Herman AI Talk — AI Infra 狂热后的风险反思

> Disclaimer：本文是对 Herman 观点的整理与结构化复盘，不构成投资建议；其中涉及的市场判断、公司判断、收入数字、CapEx 数字等，除非另行核验，均应视为 Herman 当时分享中的观点或转录内容。

0. TL;DR

- Herman 的核心语气不是“AI 不行了”，而是：**市场已经从早期机会进入高度拥挤阶段，很多人赚了很多钱，现在该静下来反思仓位和风险**。他提到从较早参与 AI / 半导体投资的人来看，不少账户已经有非常高的收益，这时继续 all in 的边际风险开始变大。
- **不能轻易做空。** 即使你看出 AI infra 有泡沫、有风险，也不代表能靠做空赚钱。Herman 用 2007/2008 类比：很多人可能在 2007 年就看出问题，但真正的大崩溃在 2008/2009，过早做空的人未必能活到兑现时刻。
- 当前 AI 主线有一个“**铜墙铁壁 / 金钟罩**”：只要市场相信 **OpenAI / Anthropic 的收入增长能够 justify hyperscaler CapEx**，那么 Blackwell delay、交换机过热、光学链路 delay、油价、通胀、利率、CTA 卖出等外围负面，都很难真正打穿主线。
- Herman 反复强调：**每个底层公司都真实缺货，订单是真的，涨价是真的，盈利也是真的；但高层 thesis 仍然有风险。** 不能因为微观细节都对，就忽略整个系统最上层的资金回报闭环。
- 他的一个强表达是：**“低 PE 的泡沫不代表它不是泡沫。”** 低 PE 反而可能让人每次下跌都敢加仓、敢加杠杆，因为看上去盈利真实、估值便宜、缺货持续。但一旦 PE 背后的盈利预期被打破，逃跑可能很难。
- 流动性是当下利好，也是未来风险。Herman 认为 eSLR / deregulation 释放银行可购买资产，类似一次性“开闸放水”；但不同于持续 QE，流动性快感会递减，像多巴胺一样，需要越来越多刺激才能维持同样的市场反应。
- 半导体供应链正在从原来 TSMC / ASML / Apple / Qualcomm / MediaTek 这类“计划经济式分蛋糕”的秩序，变成 AI 需求拉动下的“**群雄并起 / every dog has its day**”：存储、MLCC、CPU、光模块、晶圆代工、GFS、中国成熟制程等都可能涨价。
- “群雄并起”短期是利好，因为每个环节都可能出现瓶颈、涨价、盈利弹性；但它同时会推高 hyperscaler 的总 CapEx，伤害最终 ROI，使主线更依赖模型收入继续高速增长。
- Herman 说现在整个系统走在一条“**窄钢丝**”上：市场预期 OpenAI / Anthropic 收入继续高速增长，从而让 hyperscalers 愿意继续扩大 CapEx。如果收入增速只是“加速度下降”，不一定要负增长，就可能让市场震动。
- Herman 特别关注 **Anthropic / Claude**：因为 Anthropic 今年收入增势很猛，某种意义上是 AI 资本开支故事里的“主帅”之一。但他观察到 Claude 在高调用压力下模型质量下降、编程体验变差、调用次数增加但满意度下降，这可能反过来影响需求增速。
- **模型商品化** 是真正危险的叙事。如果 Kimi / 中国模型 / 其他低成本模型能以更低价格提供接近的效果，就会动摇 OpenAI / Anthropic “值得持续烧大量算力和训练成本”的梦。AI infra 的中间需要一个“模型越来越强”的梦来连接半导体 CapEx 和未来生产力。
- 操作启发不是“立刻做空”，而是：赚到很多钱后，要从底层公司细节往上拔一层，看收入、CapEx、流动性、杠杆、模型质量、供应链涨价是否还能闭环，再决定仓位和风险管理。

1. 当前语境：为什么现在要反思仓位

Herman 开场的基调很明确：**现在市场很疯狂，涨得很多，AI 已经从一个早期少数人理解的主线，变成了所有人都在谈的拥挤共识。**

他提到，早期一起参与 AI / 半导体投资的人，很多已经赚了非常多。重点不在具体倍数，而在这个状态：

> 当一个方向已经带来巨大收益，投资人容易把最初的 thesis 变成信仰；仓位、杠杆、风险承受度都会悄悄发生变化。

这也是他认为需要“静下来反思”的原因。不是因为他否定 AI 长期，而是因为：

1. **市场共识已经高度一致。**
   “AI is everything”“all in AI” 变成流行语言后，原来的信息优势和仓位优势都在下降。
2. **半导体研究已经全民化。**
   他提到现在市场上“无数半导体专家”，很多人都能讲某个细分环节的缺货、涨价、订单、产能瓶颈。这说明主线已经被充分挖掘。
3. **底层信息越看越真实，反而容易忽略高层风险。**
   越钻进单个公司、单个 sector，越会看到真实订单、真实缺货、真实涨价、真实盈利，于是很难主动撤退。但 Herman 认为，真正的风险不一定在底层公司，而在更上层的 CapEx 与模型收入闭环。
4. **仓位和杠杆比观点本身更危险。**
   当大家都赚钱、都相信、都加仓时，风险不是“有没有长期价值”，而是“市场是否已经把太多未来回报提前定价”。

所以这段 talk 的核心不是“AI 泡沫马上破”，而是：**在高收益之后，从 bottom-up 的公司细节中抽离出来，重新看 top-down 的主线是否仍然稳固。**

2. 不要轻易做空：即使看出风险，也可能空在 2007

Herman 第一件事就强调：**不能去做空。**

他的逻辑不是说市场没有风险，而是说“看出风险”和“靠做空赚钱”中间隔着非常远的距离。他用 2007 / 2008 做类比：

- 2007 年时，部分人已经看出美国金融系统和房地产泡沫有问题；
- 但真正的大危机在 2008 / 2009 才全面爆发；
- 如果过早做空，可能在泡沫最后一段上涨中先被挤爆；
- 很多做空的人“没有活到”风险兑现的时刻。

这和 AI infra 当前状态的类比在于：

> 即使你认为 AI infra 存在泡沫、CapEx 有过度扩张、半导体涨价不可持续，也不代表你可以现在就空 NVIDIA、空半导体、空光模块、空存储，甚至空整个 AI trade。

原因有几个。

第一，**主线仍然有真实基本面支撑。**
每个底层公司确实在缺货，很多环节确实在涨价，订单也确实存在。做空一个基本面还在加速、叙事还在强化、流动性还宽松的方向，本身就是逆势。

第二，**市场可以在“明显贵”之后继续贵很久。**
泡沫最危险的地方不是它完全没有盈利，而是它有盈利、有增长、有故事，因此每一次回调都被买入。

第三，**做空需要非常特殊的工具和时点。**
Herman 提到，如果要 hedge 或做空，可能需要寻找“杠杆极高、面对大跌有极高凸性”的产品。换句话说，普通线性做空或普通 put 很可能会在时间价值、波动率、挤仓中被消耗掉。

因此，他的建议更接近：

> 不要把“风险识别”直接翻译成“立刻做空”。更合理的动作是管理仓位、降低杠杆、寻找对冲结构、观察核心指标，等真正命门出现裂缝。

3. AI 主线的“金钟罩”：OpenAI / Anthropic 收入增长能否 justify hyperscaler CapEx

Herman 认为，当前 AI 交易真正的主线不是单一芯片、单一光模块、单一存储公司，而是一个更高层的闭环：

> **OpenAI / Anthropic 等模型公司的收入增长足够快，因此能够 justify Microsoft / Amazon / Google / Meta / Oracle 等 hyperscalers 的 CapEx；hyperscalers 的 CapEx 又带动 NVIDIA、存储、光模块、晶圆代工、MLCC、PCB、服务器等整个 AI infra 链条。**

这就是他所说的“命脉”。

只要这条命脉不被打穿，很多负面信息都只是扰动，而不是主线破坏。Herman 举了很多例子：

- SVB 出问题；
- 关税战；
- 伊朗危机；
- CTA 卖出；
- 2024 年 NVIDIA Blackwell / 液冷相关问题；
- 交换机可能过热；
- 光学链路可能 delay；
- 某一代卡可能 delay；
- 油价、通胀、利率、QT、降息节奏变化。

这些事情当然会影响市场情绪，也可能造成短期下跌。但在 Herman 看来，只要市场仍然相信 AI 需求持续强、模型收入能支撑 CapEx，这些外围风险很难形成真正的趋势反转。

他所谓“金钟罩 / 铜墙铁壁”的意思是：

> 市场已经形成一个强信念：AI 模型收入增长会足够大，足以解释今天看似夸张的算力投入。这个信念像护城河一样挡住了很多宏观和技术扰动。

但这也带来反向问题：**如果真正打到这条主线，市场会非常脆弱。**

Herman 特别强调，从 bottom-up 看，每个环节都是真实的：

- 每个公司都有订单；
- 每个缺货都是真的；
- 每个涨价都有现实原因；
- 产能扩张需要时间；
- 盈利也在兑现。

但这些真实的局部，不等于系统整体没有风险。系统真正需要回答的是：

> 模型公司收入增长，能不能长期支撑 hyperscaler CapEx?
> 如果 CapEx 继续扩大，ROI 是否还能被华尔街认可?
> 如果供应链每个环节都涨价，最终是否会反过来压垮主线?

4. 风险一：低 PE 泡沫

Herman 的一个核心表达是：

> **低 PE 的泡沫，不代表它不是泡沫。**

这句话很重要，因为很多人理解泡沫时，会默认泡沫等于高 PE、高 PS、没盈利、讲故事。但 Herman 认为，AI infra 这轮的危险恰恰在于：很多公司有真实盈利，甚至 PE 看起来不高。

低 PE 泡沫的机制是：

1. 公司盈利真实增长；
2. 估值看上去不贵；
3. 每次下跌都像“给机会”；
4. 投资人敢于买跌；
5. 因为基本面强，甚至敢于加杠杆；
6. 股价继续上涨，形成正反馈；
7. 更多资金进入，仓位越来越拥挤。

Herman 用了类似 P2P 高息揽存的比喻：
每天都给你正反馈，每天看到利息，每天看到业绩，每天看到缺货，每天看到涨价，于是你越来越相信它不会出问题。

4.1 韩国 / SK Hynix 杠杆例子

Herman 提到，韩国市场上有人大规模加杠杆买 Samsung / SK Hynix，甚至有原来币圈社群转去高杠杆买 SK Hynix。转录中提到“五倍杠杆”“抵押房子”等说法，这些属于 Herman 听到的市场现象，应作为待核验 anecdotal evidence，而不是已证实事实。

但这个例子的意义不在于具体杠杆倍数，而在于说明：

> 当一个低 PE + 高盈利 + 强缺货的资产持续上涨时，它会吸引越来越激进的杠杆资金。

4.2 华强北 NAND / 存储涨价

Herman 提到华强北 NAND / 存储价格短期连续上涨，转录中出现“过去三天涨约 5%”“前一周又涨约 5%”这类数字。这些数字需要单独核验。

他想表达的是，存储价格上涨不是空穴来风，而是现货市场、协议价、供需紧张共同作用的结果。存储具有商品属性，小规模现货市场的价格波动可能影响更大规模的协议价谈判，有点类似铁矿石。

4.3 晶圆代工涨价与成熟制程回流

Herman 还提到，因为 TSMC / GlobalFoundries 等成熟或相对低端制程产能紧张，一些客户被挤出，回到中国大陆成熟制程产能。转录中出现“原来 25 美元，现在 50 美元”“涨价一倍”等说法，也需要核验。

这里的核心含义是：

> AI 需求不仅拉动最先进制程，也会通过产能挤压传导到成熟制程、晶圆代工、光罩 / wafer 相关环节，导致原来不赚钱的产能也可能变得赚钱。

4.4 低 PE 幻象为什么危险

Herman 认为，低 PE 泡沫真正危险的地方是：**一旦 PE 背后的 E 被证明不可持续，市场不会给你慢慢撤退的机会。**

他用了类似“P 是幻象”的表达。更准确地说，这里的 P/E 幻象不是价格 P 幻象，而是盈利 E 的可持续性和终局假设可能是幻象：

- 如果订单来自短期抢产能；
- 如果涨价来自供应链挤兑；
- 如果 CapEx 最终不能被模型收入 justify；
- 如果模型收入增速放缓；
- 如果 hyperscaler 降低 CapEx；
- 那么今天看起来低的 PE，可能突然变成高 PE。

所以，低 PE 不一定是安全垫。它也可能是泡沫最有效的麻醉剂。

5. 风险二：流动性洪水与快感递减

Herman 的第二个风险是流动性。他说的有点反直觉：**他讲的风险，很多都是当下的利好。**

当前市场的流动性利好，主要来自：

- eSLR / SLR 规则调整预期；
- deregulation / 去监管；
- 银行资本金要求放松；
- 银行可购买资产空间增加；
- 银行间流动性宽松；
- SOFR 等利率指标偏低；
- bank reserves 较高。

转录中提到“一次性放出四万多亿 / 四万五千亿银行可购买资产”之类数字，应视为 Herman 的口径，需另行核验。

5.1 和 2008 后 QE 的区别

Herman 区分了几种流动性：

2008 后 QE：
金融系统去杠杆，监管加强，央行资产负债表扩张，相当于央行印钱来对冲金融体系杠杆下降。

疫情期间 QE：
每个月持续购买资产，像“一条河流”不断流入池塘。转录中提到疫情期间 QE 约每月 1200 亿美元，这个数字也需要核验，但比喻很清楚：持续水龙头不断注水。

当前 eSLR / deregulation：
更像一次性开闸放水，把原来受监管限制的银行资产购买能力释放出来。它的冲击可能很猛，但未必是持续每月新增的水龙头。

5.2 多巴胺类比：流动性快感会递减

Herman 用“多巴胺”类比市场对流动性的反应。他的意思是：

- 流动性进入市场，会带来上涨和风险偏好提升；

- 但市场很快会把它变成新的基准预期；
- 要维持同样快感，需要更多、更大的刺激；
- 如果没有新增刺激，流动性带来的边际推动会减弱。

这和毒品、烟草、短视频刺激类似：第一次很强，后面需要更大剂量才能达到同样反应。

所以 Herman 并不是说流动性不重要。他的意思是：

> 流动性是当前 AI / 半导体上涨的重要助推器，但它不是无限可重复的发动机。一次性释放后，市场对它的反应可能逐渐钝化。

5.3 需要观察的流动性指标

指标	观察意义
SOFR	银行间美元融资是否宽松；若持续低于相关基准，说明短端流动性较宽
Bank reserves	银行体系准备金水平，反映系统流动性余量
Treasury issuance / 发债吸收流动性	大量发债可能抽走流动性，但若市场仍宽松，说明资金面仍强
eSLR / SLR 政策进展	资本金约束是否放松，银行能否扩大资产负债表
银行可购买资产规模	若确实释放大量购买能力，短期利好风险资产
美联储 QT / 降息节奏	如果通胀、油价或利率扰动影响降息预期，可能削弱流动性叙事

6. 风险三：半导体供应链从秩序市场变成“群雄并起”

Herman 对半导体产业链有一个很形象的判断：过去的半导体更像一个高度垄断、有秩序、甚至“黑社会式”的卖方市场。

这里的“黑社会”不是道德评价，而是指：

- 产能高度集中；
- 技术门槛极高；
- 供应链头部掌握分配权；
- 每年像“分蛋糕”一样安排产能；
- 客户之间有明确优先级。

6.1 原来的秩序：计划经济式分蛋糕

在消费电子时代，半导体产业链有较稳定的秩序：

- ASML 决定先进光刻机供给；
- TSMC 决定先进制程产能；
- Apple 作为最大客户优先拿最先进节点；
- 之后再轮到 Qualcomm、MediaTek 等；
- 每个节点、每个客户、每年产能怎么分，都比较有秩序。

Herman 的说法是，这更像一种“计划经济”。需求大致可预测，产能分配有层级，龙头有定价权，产业链虽然紧张，但秩序仍在。

6.2 AI 并行计算需求打破旧秩序

AI 需求不同于传统消费电子。消费电子需求是周期性的、终端驱动的、相对可预测的；AI 需求则来自并行计算、token 调用、模型训练和推理，增长速度更快，结构更复杂。

Herman 用“老车往前猛开”的比喻：

- 不是只有发动机会坏；
- 轮子、刹车片、传动系统都可能出问题；
- 整个系统从未经历过这种量级的拉动；
- 于是到处都变成瓶颈。

这就是 “every dog has its day”：

> 每个环节都有出头日，因为每个环节都可能突然缺货、涨价、获得议价权。

6.3 Every dog has its day：哪些环节会被拉动

Herman 在 talk 里提到或隐含了很多可能“出头”的环节：

环节	Herman 逻辑
存储 / NAND / DDR / HBM	AI 服务器和数据中心需求拉动存储；现货和协议价可能互相强化
MLCC / 陶瓷电容	高性能服务器、GPU 板卡、供电系统复杂度上升，带动被动元件需求
CPU	AI 服务器不是只有 GPU，CPU、host、系统配套也可能涨价
光模块	数据中心内部和集群间通信需求上升，光模块供应可能被卡住
TSMC	先进制程和先进封装核心瓶颈，仍有强定价权
GlobalFoundries / GFS	成熟制程被挤压后也可能获得涨价空间
中国成熟制程	外部产能紧张时，部分需求回流，原来追赶型、不赚钱产能可能变得有盈利弹性
Wafer / 光罩 / 代工加工	被产能挤压后价格可能大幅上行
服务器 / 电源 / 散热 / PCB	AI rack-scale 复杂度上升，配套环节也可能成为局部瓶颈

6.4 为什么这既是利好，也是风险

短期看，“群雄并起”是强利好：

- 越多环节缺货，越多公司有业绩；
- 越多环节涨价，越多股票有弹性；
- 越多瓶颈出现，AI infra trade 越能扩散；
- 原来不被市场关注的小环节，也可能被重新定价。

但从系统层面看，它也是风险源：

> 如果每个环节都涨价，那么 hyperscaler 的 CapEx 会越来越高；如果 CapEx 被推高，而模型收入没有同步上修，整个 ROI 闭环就会变差。

这就是 Herman 反复回到的命门：**CapEx 与模型收入必须匹配。**

7. 风险四：CapEx 与收入之间的窄钢丝

Herman 说，现在市场走在一条非常窄的钢丝上：

> **justifiable CapEx** 必须对应 **可被验证的收入增长**。

市场不是简单相信“AI 很强”，而是相信：

- OpenAI 收入继续高速增长；
- Anthropic 收入继续高速增长；
- hyperscalers 的 CapEx 会被模型需求吸收；
- 数据中心投资能在未来几年赚回来；
- 华尔街愿意认可借债、发股、自由现金流转负之后继续投入。

7.1 Token 需求指数增长 vs 半导体产能线性增长

Herman 的长期牛市逻辑本来是：

- token 需求是指数级增长；
- 半导体产能扩张是线性的；
- 线性供给追不上指数需求；
- 所以整个半导体会长期紧缺，价格和盈利被推上去。

他说这个逻辑在过去两年已经被验证：AI 需求确实把半导体推到了非常高的位置。

但他现在提出的问题是：

> token 需求曲线真的完全不会回调吗?
> 需求曲线和模型质量有没有关系?
> 如果模型质量下降，需求增速会不会放缓?

这是关键变化。

7.2 Hyperscaler CapEx 压力

Herman 提到，今年 hyperscaler CapEx 规模已经非常大，转录中出现“七千五百亿到七千七百亿”“明年搞到一万亿”这类数字。由于 ASR 可能有误，且口径可能是总量、年化、美元或人民币表达，需单独核验。

但方向很清楚：

- 今年 hyperscalers 已经投入巨大；
- free cash flow 可能承压甚至转负；
- 继续投入需要借债或发股；
- 明年若进一步扩大 CapEx，就需要 OpenAI / Anthropic 收入继续暴涨来支撑市场信心。

这意味着市场容错率越来越低。

7.3 Oracle / OpenAI backlog / CDS 例子

Herman 用 Oracle 举例说明市场容错率很低。

他的说法大意是：

- 去年某个阶段，OpenAI 模型相对优势受到挑战；
- 市场开始不认可 OpenAI 给 Oracle 的 backlog；
- Oracle 的 CDS 曾经大幅上行，转录中提到“五百多个点”，需核验；
- Oracle 股价受到明显压力，后来即使 OpenAI 重新变强，相关资产也未必完全修复。

这个例子说明：

> 市场不是无条件相信 AI CapEx。只要模型质量、收入兑现、backlog 质量出现怀疑，信用市场和股票市场会迅速重估。

7.4 加速度下降也可能足够危险

Herman 强调，不一定需要 OpenAI / Anthropic 收入下降，甚至不一定需要增速变为零。只要：

- 收入继续增长，但增长加速度下降；
- 月度收入增速开始变慢；
- 市场发现模型质量问题影响需求；
- hyperscaler 对下一年 CapEx 指引变谨慎；

就可能足以让市场震动。

这就是“窄钢丝”的含义：

> 当前估值和仓位隐含的是非常高的收入增长连续性。只要增长曲线从指数型变得没那么陡，市场就可能重新计算整个 AI infra 链条的估值。

8. Anthropic / Claude：Herman 认为的关键命门

Herman 最后把命门集中到模型公司，尤其是 Anthropic / Claude。

8.1 为什么 Anthropic 是关键观察点

Herman 认为，Anthropic 今年收入增长势头非常猛，是 AI infra 主线里的重要“主帅”之一。转录中提到：

- 去年某阶段市场预期 Anthropic 年收入约 90 亿或 100 亿；
- 今年二三月份预期年底约 300 亿；
- 五月中传出数据约 550 亿。

这些数字高度依赖转录和口径，必须核验。尤其要确认：

- 是 ARR、年化收入、run-rate revenue，还是实际年度收入；
- 单位是美元、人民币，还是 ASR 识别错误；
- 来源是媒体、二级市场传闻、融资材料，还是公开披露。

但 Herman 想表达的是：

> Anthropic 的收入预期上修速度很快，因此它成为市场相信 hyperscaler CapEx 能继续扩张的重要证据。

如果 Anthropic 的收入增速开始变慢，市场就会质疑：AI 模型需求是否真的能持续指数增长？

8.2 Claude 质量下降与调用量上升

Herman 分享了自己使用 Claude / AI 编程的体验。他认为 Claude 在二三月份时效果非常强，用于代码 audit 几乎“完美无瑕”；但后面开始明显出错，现在每个模块都会出现严重错误。

这里要注意，这属于 Herman 的个人使用体验，不是系统 benchmark。

但他提出的机制很重要：

- Claude 好用，所以用户增加；
- 用户调用增加，系统资源紧张；
- 供应不足导致模型服务质量下降；
- 质量下降后，用户为了达到同样结果，不得不反复调用；
- token 用量反而上升；
- 但用户满意度下降；
- 最后可能导致需求增速放缓。

这个逻辑很像一个工厂：

> AI 不是普通软件。普通软件下载安装后边际成本很低；AI 更像一个工厂，每次使用都在向工厂下订单。订单暴增时，如果产能不够，交付质量可能下降。

市场现在可能只看到 token 用量增长，却没有充分定价“质量下降会不会反噬需求”。

8.3 “调用更多”不一定等于“需求更健康”

这是 Herman 特别敏锐的点。

如果用户因为模型更差而反复调用，那么短期 token 数量会上升，收入也可能上升。但这不是健康增长，而可能是“低效率调用”：

- 单次任务需要更多 token；
- 用户为了纠错来回调用；
- 成本和收入一起上升；
- 用户满意度下降；
- 未来可能转向其他模型或减少使用。

因此，观察 AI 公司时不能只看调用量，还要看：

指标	为什么重要
月度收入增速	直接决定 CapEx thesis 是否能持续
收入增速的加速度	市场定价的是高速上修，放缓也可能是负面
用户留存	判断调用是否来自真实满意需求
单任务 token 消耗	区分高价值使用与低效率纠错
模型质量 / coding benchmark / 实际用户反馈	判断需求是否可持续
推理毛利率	token 增长是否带来真实利润
价格变化	是否靠降价维持用量
API 客户增长	企业需求是否持续扩散

8.4 模型商品化风险

Herman 提到“模型商品化”这个词，并提到 Oracle CEO Larry 相关访谈。这里应理解为：当模型之间差距缩小，训练边际收益下降，模型能力更接近 commodity，模型公司就更难证明自己值得持续巨额烧钱。

模型商品化的风险在于：

- 如果 Kimi / 中国模型 / 开源模型 / 蒸馏模型能以更低成本提供接近效果；
- 如果用户发现 Anthropic / OpenAI 的服务质量在高负载下并不稳定；
- 如果模型领先优势缩小；
- 那么高价订阅、高价 API、巨额训练和推理投入就会被质疑。

Herman 的关键问题是：

> 如果中国模型公司用十分之一成本做出差不多的效果，为什么用户还要支付那么多钱给 OpenAI / Anthropic？为什么华尔街还要认可 hyperscaler 继续投入更多 CapEx？

8.5 “多投半导体就能解决问题”为什么仍有金融约束

Herman 提到一个对冲基金朋友的反驳：所有问题都可以用增加半导体投入解决。只要算力更多，模型质量、供应不足、调用拥堵都能改善。

Herman 承认技术上这个方向可能对，但他反问：

> 钱呢？谁出钱？华尔街认不认账？

这句话是整段 talk 的核心之一。

在工程视角里，答案可能是“加 GPU、加数据中心、加推理产能”。
但在金融市场视角里，答案必须包括：

- 谁承担 CapEx？
- 资产负债表能不能承受？
- free cash flow 会不会转负？
- 借债成本多少？
- 股东愿不愿意稀释？
- 华尔街是否相信未来收入能覆盖今天投入？
- 如果投入从一万亿变一万五千亿，市场是更兴奋，还是开始恐慌？

所以 Herman 不是否定 AI 能靠更多算力变强，而是强调：

> 金融市场定价的是“投入—收入—利润—现金流”的闭环，不是单纯的技术可行性。

9. 长期判断：AI 是集中权力，而不是分散权力

Herman 最后提出一个更长期的判断：**AI 最终是集中权力，而不是分散权力。**

早期 AI 看起来像是在赋能个人：

- 小团队可以用 AI 写代码；
- 个人投资者可以用 AI 做研究；

- 小机构可以借 AI 提升生产力；
- 独立开发者可以更快构建产品；
- Herman 自己也提到，使用 AI 让一些程序和工具能力接近 Jump / Jane Street 等顶级机构的某些效果。

但他认为，这可能只是早期阶段。

长期看，真正能支付最高 token 成本、训练成本、推理成本、数据成本的人，是：

- 大公司；
- 大型金融机构；
- 政府；
- 军事 / 情报 / 主权级机构；
- 高支付能力企业客户。

这些主体能用更高价格购买更强模型、更低延迟、更高稳定性、更大上下文、更专属算力。普通个人和小机构在早期获得的效率提升，长期可能又被更强支付能力的机构重新拉开。

Herman 举了一个量化交易的例子：

- 小团队用 AI 可以把很多程序能力提升到接近顶级机构；
- 但顶级机构可以花更多钱买更强 AI；
- 他们愿意花几百万美元替代几个人；
- 小团队未必付得起同等规模的 token / AI 成本。

这意味着：

> AI 的长期形态可能不是“所有人变得一样强”，而是“最有资源的人获得更强的智慧杠杆”。

这对投资含义是：下一轮牛市不一定只看半导体，也要看谁能掌握模型、算力、数据、分发、企业客户和支付能力。

10. 投资操作启发：应该观察什么，而不是立刻做什么

Herman 的操作启发可以概括为一句话：

> **不要轻易做空，但要认真管理仓位；不要只盯底层公司，而要观察主线命门。**

10.1 不要轻易做空

原因前面已经讲过：

- 需求仍强；
- 缺货真实；
- 盈利真实；
- 流动性仍宽；
- 拥挤交易可以更拥挤；
- 泡沫可能持续很久；
- 线性做空容易死在最后一涨。

所以更合理的是：降低杠杆、控制仓位、保留现金、选择性 hedge，而不是直接大空。

10.2 仓位管理：从底层细节拔到高层风险

Herman 反复强调：越看底层，越乐观。

因为你看任何一个具体公司，都能看到：

- 订单；
- 产能不足；
- 涨价；
- 客户催货；
- 毛利率改善；
- 利润上修；
- 分析师调高目标价。

但投资人赚了很多钱后，应该把视角往上拔：

- 模型公司收入能不能继续高增?
- hyperscaler CapEx 是否会继续上修?
- 供应链涨价是否侵蚀 ROI?
- 流动性快感是否递减?
- 杠杆资金是否过度拥挤?
- 模型质量是否影响真实需求?
- 华尔街是否还愿意给 AI infra 故事买单?

10.3 核心观察指标

类别	指标	Herman 框架下的意义
模型收入	OpenAI / Anthropic 月度收入增速	最核心命门，决定 CapEx 是否能被 justify
模型收入	收入增速的加速度	即使仍增长，只要加速度下降也可能冲击市场
模型质量	Claude / OpenAI / Kimi coding 体验、benchmark、用户反馈	判断模型商品化和需求放缓风险
用户行为	调用量 vs 满意度 vs 留存	区分健康需求和因出错导致的反复调用
Hyperscaler CapEx	Microsoft / Amazon / Google / Meta / Oracle CapEx 指引	判断 AI infra 订单能否延续
现金流	FCF、债务融资、发股、信用利差、CDS	判断华尔街是否继续认可投入
供应链价格	存储、MLCC、wafer、光模块、CPU、PCB 等涨价	短期利好供应链，长期可能推高 CapEx
流动性	SOFR、reserves、SLR/eSLR、银行资产购买能力	判断风险偏好是否仍有水推动
市场结构	杠杆 ETF、融资余额、散户/币圈资金迁移、拥挤度	判断低 PE 泡沫是否进入危险阶段
技术扰动	Blackwell、交换机、光学链路、液冷等 delay	单独看未必致命，但若影响 CapEx ROI，会放大

10.4 哪些信号支持继续持有

在 Herman 框架下，如果出现以下信号，AI infra 主线仍较稳：

- OpenAI / Anthropic 月度收入继续上修；
- Claude / OpenAI 模型质量维持或改善；
- 企业 API 客户和真实付费使用增加；
- hyperscaler 继续明确上调 CapEx；
- Oracle / Microsoft / Amazon / Google / Meta 对 ROI 表达更有信心；
- 信用市场没有明显恶化；
- 供应链涨价没有明显压垮下游预算；
- 流动性仍宽松，且没有明显边际恶化。

10.5 哪些信号提示降低仓位

需要更谨慎的信号包括：

- Anthropic / OpenAI 收入增速低于市场预期；
- 增速仍高但加速度明显下降；
- Claude / OpenAI 质量被开发者广泛吐槽，且影响付费意愿；
- 低成本模型明显替代高价模型；
- hyperscaler CapEx 指引不再上修，甚至开始强调 ROI / discipline；
- Oracle / AI data center 相关信用利差扩大；
- 存储、MLCC、光模块等涨价过快，开始伤害整机 / 数据中心 economics；
- 杠杆资金极端拥挤；
- eSLR / deregulation 利好兑现后，流动性指标边际转弱。

10.6 哪些情况下等待大回撤后再利用 AI 确定性

Herman 不是长期否定 AI。他甚至认为最终会涨回来，因为模型差距、算力需求、半导体产能扩张都是真实趋势。

所以一种更合理的策略是：

> 不在狂热阶段继续无限加杠杆，而是保留资本，在主线因模型收入、CapEx、流动性或信用冲击出现大回撤时，再利用 AI 长期确定性重新布局。

11. 术语与公司速查表

术语 / 公司	在这段 talk 中承担的角色
Anthropic	Herman 认为的核心命门之一；收入增速和 Claude 质量直接影响 AI CapEx 故事
Claude	Anthropic 的模型产品；Herman 观察到调用压力下质量下降，可能影响用户需求
OpenAI	AI 主线的核心模型公司之一；其收入增长是 hyperscaler CapEx 的关键 justification
ChatGPT / GPT	OpenAI 模型产品；与 Claude、Kimi 等一起构成模型质量比较对象
Hyperscalers	Microsoft、Amazon、Google、Meta、Oracle 等云和数据中心巨头；AI CapEx 的主要出资方
CapEx	资本开支；AI infra 投资链条的核心变量，需要被模型收入 justify
Justifiable CapEx	Herman 的核心框架：资本开支必须能由未来收入、利润、现金流解释
SOFR	短端美元融资利率指标；用于观察银行间流动性是否宽松

术语 / 公司	在这段 talk 中承担的角色
eSLR / SLR	银行杠杆率 / 补充杠杆率相关监管指标；放松可能释放银行资产购买能力
Deregulation	去监管；Herman 认为它带来一次性流动性释放
Bank reserves	银行准备金；衡量银行体系流动性余量
QE	量化宽松；Herman 用它对比当前一次性开闸式流动性释放
QT	量化紧缩；如果通胀或利率扰动导致 QT / 降息节奏变化，会影响市场
TSMC	半导体旧秩序中的核心分配者；先进制程和产能分配中心
ASML	光刻机供应源头，决定先进制程产能上限的重要环节
Apple	原来消费电子秩序中优先拿先进制程的头部客户
Qualcomm	Apple 之后的重要移动芯片客户，体现 TSMC 产能分配秩序
MediaTek	与 Qualcomm 类似，是旧消费电子半导体秩序中的重要客户
GlobalFoundries / GFS	成熟制程代工厂；Herman 认为在产能紧张时也可能涨价
NVIDIA	AI GPU 主线核心公司；Blackwell、交换机、液冷等技术扰动会影响市场情绪
Blackwell	NVIDIA 新一代 AI GPU 平台；delay 或技术问题是市场关注点，但不一定打穿主线
Oracle	AI data center / cloud CapEx 重要参与者；Herman 用其 backlog 和 CDS 说明市场容错低
Microsoft	Hyperscaler 之一，OpenAI 生态的重要支持者和 AI CapEx 主体
Amazon / AWS	Hyperscaler 之一，AI 云和数据中心 CapEx 主体
Google	Hyperscaler 之一，拥有云、TPU、模型和数据中心投入
Meta	Hyperscaler 之一，AI 训练和推理 CapEx 重要出资方
Kimi	Herman 用来比较 Claude 编程效果和低成本模型竞争的中国模型代表
SK Hynix	存储 / HBM 主线代表；Herman 提到韩国资金加杠杆买入现象
Samsung	存储和半导体巨头；与 SK Hynix 一起代表韩国存储交易
NAND	存储商品之一；Herman 提到华强北现货涨价带来的正反馈
DDR	存储类型；AI / 服务器 / 消费电子需求变化都会影响其价格
HBM	AI GPU 关键高带宽存储；虽然转录中不是主讲重点，但属于 AI infra 存储核心环节
Wafer	晶圆；代工产能紧张和加工涨价的基础单位
光罩 / Photomask	晶圆制造相关成本项；Herman 提到成熟制程加工价格变化
MLCC	多层陶瓷电容；AI 服务器复杂度上升带来的潜在涨价环节
Optical modules	光模块；数据中心通信需求增长带动，供应受限时具有涨价潜力
AAOI	光通信相关公司；Herman 用作 AI 光链条上涨扩散例子
Lumentum	光器件 / 光通信公司；Herman 提到光模块产能和涨价可能性

术语 / 公司	在这段 talk 中承担的角色
AMD	Herman 提到过去买过的大仓位半导体标的；说明当主线强时买哪张可能都涨
Arm	半导体 IP 公司；Herman 用来说明 AI beta 扩散时产品兑现不一定是短期股价核心
“低 PE 泡沫”	Herman 的核心风险表达：低估值表象可能强化买跌和杠杆
“铜墙铁壁 / 金钟罩”	市场相信模型收入能支撑 CapEx，因而抵御外围负面
“门外的野蛮人”	高杠杆、宏观风险、流动性反噬等暂时被挡在主线外的风险
“群雄并起”	AI 需求打破半导体旧秩序，每个环节都可能涨价
“窄钢丝”	CapEx 和模型收入之间容错极低的平衡
“模型商品化”	如果模型差距缩小、低成本替代出现，OpenAI / Anthropic 的高投入故事会被质疑

12. 待验证数据点 / 后续研究问题

以下数据或判断来自 Herman 分享或 ASR 转录，很多需要单独核验。整理时不应直接当作事实引用。

待验证点	原话含义	为什么重要	建议核验来源 / 方式
早期 AI / 半导体投资者“十倍以上”收益	Herman 说较早一起投的人很多收益很高	判断市场拥挤度和仓位风险	身边样本不可外推；可对比 AI 半导体核心标的 2023-2026 涨幅
韩国资金高杠杆买 SK Hynix / Samsung	转录提到五倍杠杆、抵押房子等现象	衡量存储 trade 是否散户化、杠杆化	韩国券商融资数据、杠杆 ETF 规模、媒体报道
香港 7709 / 两倍杠杆 Hynix ETF 规模	Herman 提到相关 ETF 已很靠前	衡量杠杆资金拥挤度	HKEX ETF AUM、成交额排名
华强北 NAND / 存储价格三天涨约 5%、前一周涨约 5%	现货价格快速上涨	影响存储协议价和市场情绪	TrendForce、DRAMeXchange、渠道报价、华强北报价跟踪
国内成熟制程加工价格从约 25 美元到约 50 美元	产能紧张导致成熟制程涨价	判断中国成熟制程盈利弹性	代工厂报价、产业链调研、券商产业链纪要
“光罩 / wafer 涨价一倍”	半导体加工环节涨价	判断供应链涨价是否广泛扩散	foundry 报价、mask shop 报价、产业链访谈
TSMC / GFS 把部分客户挤出	成熟或低端制程产能不再给部分客户	判断需求是否外溢到中国厂	TSMC / GFS 产能利用率、交期、客户转单情况

待验证点	原话含义	为什么重要	建议核验来源 / 方式
eSLR 释放约 4 万亿 / 4.5 万亿银行可购买资产	去监管带来一次性流动性释放	判断风险资产上涨的流动性来源	Fed、银行监管文件、BIS、卖方宏观报告
SOFR 低于相关基准、bank reserves 高	银行间流动性宽松	判断流动性是否仍支持风险资产	NY Fed SOFR、FRED reserves、repo market 数据
疫情 QE 每月约 1200 亿美元	持续 QE 像水龙头不断注水	对比当前一次性开闸放水	Fed QE 历史购买规模
Hyperscaler 今年 CapEx 约 7500-7700 亿、明年可能到 1 万亿	AI CapEx 规模巨大且继续上升	这是 AI infra 的直接需求来源	Microsoft / Amazon / Google / Meta / Oracle 财报和 CapEx 指引
Oracle 回收卡成本周期约 1.5-2 年	AI data center 投资回收期假设	判断 CapEx ROI 是否合理	Oracle 财报电话会、投资者日、债券材料
Oracle CDS 曾到 500 多 bps	市场曾质疑 Oracle AI backlog / 信用风险	判断 AI CapEx 的信用市场容错率	Bloomberg / Markit CDS、债券利差数据
Anthropic 去年预期 90-100 亿、后来 300 亿、五月传 550 亿	Anthropic 收入预期快速上修	Herman 认为这是主线命门	The Information、WSJ、融资文件、二级市场材料；确认 ARR / run-rate / revenue 口径
Claude 质量下降	Herman 个人使用体验，尤其编程任务	判断模型需求是否可持续	SWE-bench、Aider benchmark、用户留存、开发者社区反馈
Claude 调用量上升但满意度下降	低质量导致更多 token 调用	区分健康增长与低效率消耗	API 用量结构、企业客户反馈、推理毛利率
Kimi / 中国模型接近 Claude 编程效果	低成本模型替代风险	关系到模型商品化	第三方 benchmark、实际 coding workflow 测试、价格对比
“模型商品化”由 Larry / Oracle CEO 提到	模型差异缩小、算力和数据更重要	影响 OpenAI / Anthropic 长期护城河	Oracle CEO 访谈原文、财报电话会文字稿
AI 最终卷掉几亿人工作	Herman 的长期宏观判断	关系到 AI 生产力兑现和社会影响	劳动力替代研究、McKinsey /

待验证点	原话含义	为什么重要	建议核验来源 / 方式
			Goldman / IMF / OECD 报告
AI 是集中权力而非分散权力	长期格局判断	影响下一阶段投资主线	观察企业 AI 支出、政府 AI 采购、模型价格分层、专属算力趋势

13. 短版摘要

Herman 这段分享不是说 AI 不行了，而是提醒大家：AI / 半导体这轮已经涨得太多、太拥挤，很多人赚了很多钱，现在该从底层公司细节里拔出来，看更高层的风险。

他的核心框架是：现在 AI infra 的“金钟罩”在于市场相信 OpenAI / Anthropic 的收入增长能 justify Microsoft、Amazon、Google、Meta、Oracle 这些 hyperscaler 的巨大 CapEx。只要这个信念不破，Blackwell delay、交换机过热、光模块 delay、通胀、油价、利率这些都只是扰动。但如果模型收入增速放缓，或者 Claude / OpenAI 的质量和需求出问题，就会打到主线命门。

他讲了几个风险：第一是“低 PE 泡沫”，低 PE 不代表安全，反而会让人每次下跌都敢加仓、加杠杆；第二是流动性，eSLR / deregulation 像一次性开闸放水，但快感会递减；第三是半导体从原来 TSMC 主导的秩序市场，变成“群雄并起”，存储、MLCC、光模块、GFS、中国成熟制程都可能涨价，短期利好，长期会推高 CapEx；第四是 CapEx 和模型收入之间的钢丝很窄，hyperscaler 明年继续投入，需要 OpenAI / Anthropic 收入继续高速增长。

最值得盯的是 Anthropic / Claude：如果调用量上升但质量下降，用户为了修错反复调用，短期 token 变多，但满意度下降，未来收入增速可能放缓。再叠加 Kimi / 中国模型等低成本替代，就会出现“模型商品化”风险。

操作上，他不是建议立刻做空，反而强调不要轻易做空，因为每个底层公司缺货和盈利都是真的。真正该做的是：赚了很多后重新审视仓位、杠杆和风险，重点跟踪 OpenAI / Anthropic 月度收入、模型质量、hyperscaler CapEx 指引、信用市场、SOFR / reserves 和供应链涨价是否侵蚀 ROI。