

大模型的账本

几千亿的估值、几千亿的基建，钱到底靠什么赚回来

咕噜

费曼式写法 · 由多个 AI 代理撰写与互相审校

导读

几千亿的估值、几千亿的基建，钱到底靠什么赚回来

先别背估值公式，先想想一台三年就过时的印钞机。

假设有人向你推销一门生意：买一台机器，花掉 300 亿美元；它确实印钱——去年印了 130 亿，今年眼看着印 400 亿；但这台机器三年就会过时，到时候你得再买一台更贵的；而且隔壁中国车间刚证明了，类似的机器用十分之一的成本也能造。

你买不买？

2026 年 3 月，全世界最精明的一批机构用 1220 亿美元真金白银回答了「买」，给这门生意标价 8520 亿美元。两个月后，另一批机构给它的竞争对手标价 9650 亿。

他们疯了吗？还是他们知道什么我们不知道的事？

这本书就是把这笔账彻底拆开给你看的。不用任何金融黑话——每个名词出现的第一秒就翻译成成人话；不站队——多头最硬的论据和空头最狠的指控，都摆上桌自己比对；不假装能预测未来——给你的是三个剧本和一张七个读数的仪表盘，让你自己盯盘。

读完全书，那台「三年过时的印钞机」会变成一个你能自己回答的问题：**它印的钱，够不够在过时之前，把自己买回来？**

答案比「能」或「不能」都有趣得多。从第一章那笔算不过来的账开始。

第一章 · 一笔算不过来的账

先给你看一组真实数字，都来自 2026 年上半年的公开报道：

- OpenAI 在 2025 年全年的收入大约是 **130 亿美元**；到 2026 年夏天，年化收入（把最近一个月的收入乘以 12 推算全年）冲过了 **400 亿美元**。
- 同一时期，它一年亏掉的钱以**上百亿美元**计——光 2025 年的净亏损就接近 400 亿（其中一大块是账面上的非现金项目，经营亏损约 210 亿），2026 年的亏损预计仍在百亿量级。
- 而这家公司在一级市场的估值是 **8520 亿美元**——2026 年 3 月，它完成了一轮 1220 亿美元的融资，是人类历史上最大的一笔。管理层对外放话：上市时低于一万亿美元免谈。
- 它的竞争对手 Anthropic 更夸张：2026 年 5 月估值 **9650 亿美元**，超过了 OpenAI；年化收入口径下也冲到了 650 亿美元（注意这个数字是把通过云厂商转售的收入全额计入的「毛口径」，水分我们后面会拆）。

把这三行放在一起念一遍：收入几百亿，亏损上百亿，估值近万亿。

如果你的第一反应是「这不科学」，恭喜，你的直觉是健康的。这本书要回答的就是这个直觉背后的两个问题：

第一，这些公司到底靠什么赚钱？ 砸进去建数据中心的几千亿美金，得从哪条水管里流回来？

第二，在「现在明明在亏钱」的前提下，那近万亿的估值是怎么算出来的？ 是有一套算法，还是大家集体讲好了一个故事？

为什么这个问题值得认真问

你可能觉得「估值是投资人的事，与我无关」。但这一次不太一样。大模型公司的基建投入已经大到足以影响宏观经济的程度：2025 到 2026 年，美国 AI 相关的数据中心投资被多家机构估算为对 GDP 增长贡献了一个可观的份额；英伟达一家公司的市值一度超过了整个德国股市的总市值。如果这门生意最终算不过账，震荡不会只停留在硅谷的会议室里。

反过来说，如果它算得过账——如果十年后 AI 真像电力一样无处不在且利润丰厚——那今天看懂「钱从哪来」的人，和只看热闹的人，对未来的判断会完全不同。

先建立一个基本体感：这不是普通的「暂时亏损」

创业公司亏损不稀奇。亚马逊亏了快十年才稳定盈利，Uber 亏了更久。但大模型公司的亏损结构里有一个以前很少见的东西：**它的成本大头不是人，是机器，而且是一种会快速过时的机器。**

一家软件公司烧钱，烧的主要是程序员的工资——写出来的代码今年能用，十年后大概率还能用，是资产。大模型公司烧钱，烧的主要是 GPU——一种三年左右就会被新一代产品甩开身位、经济性急剧下降的芯片。这就像你开了一家餐厅，装修钱不是花在能用到 2050 年的楼面上，而是花在三年后就要扔掉重买的灶台上。

大白话

这本书反复用到的一个类比：大模型公司很像**电厂**。前期砸巨资建厂（训练模型、买 GPU），然后一度电一度电地卖（按 token 计费）。不同的是，电厂的设备能用三十年，它的「机组」三年就落伍。带着这个画面看后面的每一章，很多账会自己变得清楚。

这本书怎么讲

我们会沿着一条线往下走，每章解决上一章留下的疑问：

1. 先拆开看这门生意**卖什么**——token、订阅、企业合同分别是什么东西，钱具体怎么收（第二章）；
2. 再看**钱花在哪**——训练和推理是两种完全不同的成本，GPU 折旧是埋在账本里的定时炸弹（第三章）；
3. 然后算最小的一笔账：**卖出去一个 token，到底赚不赚钱**（第四章）；
4. 接着看牌桌上谁已经稳赚了——卖芯片的、卖云的、卖电的（第五章）；
5. 再回头问历史：以前那些疯狂烧钱的行业，哪些烧出了护城河，哪些烧成了灰（第六章）；
6. 然后正面硬刚估值问题：教科书方法为什么失灵（第七章），市场实际在用哪套算法（第八章）；

7. 再揭开一个让人不安的细节：这些收入里有多少是「自己人卖给自己人」的循环交易（第九章）；
8. 最后落到两个实用问题上：钱到底什么时候能赚回来（第十章），以及作为普通读者，你以后看到任何一家 AI 公司的新闻，应该盯住哪几个数字（第十一章）。

读完全书，你未必能给人算出精确的目标价——说实话，最专业的分析师也算不出来。但你会拥有一件更有用的东西：一张完整的账本，和一套自己验算的方法。下次再看到「某 AI 公司估值又翻倍」的新闻时，你能说出那句话到底哪部分是真的、哪部分是期权、哪部分是泡沫。

我们从最基础的问题开始：这门生意摆在货架上的，到底是什么商品？

第二章 · 大模型公司到底卖什么

要算一门生意的账，第一步永远是搞清楚：货架上摆的到底是什么？客户掏的钱，换走了什么东西？

大模型公司的货架上，其实只摆了一样商品：**预测下一个词的能力**。但这个能力被包装成了三种完全不同的卖法，它们的赚钱逻辑也完全不同。我们一种一种拆。

商品的原材料：token

所有卖法背后，计价的最小单位都是 token——可以粗略理解为「模型眼里的字块」。一个 token 大约是四分之三个英文单词，或者大半个汉字。你说「帮我写一封请假邮件」，这句话会被切成十几个 token 喂给模型；模型吐出来的邮件又是几百个 token。

为什么要切成 token 而不是直接按字算？因为模型内部根本不认字，它认的是这些小块对应的数字编号。你可以把 token 想象成乐高积木块：模型没见过「邮件」这个整体，它只见惯了成千上万种积木块，靠统计经验知道哪块后面大概率接哪块。

这个单位重要吗？非常重要。因为整门生意的定价、成本、毛利，全都挂在它身上。第四章我们算「一个 token 赚不赚钱」时，你会真切感受到这一点。

卖法一：按量计费的 API

第一种卖法是把模型能力做成 API——应用程序接口，一个给程序员用的「自动应答窗口」。开发者在自己的软件里写一行调用，就能把用户的问题转给模型、把答案拿回来。按用量付费，价目表公开挂着：比如「输入每百万 token 若干美元，输出每百万 token 若干美元」（输出通常比输入贵，因为生成比读取更耗算力）。

这就是第一章说的「电厂模式」最纯粹的形态：卖电，按度计价。成千上万个软件公司——做客服机器人的、做法律文书的、做代码补全的——都是买电的工厂。2026 年 Anthropic 收入暴涨，主力就是这个：它的编程助手 Claude Code 把无数开发者变成了按月重度耗电的大客户。

这种卖法的好处是**清晰**：用了多少，付多少，成本也随用量走，理论上每度电都能算出毛利。坏处是**没有忠诚度**：客户今天用你家电，明天别家降价就换——换家电线在代码里改几行就行。这为第四章的价格战埋下了伏笔。

卖法二：包月订阅

第二种卖法你我都很熟：ChatGPT 那样的订阅，每月 20 美元随便聊（重度用户有 200 美元的档位），Anthropic 的 Claude 同理。

注意这里定价方式变了：不再按 token 计，而是**包月自助餐**。为什么？因为普通用户根本算不清自己会消耗多少 token，「每月二十块随使用」是个能听懂的价格。自助餐的赌注在于：大多数人吃得少，少数大胃王被多数人摊平。

这个赌注在 AI 上有个麻烦：大胃王比普通餐厅里多得多。一个拿模型跑自动化任务的用戶，一个月烧掉的算力成本可以超过他的订阅费好几倍。所以你会看到订阅档里悄悄加上各种「公平使用限制」——本质就是自助餐里的「限时两小时」。

大白话

按量计费像出租车打表，包月订阅像健身卡。健身房赌你办了卡不来；大模型公司赌你订了阅不使劲用。区别在于：健身房你不去它几乎零成本，而你每多用一次大模型，它都要真花电费。

卖法三：企业大单

第三种卖法是把前两种打包卖给公司：全公司几千人的订阅席位、私有化部署、定制微调（用你的内部数据把模型调成「懂你业务」的版本）、合规承诺。这种单子金额大、合同期长、客户换供应商的成本高——换掉模型意味着内部系统重做一遍。

企业单是这门生意里最接近传统软件生意的部分：**毛利高、黏性强**。这也是为什么 Anthropic 几乎把全部筹码押在企业客户上，而 OpenAI 一边守着几亿消费者的订阅，一边拼命补企业课。

一个容易看走眼的坑：毛口径和净口径

读新闻时有个细节值得拎出来。2026 年 Anthropic 公布的「年化收入 650 亿美元」，是把通过 AWS、谷歌云转售的收入**全额**算进来的毛口径；而 OpenAI 公布的是扣多云厂商分成后的净口径。同一块电，经过云厂商这个「二道贩子」时，转售的流水在两边账上都可能出现。

这不是谁造假——会计准则允许，取决于合同怎么签。但作为读账的人，你得知道：**年化收入是「把最好的一个月乘以 12」，口径还分毛净**。两个数字直接对比，约等于拿别人的营收旺季乘以四去比你的淡季。第九章讲循环交易时，这个坑还会再变大一次。

收入的形状

把三种卖法合起来看，这门生意 2026 年的收入形状大致是：消费者订阅提供稳定的基本盘，API 和企业单提供爆炸性的增长（Anthropic 年化口径一年翻了好几倍，主要就靠后者），另有广告等新变现方式刚刚冒头（OpenAI 2026 年开始在免费用户里测试广告）。

增长速度是真的惊人——OpenAI 从 2023 年的 20 亿跑到 2026 年年化 400 亿量级，历史上几乎没有公司做到过。问题是：**收入涨得再快，成本涨得更快怎么办？**这就必须翻开账本的另一页了。下一章我们去看钱花在了哪里，以及那台「每天都在贬值的印钞机」到底是怎么回事。

第三章 · 一座每天都在贬值的电厂

上一章看了钱从哪进来。这一章翻开账本的另一页：钱花到哪去。大模型公司的成本结构里藏着这门生意最反直觉的部分——它看起来像软件公司，其实资产负债表像重工业。

两笔完全不同的钱：训练和推理

所有算力开销可以分成两类，它们的性格完全不同：

训练 (training)：教模型学会本领的过程。把互联网上能找到的文本喂给它，让几万台机器连轴转几个月，调整模型内部几千亿个参数（可以粗略理解为模型「大脑」里可被拧动的旋钮数量），直到它预测下一个词的准确率高到够用。一次前沿模型的训练，2025 年以后的公开估算普遍在**数亿到数十亿美元**一次。

推理 (inference)：模型学成之后，实际回答每一个用户问题的过程。你每问一次，机器就要真跑一遍。2026 年 OpenAI 光推理一项的年开销就被估算在 **140 亿美元量级**——注意，这不是建未来的产能，只是伺候好眼下的用户。

大白话

训练像制药公司研发新药：一次临床试验几个亿美金砸下去，成了就有一款药，败了就归零——而且新药专利还没到期，竞品可能已经比你好了。推理像药厂的生产线：每卖出一粒药都要真花成本。大模型公司难就难在，它**同时**背着这两种成本里最贵的版本。

基建的规模：数字大到失去直觉

先看几个 2025 到 2026 年的真实合同，感受一下量级：

- OpenAI 牵头的「星际之门」(Stargate) 计划，公开口径是几年内投入 **5000 亿美元**建数据中心；
- Anthropic 2026 年 4 月与亚马逊 AWS 签了 10 年 **超过 1000 亿美元**的算力承诺，锁定最多 5 吉瓦的电力容量——吉瓦是核电站机组的功率单位，5 吉瓦大约是几座大型核电站的出力；

- 微软、谷歌、亚马逊、Meta 四家 2025 年合计资本开支（capex，指买设备、建厂房这类一次投入、分多年摊还的长期投资）超过 **3000 亿美元**，2026 年继续加码，其中大头流向 AI 数据中心。

这不是「互联网公司的机房大一点」的问题，这是人类历史上最快的基建狂潮之一——上一章的读者此刻应该追问：好，厂房建了就建了，传统电厂也这么建，问题在哪？

问题在折旧：一台每天都在贬值的印钞机

传统电厂的机组能用三四十年，成本摊到漫长的寿命里，每年薄薄一层。AI 数据中心里的核心资产 GPU 不一样：它是人类历史上性能进步最快的工业品，新一代芯片一出来，老芯片的「每度电产出」立刻相形见绌。

这里有一场 2025 年底吵得沸沸扬扬的会计之争。空方代表（比如因预测 2008 年次贷危机出名的 Michael Burry）指控：大型云厂商把 GPU 按 5 到 6 年折旧——即假设一台芯片能体面地干满六七年——从而把成本摊薄、把当期利润做高；而实际上，AI 芯片两三年后在新任务上就明显落伍，真实经济寿命可能只有 3 年左右。如果按 3 年重算，这些公司的利润会明显缩水。

多方反驳：老芯片不会报废，只是降级——跑不了最前沿的训练，还能跑要求低的推理任务，就像退役的客机改货机继续飞。

谁对？诚实的答案是**没人知道，因为这批芯片还没活到头**。但这场争论本身就是全书最重要的信号之一：

这门生意的盈利模型，建立在一个未经检验的假设上：今天买的芯片，五年后还能赚到钱。折旧年限差两年，分摊到几千亿的基建上，就是每年几百亿的利润差异。一本账的生死，系于一个没人验证过的数字。

成本结构的全貌

把成本摊开，一家大模型公司大致背着五座山：

1. **算力 capex**：芯片、数据中心、电力容量——最大头，且前置（先付钱，后收租）；

2. **推理运营成本**：随用户数线性增长的电费式开销，用户越多亏得越多还是赚得越多，取决于第四章的单位经济；
3. **人才**：顶级研究员年薪包开到千万美元级，2025 年起几家实验室之间互相挖角的价码屡见报端；
4. **数据与训练实验**：大量训练跑完发现不行，直接归零；
5. **获客与补贴**：免费用户、低价档、企业折扣——本质都是贴钱换规模。

注意这五座山的性格：前四座跟「收入多少」几乎无关——**不管你卖不卖得出去，山都在那儿**。这就是经营杠杆高的生意——成本里一大块是固定的，不管你卖不卖得出去，山都在那儿：做大了利润惊人，做不大亏损也惊人。它解释了为什么所有人都在疯狂抢规模——这门生意里，规模不是追求，是活路。

下一章我们把镜头推到最小的一格：不看几百亿的总账，就看卖出去一个 token 的那一瞬间，这笔交易本身，到底是赚还是赔？

第四章 · 一个 token 到底赚不赚钱

把几百亿的总账放一边，这一章只算一笔最小的账：用户花一分钱买走一批 token 的那一瞬间，大模型公司是赚了还是赔了？这决定了「收入涨得快」到底是好事还是坏事。

先搞清一个 token 的成本由什么决定

生成一个 token，本质是一次计算：模型把几千亿个旋钮都过一遍。成本大体取决于三件事：

- **模型多大**：旋钮越多，每个 token 过的计算量越大，越贵；
- **芯片多快多省电**：同样的计算，新一代芯片便宜得多；
- **机器闲着还是满负荷**：GPU 是分期付款买来的，空转就是白烧钱——这跟航空公司最怕空座位是一个道理。

这三件事过去三年全都在剧烈变动，于是出现了这门生意最独特的一条曲线：

一条反常的曲线：价格自由落体，成本也自由落体

2023 年初，让最强模型生成一百万 token，开发者要付几十美元；到 2025 年，同等智能水平的服务，价格掉了一到两个数量级。换句话说，「买同样多的智能」这件事，每年便宜大约 90%。

这背后是成本在同速下降：芯片一代比一代快，工程师把模型蒸馏（把大模型的本领压缩进小模型）、剪枝、优化调度，能榨的都榨。价格追着成本往下掉——但问题来了：**谁先掉、掉多快？**

2025 年 1 月，中国的 DeepSeek 把这个问题炸到了台面上：它用远少于美国同行的训练成本做出了接近第一梯队的模型，并且把 API 价格打到地板。这件事当天就砸掉了英伟达几千亿美元市值（单日跌幅创当时美股纪录），因为它向所有人证明了两件事：第一，这层生意的「技术溢价」可以被快速抹平；第二，**只要有人愿意把价格打到成本线，这就会变成一门苦哈哈的商品生意。**

所以一个 token 到底赚不赚钱？

分情况，而且分层：

卖给付全价的客户：赚。多家机构的测算是，按 2025-2026 年的 API 标价，头部模型公司卖出去一个 token 的收入高于那一瞬间的推理成本，单看这一笔，毛利可以是正的，甚至不低。也就是说「每多卖一份就多亏一份」的说法，对付费业务并不成立。

但「那一瞬间的毛利」不是全部。完整的账还要摊进：免费用户白用的算力、上一章说的芯片折旧、天价研究员的工资、烧掉的训练实验。把这些全摊进去，整个公司就深度亏损了。这就是为什么 OpenAI 一边「每个 token 有毛利」，一边一年亏上百亿——**亏的不是这一度电，是电厂本身和研发下一代电厂的钱。**

大白话

想象一家航空公司：每张机票的票价都高于这趟航班的燃油费（单笔是赚的），但公司买了太多新飞机、还养着一支研发超音速客机的团队，所以整体巨亏。你问「一张机票赚不赚钱」和问「这家公司赚不赚钱」，是两个答案。

价格越便宜，大家用得越狠：杰文斯悖论

如果故事只是「价格暴跌」，那这门生意早死了。真正的剧情是：价格每掉一个数量级，使用量涨得更多——原来用不起的场景（让 AI 读完整个代码库、让客服每句话都过一遍模型）突然变得划算，新的需求被激活。19 世纪经济学家杰文斯发现过同样的事：蒸汽机效率提高、煤更省了之后，英国的总耗煤量反而暴涨，因为用得起蒸汽机的地方变多了。这个规律被直接称为杰文斯悖论：效率提高不一定省资源，反而可能让总消耗变大。

2026 年我们正亲眼看着它上演：token 单价一路跌，全行业卖出的 token 总量和总收入却在狂飙——第二章那几条陡峭的收入曲线，就是杰文斯悖论的账单版。

悬在头顶的剑

但这条曲线里藏着一个致命的「如果」：**如果成本下降的速度慢下来了，或者「同样的智能」不再有溢价了呢？**

价格战已经在事实上发生——DeepSeek 之后，各家 API 价格轮番下调。商品化的幽灵在于：电力生意里，一度电就是一度电，没人愿意为「更有品位的一度电」付溢价。如果大模型最终变成这样，今天的高毛利就是昙花一现；如果模型之间始终存在「聪明程度」的代差，头部玩家就能一直收到溢价。

这个「如果」，是第十章三种剧本的分水岭。而在此之前，我们先去看看这场淘金热里已经落袋为安的人——那些不等答案揭晓就已经赚到钱的家伙们。

第五章 · 真正先赚到钱的是卖水人

1849 年加州淘金热里，真正稳定发财的不是淘金客，而是卖铲子、卖牛仔裤、开旅馆的人。每一轮资源狂热都会重演这个结构。大模型淘金热进行到 2026 年，我们已经有足够的账目可以看看：钱到底先流进了谁的口袋？

头号卖水人：英伟达

英伟达 2026 财年的数据中心业务收入已经到了**每季度几百亿美元**的规模，毛利率常年维持在 **70% 上下**——这是什么概念？它每卖出一块芯片，七成是利润。制造业里这是天文数字，只有垄断级的生意才配拥有。

它的护城河不是「芯片最快」这么简单，而是 CUDA——一套让程序员指挥 GPU 干活的软件层，积累了近二十年。全世界的 AI 代码都写在它上面，换芯片等于让所有程序员重新学一门语言再重写一遍系统。芯片会过时，软件生态不会。这就是为什么每一轮模型竞赛打得越凶，英伟达越稳——**它不为任何一方的输赢下注，它向所有下注的人收桌费。**

二号卖水人：云厂商

亚马逊 AWS、微软 Azure、谷歌云扮演着「二房东」的角色：自己买芯片建数据中心，再把算力转租给大模型公司和万千企业。这笔生意的账很有意思——它们同时站在三个位置上：

- **收租的**：把算力租给 Anthropic、OpenAI 们，旱涝保收；
- **参股的**：微软是 OpenAI 最大的外部金主，亚马逊和谷歌是 Anthropic 的大股东——租金收一道，股权升值再吃一道；
- **下场的**：同时自己也在做模型（谷歌的 Gemini 甚至是第一梯队）。

一个玩家同时是裁判、房东和运动员——第九章我们会看到，这种多重身份正是「循环交易」疑云的温床。但单看眼前，云厂商的 AI 生意有个关键区别：**它们花的钱本来就要花。**数据中心即使 AI 需求不及预期，还能转租给传统云计算客户，退路比纯大模型公司宽得多。

最意外的卖水人：电力

2025 年起，这个行业出现了一句半开玩笑的共识：「AI 的尽头是电力」。数据中心的胃口大到开始扭曲能源市场：美国部分地区的电价因数据中心抢电而明显上涨；科技公司直接签下核电站的购电协议（微软为重启三里岛核电站一座机组而签的 20 年长约是最著名的例子）；得克萨斯州荒地里的风电光伏项目，因为「离数据中心近」而身价倍增。

电力的故事提醒我们一件事：**这场基建狂潮的瓶颈不是钱，是物理世界**——变压器要排队几年，电网接入要排队，燃气轮机要排队。钱能在一周内到位，一座变电站不能。

一个新物种：算力二房东

还冒出了一批专门的「算力房东」，最典型的是 CoreWeave：借钱买英伟达芯片，整栋楼转租给大模型公司，拿长期租约去银行抵押再买更多芯片。它在 2025 年上市，一度是 AI 基建狂热的晴雨表。

这类公司的账本值得多看一眼，因为它把整场游戏的风险结构暴露得最清楚：**资产是三年就可能落伍的芯片，负债是十年期的贷款，收入是「大模型公司持续扩张」这一个假设**。如果扩张停止，链条会倒着爆：租约违约→芯片抵押物贬值→贷款爆雷。它是整个生态里杠杆最高、缓冲最薄的一环。

卖水人赚到的钱，说明了什么？

把这一章的钱流画成一张图：用户和企业付钱给大模型公司 → 大模型公司付钱给云厂商和算力房东 → 云厂商付钱给英伟达和电力公司。越往上游，利润越真实；越往下游，故事越多。

这不是说下游必死——淘金热里也有真挖到金子的。但这张图给了我们一个估值的锚：**整条产业链下游公司未来要赚到的钱，至少要覆盖上游已经落袋的利润加资本成本，这场游戏才算闭环**。英伟达每赚走一块钱利润，下游就得多赚回一块多才能打平。上游的季报越漂亮，下游欠的「未来」就越多。

记住这个锚。第六章我们暂时离开账本，去历史里找三个参照系：以前那些把基建烧到天上的行业，最后都怎么收场的？

第六章·烧钱是病还是药

前五章把账拆完了：收入狂飙、成本更猛、上游已经落袋、下游整体巨亏。这样的剧情，历史演过不止一遍。这一章我们请出三位「前浪」，看看疯狂烧钱建基建这件事，什么时候是远见，什么时候是葬礼。

前浪一：铁路——烧出了未来，烧死了股东

1840 年代的英国铁路狂热：国会在几年内批准了相当于当时全国 GDP 好几成的铁路投资计划，股价疯涨，全民炒股。结局是 1847 年泡沫破裂，股价腰斩再腰斩，无数人家破。

但注意结局的另一面：**铁路留下来了**。泡沫期铺的轨道，撑起了之后半个世纪的英国经济。投资人亏光，社会血赚。美国铁路史更直白：19 世纪铁路公司成片破产，但破产清算只是换了股东，铁轨一根没被拆走，后来成了美国工业化的骨架。

铁路给大模型的启示是冰冷的：**「对社会有价值」和「对股东有回报」是两回事，中间可以隔着一整代人的亏损**。就算 AI 基建最终证明极有价值，也不保证今天出钱的这拨人能活到分红那天。

前浪二：光纤——需求是对的，节奏是错的

2000 年前后的电信泡沫更像今天。逻辑无懈可击：互联网流量在爆炸，带宽永远不够，埋光纤稳赚。Global Crossing、WorldCom 们借债几千亿美元，把海底和大陆都缠满了光纤。

结局：需求确实爆炸了，但光纤埋得比需求快得多，带宽价格暴跌 90% 以上，公司成片破产——WorldCom 的破产是当时美国史上最大。而那些贱卖清算的光纤，后来被谷歌、Netflix 们以白菜价租走，撑起了宽带时代。**先知们看准了未来，却死在了沙滩上，尸体喂饱了下一代**。

光纤留给我们的问题是全书最锋利的一个：**大模型公司是「先知」还是「捡尸体的」**？今天砸几千亿建的数据中心，会不会五年后以三折价格，便宜了下一批玩家？芯片和光纤有个关

键区别：光纤埋下去几十年不过时，芯片三年就落伍——如果这一轮赌错，连「留给下一代的资产」都会比光纤时代少得多。

前浪三：亚马逊——烧钱烧对了的样子

对照组是亚马逊：连亏将近十年，华尔街骂了将近十年。但它烧的每一块钱都进了两个地方：物流网络（别人要花同样多的钱和时间才能复制）和后来长出来的 AWS（亏出来的规模摊薄了成本，让它能把云计算卖成现金牛）。它烧钱的每一年，单位经济都在改善：同样的包裹，履约成本一年比一年低。

2000 年泡沫破裂时亚马逊股价跌了 95%——但它没死，因为它烧出来的东西**是资产，不是费用**，而且现金流始终掐在自己手里：用户先付钱，供应商后结账。

三堂课合起来：一张鉴别烧钱的试纸

把三个故事摆在一起，可以提炼出区分「药」和「病」的四个问题：

1. **烧出来的东西耐用吗？** 铁路、物流网络耐用；芯片三年过时——大模型这条偏弱，除非烧出来的是模型能力和生态这种无形资产。
2. **单位经济在改善吗？** 亚马逊是；第四章告诉我们，大模型的单个 token 经济在改善，但全公司口径还没见到。
3. **需求是真实的还是被补贴出来的？** 光纤的需求是真的（只是节奏错）；大模型的需求爆炸也是真的——这一条，大模型比光纤强。
4. **烧出的护城河挡得住复制吗？** 铁路有排他的路权，物流有实体网络；大模型的护城河是模型代差和生态——DeepSeek 证明了代差可以被打得很快。

四项对照下来，大模型公司是个混血儿：需求真实性像亚马逊，资产耐用性像光纤（甚至更差），护城河的牢靠程度未定。**它不是注定的泡沫，也不是注定的亚马逊——它还悬在中间。**

理解了「烧钱本身是中性工具」，我们才能正面进入估值问题。下一章先看教科书怎么算——以及它为什么在这门生意面前缴械。

第七章 · DCF 为什么套不上

终于到估值了。先别管大模型，我们先把教科书里最标准的那套方法——现金流折现——老老实实干一遍，看它在哪里卡住。看清它怎么死的，才能明白市场实际在用哪套「野路子」。

现金流折现：一句话版

现金流折现（DCF）的思想朴素得像常识：**一家公司值多少钱 = 它这辈子能赚回来的所有现金，折算成今天的钱，加总。**

为什么要「折算」？因为明年的 100 块不等于今天的 100 块——今天的 100 块拿去存银行，明年就变 105 了。反过来说，明年的 100 块，今天只值 95 块左右。越远的钱，折得越狠。这个「打折的狠度」叫折现率：无风险利率打底，风险越大加得越多。

所以 DCF 只需要两个输入：

1. **未来每一年的现金流预测**；
2. **一个折现率。**

听起来很客观？马上你就会发现，这两个输入在大模型公司身上，一个比一个没法填。

第一个填不上：现金流预测

给一家成熟公司做预测，锚点是历史：过去五年利润涨 8%，外推一下，差不到哪去。大模型公司没有历史——OpenAI 的收入从 2023 年 20 亿到 2026 年年化 400 亿量级，这不是增长曲线，这是爆炸。

拿它自己的官方预测感受一下：OpenAI 流出的内部预测说 2026 年亏 140 亿美元左右，2029 年到 2030 年才转正，2029 年收入目标 1000 亿美元以上。注意这些数字的性质：**它不是分析，是许愿**——是公司自己对投资人讲的剧本。五年后的收入，连公司自己都只能猜；而 DCF 要求你猜十年、二十年。

第二个填不上：折现率

一家随时可能被 DeepSeek 式黑马打穿的公司，风险该加多少？没有答案。成熟市场给稳定公司用 8%-10% 的折现率；风投给早期公司心里用 30% 甚至更高——差几个百分点，算出来的估值能差出好几倍。**折现率在这里不是一个测量值，是一个态度。**

就算都填了，还有一个结构性问题：终值吃掉了一切

DCF 实战里有个公开的秘密：对亏损成长期公司，估值的绝大部分来自终值——「十年以后每年的现金流打包折算」那一坨。原因是前面十年的现金流要么为负、要么很小，真正的钱全在「假设它最终长成大公司」那一步。

做个粗算：假设 OpenAI 2035 年做到 2000 亿美元收入、25% 利润率，即年赚 500 亿；之后永续增长 3%，折现率 12%。这套假设下终值折到今天大约是几千亿美元量级——和实际估值能对上。但你注意到没有：**这串数字里没有一个测出来的，全是摆上去的。**利润率从 25% 改成 15%，估值少三分之一；折现率从 12% 改 15%，再少三分之一。同一家公司，诚实的 DCF 可以给出从「两千亿」到「两万亿」的任何答案，取决于你调哪个旋钮。

大白话

DCF 像一台精密的天平，但你放上去的砝码全是你自己猜的。它不会告诉你一家公司值多少钱，它只会把你的信念翻译成美元。对大模型公司做 DCF，本质是把自己的乐观程度过一遍磅。

教科书还教了什么？倍数法也失速

另一条常用捷径是市销率 (P/S，市值除以年收入)：成熟的软件巨头大概在 8-15 倍。OpenAI 8520 亿美元估值对年化 400 亿美元收入，是 20 倍出头；如果用 2025 年实际收入 130 亿算，是 60 多倍。Anthropic 看起来便宜些——9650 亿估值对 650 亿毛口径年化收入约 15 倍——但别忘了第二章拆过的口径水分：它的年化数是「毛口径、最好月份乘 12」，按 2026 年实际收入估算是 40 倍上下。

20 倍市销率意味着什么？意味着买家不是在为现在的收入付钱，而是在为「收入还会爆炸很多年」付钱。这不是估值错误，这是另一种逻辑：**这根本不是市盈率市销率的地界，这是**

期权的地界。

下一章我们就看市场实际在用的那套算法——它不在教科书里，但每个人都在用。

第八章 · 给未来买一张期权

教科书方法失灵了，但钱是真的在流：一轮一千多亿美元的融资，真金白银打进了账户。出钱的人不傻——他们是全世界最精明的机构。他们用的算法不在教科书里，却自有一套严密的逻辑。这一章就把它破译出来。

算法一：风投的幂律，放大到国家尺度

风投行业的底层数学是幂律：一个基金投二十个项目，十九个死光没关系，只要剩下那个涨一百倍，整盘就赚翻。所以风投天然不在乎「大概率亏钱」，只在乎「小概率的赢有多大」。

现在把这个逻辑放大一万倍：如果一个投资人相信「AI 会重塑整个经济，最终赢家年收入以万亿计」，那么花 8000 亿买一张「可能是那个赢家」的门票，在小概率剧本下依然划算——就像花一块钱买一张期望值两块、但九成九是废纸的彩票。只要「赢家的奖品」足够大，任何今天看起来离谱的价格都能被合理化。这就是为什么估值争论总在「你信不信那个剧本」上打转，而不是在现金流上打转。

算法二：期权，不是股票

更精确的说法：买大模型公司的股权，本质是买一张看涨期权——一种「约定未来以某个价格买入的权利」。期权的特点是：最多亏掉买权的钱，赚的却不封顶。

金融里期权有个反直觉的性质：不确定性越大，期权越值钱。因为波动只会放大上行，下行永远锁定在「全亏光」为止。这解释了一个让价值投资者抓狂的现象：大模型行业每出一次「前途未卜」的大新闻，估值反而常常涨——因为对期权买家来说，混乱本身就是价值的一部分。你担心的「看不清」，正是他们买的东西。

算法三：锚定与稀缺

还有一层更朴素的机制。一级市场没有连续报价，一轮融资的定价很大程度上锚定上一轮：「上一轮 3000 亿，这轮收入翻了三倍，怎么也得 8000 亿吧。」涨跌幅有惯性，而「错过上一轮」的懊悔（业内叫 FOMO，错失恐惧）会替高价格背书。

稀缺性再添一把火：想投 OpenAI、Anthropic 的机构排长队，额度要抢。当买家竞争的不是「值不值」而是「能不能进场」，价格就由最乐观的那个人定——这在拍卖理论里叫「赢家诅咒」的亲戚：成交价永远偏向最敢想的人。

算法四：战略买家买的根本不是股权

还有一类买家，账根本不是按股权算的。微软投 OpenAI 一百多亿的时候，它买的是：Azure 的云算力大单、Copilot 全家桶的引擎、以及「不让谷歌拿到」这件事本身。亚马逊、谷歌投 Anthropic 同理——投资款一转头就变成云服务订单流回来。

对战略买家，这不是投资，是**采购加防御**：就算股权归零，只要绑住的云收入和使用量够大，账已经平了。这也解释了为什么这些巨头对估值数字本身出奇地不敏感。

这套算法合法吗？合法。危险吗？危险

把这四种算法合起来，你会得出一个有点不安的结论：**今天大模型公司的估值，是期权定价、幂律信仰、拍卖机制和战略采购的混合体——唯独几乎不含「现金流」这一项。**它是自洽的：每个买家按自己的算法都算得过账。但它的自洽依赖一个公共前提：那个「万亿级赢家」的剧本必须最终有人演成。

期权还有一个特性：到期日。信念可以支撑估值很多年，但不可能永远——最终要么现金流兑现（期权变成股票），要么剧本破产（期权变废纸）。第十章我们会算这笔「到期日」的账。

但在此之前，必须先揭开牌桌上最让人心里发毛的一个角落：那些撑起收入数字的合同里，有相当多是在「自己人」之间签的。下一章，循环交易。

第九章 · 循环交易疑云

前面八章都在善意地算账。这一章要把灯打向牌桌底下：那些漂亮的收入数字里，有多少钱只是在圈子里打转？这个问题不阴谋论——它是 2025 到 2026 年华尔街最严肃的一场争论。

先看几笔 2025–2026 年真实发生的交易

- **英伟达 ↔ OpenAI**：2025 年 9 月，英伟达宣布向 OpenAI 投资最高 1000 亿美元；同一框架下，OpenAI 承诺部署数百万块英伟达 GPU——投资款转个身，以芯片订单的形式流回英伟达。麻省理工的教授评论这笔交易「kind of a wash」（差不多是原地打转）。
- **OpenAI ↔ AMD**：2025 年 10 月，OpenAI 承诺采购 AMD 芯片（6 吉瓦规模），作为交换，AMD 给了 OpenAI 一批认股权证——一种「未来可以按约定低价买股票」的凭证，最多可拿走 AMD 约 10% 的股份。买方买东西不掏全款，卖方倒贴股权。
- **云厂商 ↔ Anthropic**：亚马逊、谷歌给 Anthropic 投了上百亿美元，Anthropic 转身承诺十年内向 AWS 采购超 1000 亿美元算力、向谷歌采购 400 亿美元级别——投资款同样以云订单的形式回流。
- **英伟达 → CoreWeave → 大模型公司**：英伟达投资 CoreWeave，CoreWeave 拿融资买英伟达的芯片，再租给大模型公司。钱在同一个生态里转了一圈，每一站都记成了「收入」。

到 2026 年 3 月，连英伟达自己都表态「对 OpenAI 和 Anthropic 的投资可能是最后一批」——黄仁勋的解释是它们要上市了、机会窗口关了，但市场读出的另一层意思是：这种结构的压力已经大到不便再加码。

这种模式有个名字，还有个前科

这在会计和金融史上叫厂商融资（vendor financing）：卖方借钱给买方，让买方有钱买自己的货。而它的前科非常难看——2000 年电信泡沫里，北电、朗讯就是这么干的：借钱给新兴的电信公司，电信公司拿钱买它们的设备，收入报表一片繁荣。泡沫破裂后真相大白：

那些「收入」对应的客户根本没有真实的付费能力，厂商融资成了史上最贵的左口袋倒右口袋。WorldCom、Global Crossing 的坟头，第六章我们刚刚上过香。

对照之下，今天的结构在机制上同构：**生态里最强的玩家用自己的利润，给生态里最弱的玩家输订单，输进来的订单再变成自己的收入**。钱在闭环里流动，每一步都合法记账，但整个闭环对外是否真的有「新钱」流入，报表上看不出来。

但先别急着喊泡沫：三点辩护

诚实的分析必须给对方律师说话的机会。这个结构和 2000 年有三个关键区别：

1. **输血方的体质不同**。当年的北电自己是借钱放贷；今天的英伟达拿的是真金白银的利润（70% 毛利卖芯片赚来的），微软、亚马逊、谷歌拿的是每年几千亿美元现金流的家底。庄家用自己的钱坐庄，和借钱坐庄，崩盘的传染性完全不同。
2. **终端需求是真实的**。第六章的试纸说过：几亿用户真在用 ChatGPT，企业真在为 Claude 付费，第二章那条收入曲线的大部分是外部客户的真钱。循环交易是**叠加**在真实需求之上的一层放大器，而不是凭空捏造需求。
3. **比例问题而非性质问题**。没有证据表明这些公司的收入「大部分」是循环的——合理的担忧是「有一部分被放大，且放大的比例看不清」，而不是「全是假的」。

大白话

判断循环交易有多严重，就问一句话：**如果把生态内互相买的东西全扣掉，剩下的收入曲线还陡吗？**还陡，就是真实需求外面包了一层泡沫糖衣；不陡了，就是 2000 年重演。目前没人能从公开数据里精确算出这个答案——这正是第十一章清单里要你自己盯的一项。

为什么它仍然是悬顶的剑

即便按最善意的解释，循环结构也制造了两个真实的危险：

危险一：放大了繁荣的表观斜率。真实需求增长 50%，循环放大后报表可能显示 100%。按第八章的期权逻辑，估值锚定的是报表斜率——于是估值也被放大。这是一条自我强化的链条：估值越高→越容易融资→越能下大订单→供应商收入和估值越高→越有钱投回来。链条向上转时是火箭，向下转时是漩涡。

危险二：风险的隐性集中。表面上看，风险分散在芯片商、云厂商、模型公司、算力房东、电力公司几百个玩家身上；实际上它们被股权和订单互相捆成了一只刺猬——任何一个环节的需求塌方，会顺着合同链条瞬间传遍全身。2008 年次贷危机教会我们的正是这件事：**分散持有不等于分散风险，如果所有人持有的是同一件事的话。**

好，账本的两面都翻完了：会算的部分（第一到五章）、历史的部分（第六章）、方法的部分（第七八章）、阴影的部分（第九章）。现在终于可以正面回答书名里的那个问题了——钱，到底什么时候、靠什么，才能赚回来？

第十章 · 钱什么时候能赚回来

到这一章，所有零件都摆上桌了：收入的三种卖法、成本的两座大山、token 的单位经济、上游落袋的利润、历史的三种结局、失效的教科书、期权的定价法、循环的阴影。现在把它们组装起来，回答书名里的问题：**砸进去的几千亿，到底靠什么、什么时候赚回来？**

诚实的答案不是一个数，是三个剧本。每个剧本在历史上都有对应的真实案例，每个剧本成立需要的条件，我们都已经具备了检验工具。

剧本一：电力剧本（乐观）——「成为下一个亚马逊云」

剧情：模型能力持续拉开代差，头部两三家始终比追赶者聪明一截，于是溢价守住了；同时 token 成本每年降 90% 而价格只降 70%，毛利逐年走阔；企业把 AI 像当年上云一样全面接入，最后大模型变成新一代水电煤——人人都用，头部躺赚。OpenAI 自己的剧本就是这个：2029-2030 年转正，之后收入往千亿美元以上走。

成立的条件：① 模型代差能守住（DeepSeek 们追不上来）；② 企业渗透率继续暴涨（今天全球白领工作里真正深度用 AI 的仍是少数，天花板还远）；③ 推理成本下降快于价格下降。

历史上演过：AWS。亚马逊云前十年也在疯狂砸基建，最后砸成了互联网的基础设施层和亚马逊的利润奶牛。这个剧本下，今天的估值不算贵——一张万亿级基础设施层的门票。

剧本二：云计算长跑（中性）——「能赚，但要十年，而且没今天想的那么肥」

剧情：模型确实家家都要用，但差距抹平得比预期快，价格战常态化，最后活下来的三四家赚的是「规模辛苦钱」——像今天的云计算：有利润、很稳、但不是印钞机。基建靠**十年以上的长跑**摊薄回本，就像电厂：一度电一度电慢慢收，收三十年。

成立的条件：① 需求持续扩大（杰文斯悖论继续有效）；② 头部玩家不在长跑中途失血而亡——这是最难的一条，因为芯片三年过时的折旧诅咒（第三章）让「慢慢回本」这条路格外陡峭；③ 融资环境不断粮，撑到盈亏平衡。

历史上演过：美国铁路的正半场——活下来的干线铁路最终成了好生意，但那是在成片破产重组之后。这个剧本下，今天的估值偏贵：生意是真的，价格是把「电力剧本」的概率定得太高了。

剧本三：光纤剧本（悲观）——「需求是真的，股东是祭品」

剧情：开源模型（DeepSeek、Llama 一系）把「够用水平」的智能打成免费品，价格战击穿成本线；与此同时上一代芯片快速过时，数据中心资产减值；收入撑不起债务，链条上杠杆最高的玩家（算力房东、高负债二线模型公司）先爆，冲击顺着第九章的合同链条传导。技术本身照样改变世界——但跟 2000 年一样，**改变世界的是活下来的技术和捡便宜资产的下一批人，不是今天付万亿估值的股东。**

成立的条件：① 开源持续咬住闭源的后脚跟；② 模型进步放缓，「代差溢价」消失；③ 利率/融资环境收紧，烧钱链条断粮。

历史上演过：Global Crossing。需求预测全对，股东颗粒无收。

今天的证据站在哪边？

三个剧本各有一些正在发生的事实支持：

- 支持电力剧本：收入曲线的斜率（第二章）史无前例；企业渗透率确实还在早期；头部两家的代差至今存在。
- 支持长跑剧本：价格战已经打起来；OpenAI 自己的时间表就是「再亏四年」；电力和芯片的物理瓶颈在拖慢所有人的速度。
- 支持光纤剧本：DeepSeek 证明过代差可以一夜抹平；循环交易在放大表观需求；Michael Burry 们已经开始做空这个叙事。

我的诚实结论是：**「行业」几乎注定赢，「公司」未必，「今天的估值价格」悬。**AI 会像电力一样普及——这一点三个剧本都不否认；争的是谁收到电费、一度电能挣几分钱、以及今天买门票的人排不排得到分红那天。第六章铁路那堂课此时读来最刺骨：对社会有价值和对股东有回报，中间可以隔着一整代人。

那么，作为不掌握内幕的普通人，你能做什么？你不需要赌剧本——你需要的是一套仪表盘，在剧本逐渐明朗的过程中持续校准自己的判断。最后一章，我把这张仪表盘给你。

第十一章 · 一张给普通读者的检验清单

这是最后一章，也是唯一一章你可以抄下来带走用的。前十章讲的所有道理，最后都可以浓缩成几个数字——以后你每看到一条 AI 行业的大新闻，对着这张清单核一遍，就能自己判断：剧本在向哪边移动。

仪表盘上的七个读数

读数一：成本曲线和价格曲线，谁降得快？（第四章）每年「同等智能的价格」大约降 90%。如果某家的**成本降得比价格快**，单位经济在改善（电力剧本）；反过来就是流血竞争（光纤剧本）。看它的方法很简单：盯 API 价目表的更新和芯片代际发布，一年对一次。

读数二：毛利率的趋势，而不是绝对值。软件公司的及格线是 60-70%。大模型公司今天的毛利被免费用户和推理成本压着，关键不是现在是多少，而是**每个季度是不是在往上爬**。连续几个季度不动甚至下滑，说明「规模带来利润」的信仰没有兑现。

读数三：收入的成色。三个问题：是净口径还是毛口径（第二章的坑）？企业合同占比是升是降（企业单才是最黏的钱）？续费率如何——客户第二年还续不续，是「AI 是必需品还是尝鲜」的最终投票。一家公司敢公布续费率，本身就是信心的信号。

读数四：资本开支和折旧年限。（第三章）盯两件事：capex 占收入的比例有没有开始收敛（一直发散说明「先建后收」永远停在「先建」）；以及——更隐蔽也更重要——**有没有公司悄悄延长芯片的折旧年限**。折旧年限从 5 年改到 6 年，利润凭空变多，一分钱生意没多做。Burry 之争后，这已经是做空机构盯梢的第一现场。

读数五：收入的循环度。（第九章）问一句：这条收入新闻里，付钱的那一方，是不是同时拿了卖方的投资？出现「A 投 B 一百亿，B 承诺买 A 一千亿」结构的新交易越多、占收入比例越高，表观需求的折扣就该打得越大。

读数六：现金跑道。（第八章期权逻辑的到期日） $\text{现金及等价物} \div \text{年烧钱额} = \text{还能活几年}$ 。这个数低于两年，公司的所有战略都会变形——你会看到涨价、裁员、疯狂融资，或者最危险的：为了不翻车而签下更多循环交易。

读数七：代差的秒表。每次开源模型发布，量一量它落后最强闭源模型几个月。这个「时间差」就是大模型行业溢价的实体化：时间差拉大，电力剧本加分；时间差缩到一两个月以内，溢价逻辑瓦解，光纤剧本加分。

三个黄色警报

读数之外，还有三种事件本身就该让你调高警惕，无论新闻标题多兴奋：

1. **融资越来越依赖「生态内」的钱**（供应商、云厂商、互相持股），而不是外部财务投资人——说明外人已经不愿按这个价进了；
2. **「承诺」替代「交付」成为新闻主体**——几千亿的意向协议满天飞，但已建成的产能、已确认的收入增速开始掉队；
3. **高管和早期投资人开始密集套现**——他们比任何人都清楚剧本写到第几页。

回到最初那笔账

第一章我们摆出那组算不过来的数字：收入几百亿、亏损上百亿、估值近万亿。现在你手里有了全部零件，可以把那句话翻译成人话了：

收入是真的，增长是真的，亏损也是真的；估值不是对今天的定价，是一张以「万亿级未来」为标的的期权；这张期权的价格里，混着信仰、拍卖的狂热、战略买家的私心，和一层循环交易吹出来的泡沫糖衣。基建能不能赚回来，取决于三个剧本里哪个成真——而这不取决于任何人的愿望，只取决于成本曲线、代差秒表和需求真实性这三条硬线。

所以「大模型公司到底值多少钱」这个问题，正确的回答从来不是「值 X 千亿」，而是：**「在剧本 A 下值这个数，在剧本 C 下接近归零——这是七个读数，你自己盯着看。」**

这本书写于 2026 年 8 月。书里的具体数字明年就会过时——估值会变、收入会变、也许又有新的 DeepSeek 改写牌桌。但账本的结构不会变：谁在收租、谁在付租、资产几年过时、钱在不在圈里打转、成本线在哪。学会了读这本账，下一次技术狂潮再来的时候——它一定会来——你不用等别人告诉你该兴奋还是该恐惧。

你自己有表。

大模型的账本 · 咕噜