，围绕场景做更多、更好的模型改动，就很好了。

你去看 Meta，除了**扎克伯格**以外，还有几位负责产品端的人；包括收购 Manus，也是为了让自己的产品跑得更快。Avocado 这个模型能够在这么快的情况下做到这个水平，已经很不错了。Meta 和另外三家**模型公司**不一样，Meta 的人是从各个地方拉过来的，大家有不同路径、不同想法和不同 know-how。现在把它们重新揉在一起，还不能完全在原来的 LLaMA 体系里面，下一代可能就跟 LLaMA 没关系了。

在几个月时间内重新做一代，做到 Gemini 3 的水平，对 Meta 来说已经非常好了。往后我们对 Meta 的要求可能不是这个模型对外能赚多少收入，而是它能不能最适配推荐算法，提升推荐算法，在 Meta 所有产品里用起来，在 Meta 广告里看到加速，这样就能减少对 CAPEX 的担心。

Meta 这个季度股价暴跌，还是源自 CAPEX 和收入加速错配。收入虽然加速了9个点，但刚刚达到 Meta 指引的高端，而且之前提到有6个点是基数影响，4个点来自汇率、美元指数基数，两个点来自美国大选竞选时广告的基数。所以真正产品收入加速是3个点，刚刚达到高端。Meta 之前每个季度收入一般会比高端多一个点，或者多一个点以上。

这是收入端变化，收入端变化没有那么明显。另一个是 CAPEX 变化，Meta 增加了**100亿美元 CAPEX**，相当于增加**7%的 CAPEX**。

但是这**7%的 CAPEX**对下半年影响很大，因为一季度 CAPEX 比预期低很多，一季度 CAPEX 很小。你要把一季度 miss 的 CAPEX 加到后面三个季度，还要把增加的10个 billion CAPEX 加到后面三个季度。后面三个季度 CAPEX 会越来越高，最后一个季度会比原来对这个季度的假设高很多，肯定不是高**7%**，而是高百分之十几，甚至可能高**20%**。越往后，CAPEX 影响越大。

这么一算，**2027年**的 CAPEX 就比之前假设的**2027年** CAPEX 大很多，对**2027年**利润的影响也更大。**2026年**公司没有调整 OPEX 的变化，OPEX 的变化分两种：好的一部分是公司在裁员，差的一部分是 CAPEX 在提高，CAPEX 提高以后会变成折旧成本。裁员以后，研发、销售和 G&A 成本会下降，互相抵消，所以今年是一个抵消过程。

但 CAPEX 明年提了很多以后，明年还能裁这么大的幅度吗？如果从投资人角度线性推导，就会把明年利润调低。**这是 Meta 反应这么大的原因。**

我自己感觉，Meta 是这一批用 AI 提效最激进的公司。**我相信可能有很多听众是 Meta 工程师。**Meta 的 PIP 考核已经考核代码量了，从员工角度讲，使用 multi-coding-agent、用 AI 做代码，都会有激励，也给大家很大的 coding consumption。

Meta 是最 AI-coding 化的公司，但最 AI-coding 化一方面让每个工程师的有效工作时长缩短很多。原来一天可能有五六个小时有效工作时间，现在如果非常 AI 化，可能只有一两个小时。

另一方面，Meta 也做了很多举措，帮助公司蒸馏一些工程师的 skill。有效工作时间减少、效率提高以后，要么公司做更多产品迭代，要么就会往下一步裁员发展。

我们确实看到 Meta 做了更多产品迭代，很明显看到一些产品开发进度比以前快很多，比如之前提到的 overlay ads 软广，还有 Meta 大大增加了做 DSP 广告、Meta Audience Network 的工程师数量。都能看到 Meta 的产品力和生产力：coding 释放以后，更多市场力在不增加人员的情况下，开发速度比以前快很多；但同样也会看到裁员幅度比以前大很多。

所以如果明年还能有相应裁员，对 CAPEX 的担心也会缓解。**明年可能会看到，虽然 CAPEX 加了很多，但是 OPEX 仍然没有明显增长，这也是有可能的。**

曹卿云：你觉得现在硅谷的裁员节奏大概到哪一步？我们是刚刚开始，还是怎样？

周默：其实第一步，生产力变化的影响，从去年**7月份**就开始了。去年**7月份**看美国数据，第一次就业没有完成目标。我记得当时我们写了一篇报告讨论，提到以后可能每个月都是这样。从去年**7月**开始，已经看得很明显；今年肯定更明显，过去几个月就业大部分也没有达成目标。

从大厂角度看，Meta、英伟达的 coding 渗透率最高、代码量最高。但从整个生态来讲，工程师经历了几个过程。

去年底可能还会说，如果你是 AI-native 公司，有**50%的 incremental coding** 代码量由 AI 做，就很好了。到了今年一月底，对 AI-native 公司的要求是，所有代码都得是 AI 写的。

到了二月份，大家开始大规模使用 Claude 4.6以后，要求更高了：不光要有一个 coding agent，还要把 coding agent 做编排，要有多个 coding agent，甚至5个、10个。

到了3月份，进一步明显，甚至出现了很多 remote coding 的概念。很多工程师会说，这次出去玩可以不带电脑，在手机上面下命令就可以了；以前都要带电脑，随时 on call。甚至大家说，睡前一定要挂一个任务，不然睡觉的8个小时就浪费了。

原来做5个 multi-coding-agent，现在有人可以做100个，甚至2000个。我们公司有一个产品经理一直跟我说，他已经做了14000个了，所以我对到底有多少个能用，一直没有太搞明白。

你会发现，生产力提高的方法是按月度迭代的，每个月生产力衡量的标准都不一样。**生产结构出现变化以后，很难再调整回去。**以后对工程师的评价标准，是你比原来人力 coding programmer 的时候高几倍，这个倍数可以直接按照 coding API 的 consumption 看出来。

再往后，对 coding 工程师的要求会更加复合：不光代码能力强，还要有多线程能力，要不停想需求，迭代速度也要快，才能管理更多 agent，做到那么多 coding consumption。**如果还是原来的单线程工作方式，其实花不了这么多 consumption。**

所以以后对 coding 工程师的标准跟以前不一样了，这是一种新的职业工种。如果出现一种新的职业工种，就一定会出现前后的交替，这可能是不可避免的。

曹卿云：我们观察到，大部分现在 AI 的变现，本质上是效率提升，同样的事情做得更快、更便宜。但真正支持长期估值的，是出现了一个在 AI 出现之前根本不存在的市场。

画面说明

最终关键帧包括 Meta 相关访谈画面和图表：图表标题为“硅谷头部公司：AI Coding 渗透率按月迭代”，可见 AI 代码占比从**20%、45%、70%、85%**升至**95%**，以及 Multi-Agent 使用率从**10%**逐步升至**80%**；另一张关键帧为周默访谈近景，字幕可见“每一个月的生产力衡量的标准都是不一样的”。

11. AI 商业化：替代存量还是创造增量（01:27:20–01:31:29）



场景 11

本节主旨 AI 商业化：替代存量还是创造增量

完整内容

曹卿云：你有没有看到这种系统性的证据，表明 AI 正在创造一个增量价值，而不只是重新分配存量？

周默：今年确实很多增长来自替代人力成本、替代 OPEX，另一部分是替代整个 Enterprise SaaS 的市场规模。

OPEX 是一个很大的数。原来在云行业里，一个企业的 IT spending 一般占 total revenue 的3%到4%，这部分预算给各种 SaaS 公司、各种数据库。能够有10亿美元 IT spending 的公司很少很少，只有苹果或者特别大的公司才能做到。

但是 OPEX 是一个很大的概念，OPEX 占一个公司的收入可能是**30%**，是七倍的概念。**所以如果能够替代人，就是很大的收入增量。**

后面遇到的问题是：替代人以后，这部分人是永久性失业，还是因为 AI 提高工作效率，创造了更多场景需求？**现在确实还没有看到这么多场景需求，但是替代人力的变化已经非常明显。**

现在 coding 已经实现**90%**的 AGI 能力了。Moss 可能已经是第一代部分实现 self-play 的 coding 模型。如果 coding 能够实现**90%**的 AGI 能力，基本上其他领域实现 AGI 可能也就是**2028年或2029年**的事情。**2028年、2029年**就是下一代模型，到了十个千亿规模的模型，可能就已经到了下一代 AGI 范式。

如果从这个角度想，替代 OPEX 的过程从现在起应该会加速。**所以到了明年或后年，确实会有更多社会问题不得不解决，涉及福利、转移支付，或者是不是会有 AI 裁员税、token 税，这些都必须解决。**

涉及新增产品，AI 本身是一个提效产品，和短视频杀时间不一样，AI 不是一个杀时间的产品。所以从创造生态价值来讲，不能指望靠 AI 创造很多需求，可能需要想的是：如果 AI 大大提升工作效率，让所有软件预算变得更高效，解放更多 OPEX 成本，当转移支付能够跑通以后，**剩下这批人不需要完全适应工作**，有更多消费时间；如果转移支付又能让他们有更多可支配收入，工作以外就会出现一个更大的消费级市场。

只不过在解决消费级市场之前，可能要先解决很多再分配问题。**有了更大的消费级市场以后，每个人就有更多时间；有更多时间以后，就会有更多 consumer 产品，可能是互联网产品，也可能是 AI 产品，可能就有更多需求起来。**

我觉得它不是一个非常直接的关系，可能要绕两个弯。

曹卿云：变成一个更间接的关系。那今天最后，我们还是回到 **AI 收入**放到 CAPEX 背景下面来理解。四家公司说过**2026年**合计 CAPEX 超过6300亿，而目前现在可以追踪到的 AI 相关收入——

画面说明

最终关键帧为**周默**在书架前的访谈画面，字幕可见“你和短视频‘杀时间’不一样”；画面没有额外图表、产品界面或姓名条。

12. 结尾：CAPEX 与 AI 收入缺口逐步收敛（01:31:29–01:33:42）



场景 12

本节主旨 结尾：CAPEX 与 AI 收入缺口逐步收敛

完整内容

曹卿云：包括云的 AI 服务、**推理**的 API、AI 订阅，年化大约在1000亿到1500亿美元之间。这是一个大概**4到6倍的投入产出比**。未来你认为收入端和资本支出端中间的差距会怎么样变化？

周默：肯定会随着时间缩小。现在肯定是一个不太好看的数字，但是今年的数字已经比去年好看很多了。

你像今年整个云的 CAPEX，大概**7000亿美元**。然后跟 GPU、AI 相关的，你打个六到七折。再看这几个**模型公司**的 ARR：Anthropic 到年底可能是100 billion，OpenAI 可能是60到70 billion，Gemini 可能是30到40 billion。

这么算合起来其实是一个**2000亿美元的 ARR**，差不多可能占到云和 AI 相关 CAPEX 的三分之二。

但是 CAPEX 的概念是提前做未来四年的投资。CAPEX 如果停了，这些 AI ARR 可能还在增长，因为它是提前投资。所以到了明年，差距会进一步缩小。如果明年这些**模型公司**到明年年底有**4000亿美元 ARR**，那明年这个数字就不是三分之二，也不是一半，可能就到四分之三，看起来会舒服一点。

为什么？你看现在市场的反应也很明确，市场反映的是模型公司的 ARR 在加速，商业模式越来越像一个真的模式。

而去年面临的很多问题是，CAPEX 增速比 AI 收入增速大。如果看 incremental revenue 的贡献，这些 incremental cost 是 incremental revenue 的好几倍。但是现在 incremental cost 相比 incremental revenue 的比重越来越接近了。所以今年反映的是，对这个问题的 concern 越来越小。

画面说明

最终关键帧为周默访谈近景，画面底部可见节目章节条，未显示新的图表、姓名条或产品界面。