

ZHIGUAN STUDIO

AI 时代的价值分布

收费权 · 泡沫 · 人的位置

研究报告三部曲 · 合订本

质观工作室 · 二〇二六年七月

总目

01

价值分布——从算力到应用，谁掌握收费权

价值落在最后一层：资源、入口、流程三类收费权

02

泡沫诊断——价格是否跑在现金流之前

技术革命可以成立，资产价格仍可能失败

03

人的位置——AI、劳动与不可外包的责任

当生成趋近免费，人的位置转向判断、选择与责任

完整版本包含摘要、术语说明、数据注、附录与参考文献。

01

价值分布

从算力到应用，谁掌握收费权

价值落在最后一层：资源、入口、流程三类收费权

质观工作室 · “AI 时代的价值分布”三部曲之一（2026）

摘要：1980 年，微软并没有发明最好的操作系统，却通过保留授权权利，占住了 PC 产业中最关键的入口。四十多年后，AI 产业链正在重演相似的故事：英伟达数据中心业务毛利率约 75%，AI 服务器整机约 4.8%，而部分大模型云端 API 部署的毛利率为 -0.4%。利润并不按照“离 AI 核心技术有多近”来分配，而是流向最难复制、最难替代的环节。本篇依次分析五类价值位置：算力与芯片的利润集中在制造和系统工程；模型价格持续下降，价值转向推理基础设施、专有数据和可信评测；AI 原生软件增长迅速，壁垒主要来自 workflow、数据和交付；垂直行业未必需要最大的模型，却需要最深的行业理解和工程能力；超级个体的生产门槛下降后，判断、信任与分销仍决定其上限。文中把这些难以被抹平的部分概括为“最后一层”，并进一步区分出资源、入口和流程三类“收费权”。最后，文章讨论中国的政策与算力环境，并提出一套判断 AI 项目真伪的五问框架。

关键词：产业链；利润池；算力；模型即服务；AI 原生软件；垂直行业；超级个体；最后一层；收费权

本篇术语说明

- **token（词元）：**大模型处理和生成文本时使用的最小计费单位，调用量和费用通常都按 token 计算。
- **MaaS（模型即服务）：**将大模型能力封装为云端接口，并按照调用量收费的服务模式。
- **LLMflation：**业内对大模型推理价格快速下降现象的称呼，近年大致呈现每年下降一个数量级的趋势。
- **毛利率 / 经营利润率：**毛利率为“收入减直接成本”占收入的比例；经营利润率还会进一步扣除运营开支，更接近企业真实的经营盈利能力。本文比较时会注明口径。
- **ARR / NRR：**ARR 指年化经常性收入；NRR 指净收入留存率，即老客户一年后贡献的收入比例，高于 100% 说明客户仍在扩容。
- **利润池：**利润在产业链不同环节之间的分布格局。
- **智算中心：**专门服务于 AI 训练和推理的数据中心。
- **光模块 / HBM / 先进封装：**分别指数据中心高速光互联器件、贴近 AI 芯片的高带宽内存，以及多芯片高密度集成工艺。三者都是当前算力供给的重要瓶颈。
- **最后一层：**本系列提出的概念，指 AI 让生产和复制趋近免费之后，价值仍然能够停留的区域，完整界定见第八章。
- **收费权：**本系列提出的概念，指企业在产业链中占据他方难以绕开的位置，并由此获得持续定价能力。下文将其分为资源、入口和流程三类，见第九章。
- **驾驭工程（harness）：**个人将习惯、知识、价值方向与审美方向外化为 AI 工作流的工程层。本文只在个人与产业的对照中简要提及，第三篇将专章展开。

上部 一条链的解剖

一、三个毛利率的谜题

1980 年夏天，IBM 正在秘密开发个人电脑，并为它寻找操作系统。公司代表来到加州蒙特雷湾，敲响了一栋海边小楼的门。屋主是加里·基尔代尔（Gary Kildall）。当时，他几乎是微型计算机操作系统领域的霸主，CP/M 装在那个年代绝大多数像样的微机上。接下来发生的事，成了计算机史上最著名的悬案之一。流传最广的版本说，IBM 到访时，基尔代尔正驾驶私人飞机外出，把西装革履的代表晾在了门外。他的遗著和同事则坚称，他后来赶回去谈过，只是双方卡在授权条款和保密协议上，最终没有谈拢。

真相究竟是哪一种，并不影响结果：IBM 转身去了西雅图，找到一家只有几十名员工、甚至连操作系统都没有的小公司。那家公司叫微软。比尔·盖茨手里同样没有现成的操作系统，但他知道去哪里买。他花约 7.5 万美元，从街对面的西雅图计算机产品公司买下一套仿照 CP/M 开发的 QDOS。这个名字的字面意思是“又快又脏的操作系统”。改变产业格局的，不是这次购买，而是随后那份授权协议。盖茨允许 IBM 使用这套系统，却拒绝将其彻底卖断，并保留了向其他电脑厂商继续授权的权利。

IBM 当时并不在意。在这家硬件巨头看来，利润来自机器，软件只是随机附赠的说明书。几年后，兼容机厂商蜂拥而至，每一台“IBM 兼容机”都要向微软支付一笔 DOS 授权费。发明 CP/M 的基尔代尔逐渐被遗忘；制造 PC 的 IBM，后来把整个 PC 业务卖给了联想。反而是那个没有发明操作系统、只是“知道去哪里买，并且保留了授权权利”的人，守住收费亭整整三十年。（“基尔代尔开飞机去了”是流传版本，其本人回忆录对此有异议，本文按历史悬案处理。）这个故事提前呈现了本文的核心命题：

收费权取决于位置，而不只取决于技术。

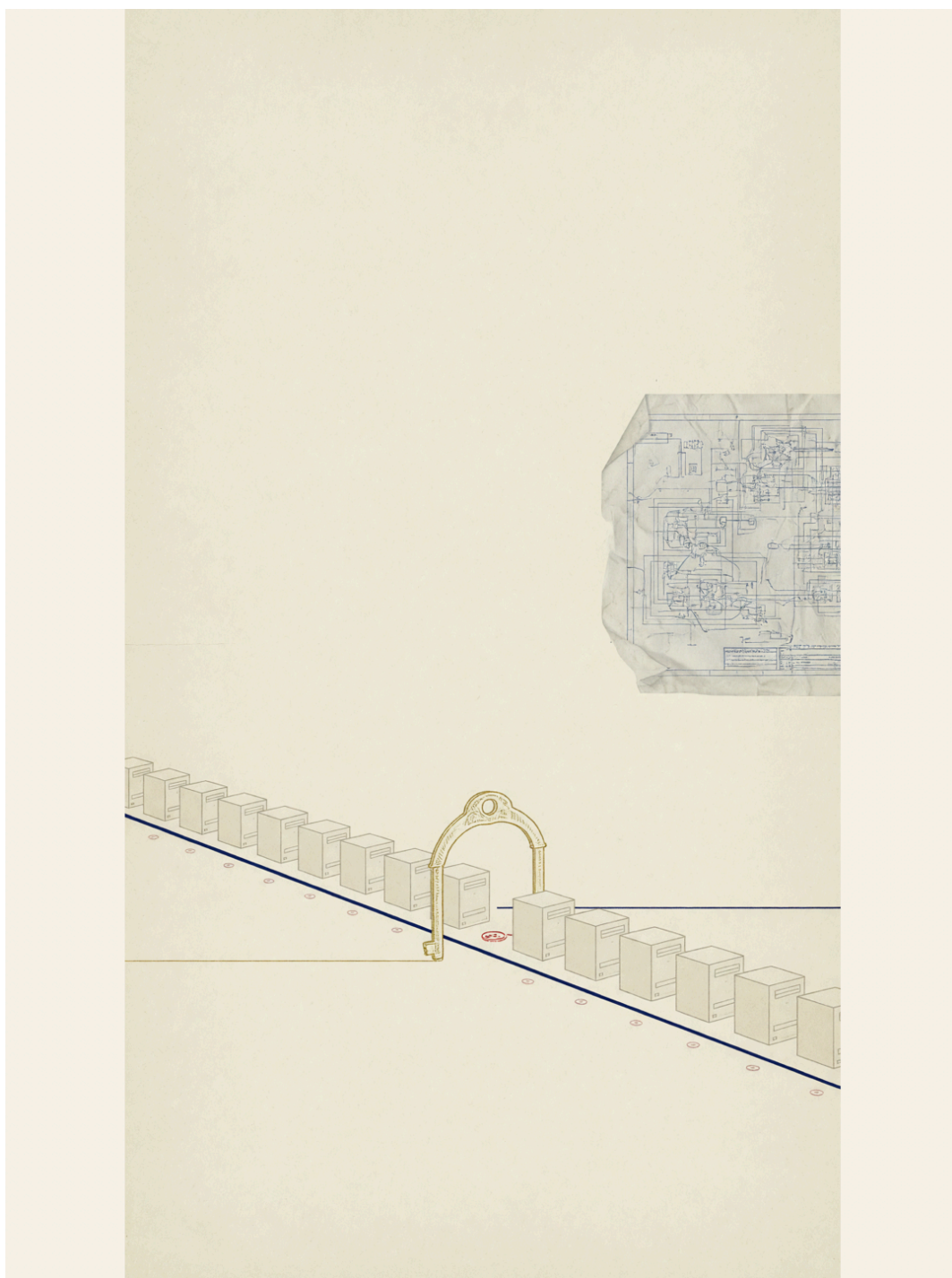


图 1 收费权是位置，而非技术本身

最早发明技术的人未必赚到最多的钱，制造机器的人也未必掌握最终利润。长期收益往往属于那个占住“别人绕不开的位置”的人。当智能本身越来越便宜，钱会停在哪里？先看三个数字。同一条 AI 产业链上，2026 年同时存在三种截然不同的毛利率：最上游，英伟达数据中心业务长期维持在 71%—75%；往下一环，把芯片装进机箱的 AI 服务器整机厂，毛利率约为 4.8%；再往下，智谱招股书披露，其云端 API 部署毛利率为 -0.4%。模型每被调用一次，企业都可能亏一点钱。同一场技术革命，有人拿走七成以上的毛利，有人只能留下不到五个点，还有人的毛利为负。利润不会均匀地流过整条产业链：它会在少数环节沉淀，也会从另一些环节迅速穿过。

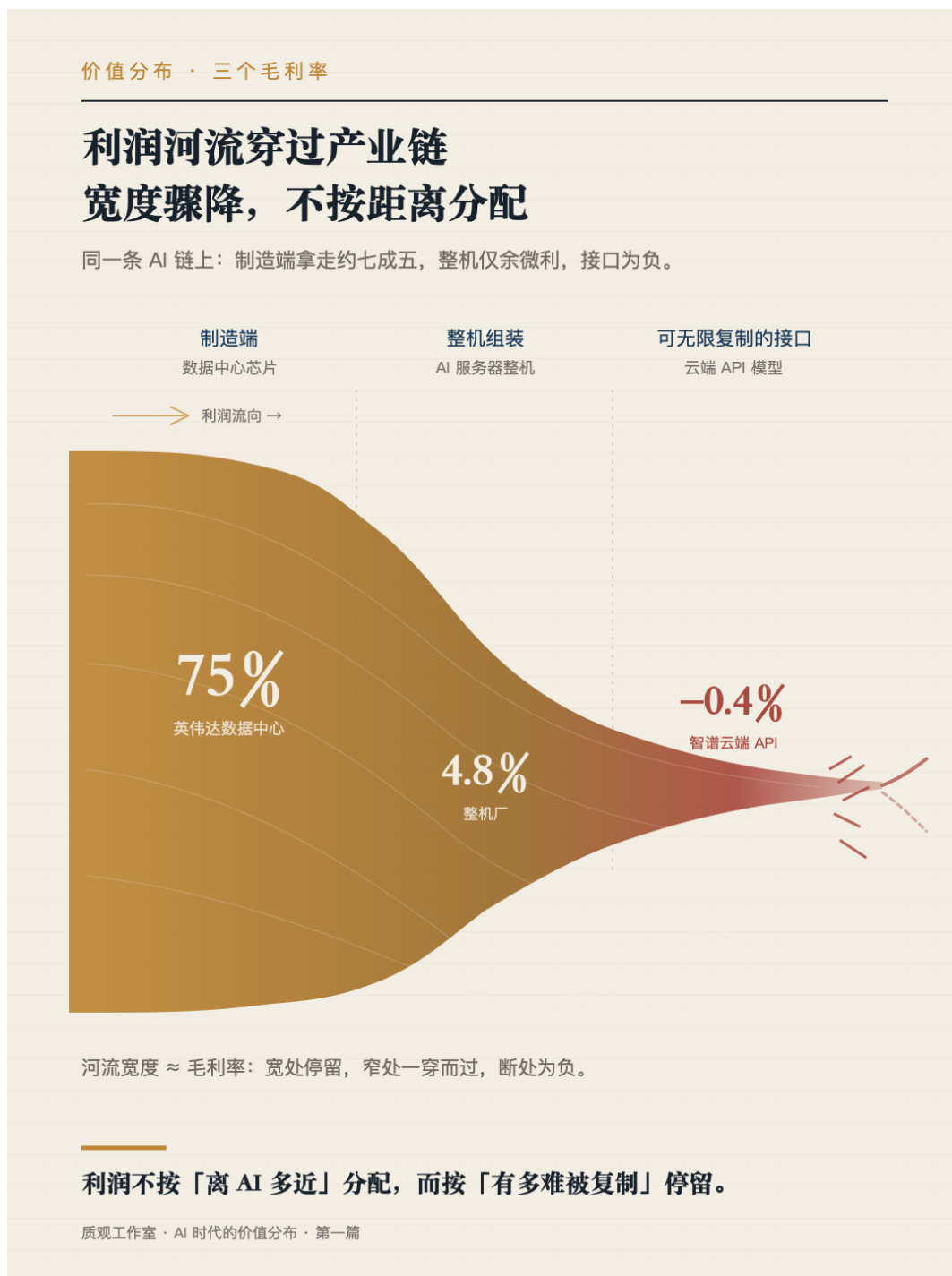


图 2 利润河穿过产业链：宽处停留，窄处穿过

可以借用一个并不严格、但有助于观察加价环节的能源类比。黄仁勋把 AI 数据中心称为“AI 工厂”，其产出物是 token。沿着这个比喻继续往下推：芯片产能像油田，算力像原油，智算中心像炼油厂，光模块像输油管道，token 像成品汽油，应用则是上路的汽车。这个类比不等同于产业分类，却提出了一个更具体的问题：一滴油从井口流到油箱，究竟在哪几段被加价？

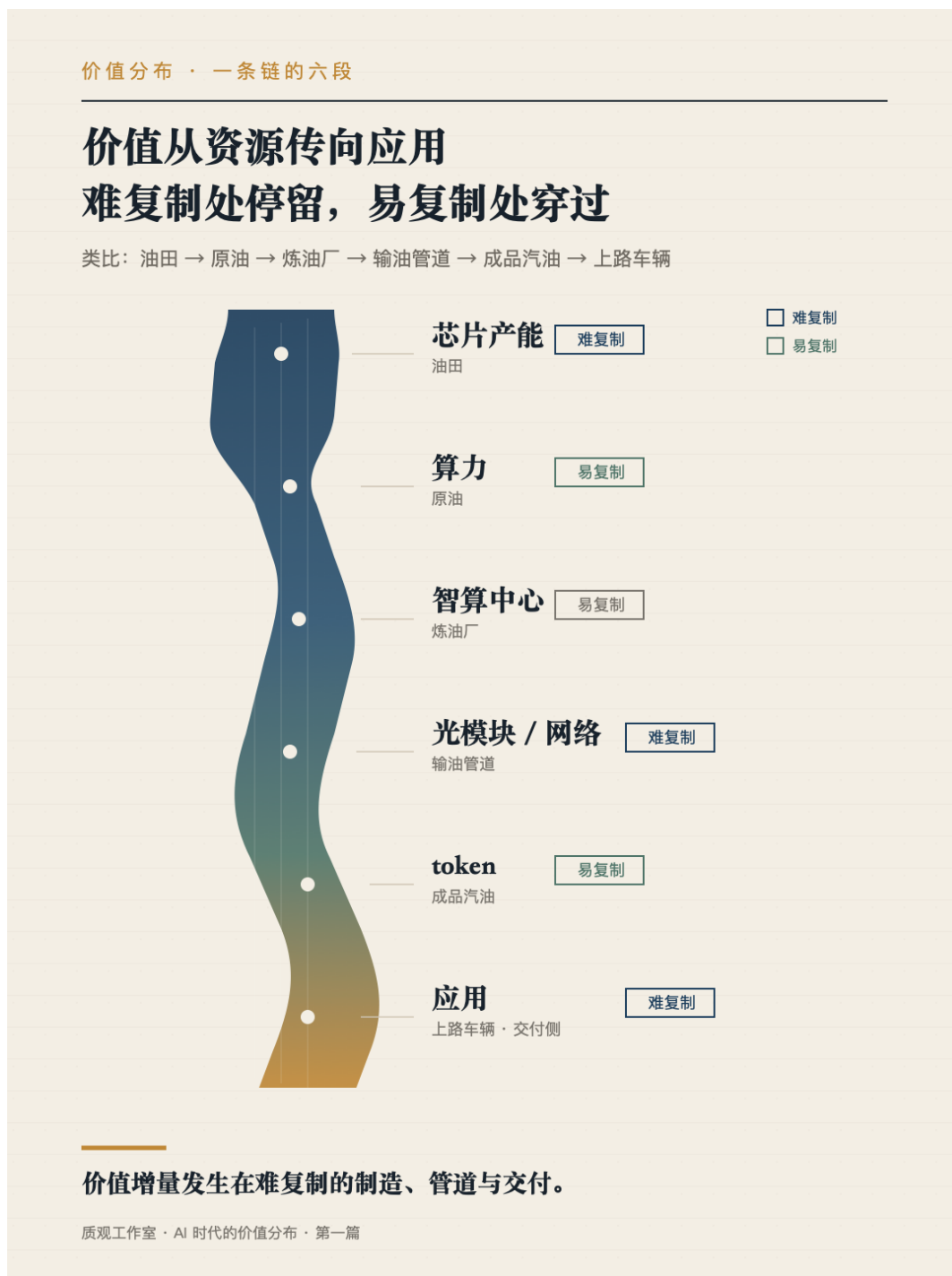


图 3 一条链的六段：难复制处停留，易复制处穿过

下文六章沿着产业链逐层分析五个核心环节。先给出一条判断利润能否留下来的标尺，随后再看各环节的差异：

判断一个环节的利润能否留下来，重点不是它离 AI 概念多近、吸引多少资本，而是它有多难被复制。

二、算力与芯片：利润在制造，不在囤积

先看算力总量。据 Epoch AI 估算，美国前五大云厂商合计掌握全球约 71% 的先进 AI 算力，中国约占 12%—14%，两者相差约五倍。若换算为 H100 等效芯片，谷歌约有 500 万张，微软约有 340 万张，而中国全国合计约 280 万张。一家头部美国云厂商的规模，就可能与中国全国总量接近。

最先进的一档芯片又长期受到出口管制影响。2025 年下半年，围绕 H20、H200 的政策数次收紧和松动。长期来看，中美算力生态受此影响越来越可能走向相对独立：中国很难单纯依靠采购最先进芯片来追平总量，更现实的路径，是用国产芯片、系统工程和充裕电力，在“性能够用”的基础上扩大规模。不过，比总量更能说明问题的，是利用率。据《科技日报》等报道，全国智算中心平均利用率约为三成。问题不在建设数量，而在建成之后没有充分转化为产出。原因主要有三类：第一，建设过热。部分号称“万卡集群”的项目，实际上把芯片分散在多个机房，无法作为一个整体投入训练。

第二，运营成本过高。西部千卡级中心每年运营开支可能达到约 3000 万元，长期处于亏损状态。第三，异构调度困难。英伟达与昇腾的软件栈互不兼容，不同芯片很难被统一纳入同一个资源池。进入 2026 年后，市场进一步表现为结构性错配：低质量的中小智算中心仍在空转，高端芯片和稳定的大集群却重新变得紧张。阿里巴巴 CEO 曾表示“几乎没有一张卡是空闲的”；年内算力租赁价格回升 20%—40%，H200 整机报价也由约 230 万元升至 380 万元。这组数据表明，约束已经从算力保有量转向稳定、可用的产出能力。

在这一环，定价者是英伟达。其 2026 财年收入约 2159 亿美元，其中数据中心业务约 1937 亿美元，第四季度 GAAP 毛利率达到 75%。这份高毛利主要来自三层壁垒：

- 算力瓶颈：训练、推理、视频生成和代码生成都在持续消耗算力。
- 系统绑定：客户一旦大规模部署 CUDA 生态，迁移成本极高。
- 供给锁定：CoWoS 先进封装和 HBM 本身就是稀缺产能，控制稀缺环节会进一步强化议价能力。

三层壁垒叠加之后，整条产业链新增的算力需求，往往首先变成英伟达的订单。风险同样清楚：英伟达最大的客户，也恰恰是最积极的替代者。微软、谷歌、亚马逊和 Meta 一边大量采购，一边自研 ASIC。未来若算力需求从“训练稀缺”转向“推理规模化”，同时推理框架逐渐抹平 CUDA 的差异，英伟达的定价地位就需要被重新评估。判断英伟达需要看期限：它现在收取的，究竟是阶段性的稀缺溢价，还是能够长期延续的基础设施收益？把利润沿着产业链摊开，衰减速度很快：制造端毛利率可达 71%—75%，服务器整机约为 4.8%，纯算力租赁则接近无利可图。

美国 CoreWeave 专营 GPU 租赁。其收入一年翻倍，经营利润率却从 17% 跌到 1%，负债逼近 250 亿美元。租赁环节缺乏足够的技术壁垒，价格战很快就会抹平利润；制造端则掌握制程、封装、生态和供应链等多重壁垒，因而能够把利润留下来。至此，谜题的第一格解开：

在算力这一环，买入并囤积资源并不能构成好生意；难以复制的制造和系统能力，才是生意。

那么，中国能够占住哪一段？答案仍然集中在制造和工程。出口管制造成了国产替代的刚性需求，寒武纪毛利率约为 55%。华为则选择“用系统补单点”：单颗昇腾芯片的性能约为英伟达顶配产品的三分之一，但 CloudMatrix 384 通过全光互联，将 384 颗芯片连接为一个整体。按 SemiAnalysis 的建模，这套系统用约五倍的芯片数量，实现约 1.7 倍的综合性能反超，代价是约四倍功耗。这条路线成立的前提，可以概括为“缺硅不缺电”。中国电力相对充裕，工业电价较低，绿电产能较大；海外超大规模数据中心却开始自建电厂、争夺核电资源。

在整条链上，中国最硬的一环之一是光模块。中际旭创全球份额约 23%，连续三年位居第一，毛利率约 46%；全球前十家光模块厂商中，中国企业占七席。短板则在最核心的高速光芯片，国产化率仍不足 15%。液冷同样是确定性较强的工程方向。随着单机柜功率持续上升，传统风冷越来越难以支撑，冷板、管路、冷媒和整机柜集成会形成一整套工程体系，而这恰好落在中国制造的长板上。能源、光模块、液冷、异构调度，这一环留给多数中国主体的机会，都属于“难以速成的制造与工程”，而不是“多买几张卡”。

三、模型与推理：价格每年下降一个数量级，钱往哪里流

从芯片继续往下，最重要的变量，是推理成本下降得有多快。a16z 将这种现象称为 LLMflation：过去三年，在达到同等能力的前提下，推理成本大约每年下降十倍。以 GPT-3 级能力为例，2021 年底每百万 token 的价格约为 60 美元，三年后，最便宜的同档模型已降至约 0.06 美元，累计下降约一千倍；GPT-4 级高性能模型按相同口径计算，也下降了约 62 倍。

Epoch AI 给出了另一种估算。它以博士级科学问答为基准，测得推理价格平均每年下降约 40 倍，同时强调不同任务之间差异很大，年降幅可能从 1.7 倍到 900 倍不等，且最近一年的极快降幅未必能够持续。两个口径虽然不同，却指向同一个方向：模型能力正在快速商品化。价格下降并没有缩小这门生意，反而扩大了总需求。Agent 工作流会把单个任务的 token 消耗提高 10—100 倍，企业的总支出不降反升。业内把这种现象称为“token 税”；在经济学上，它对应杰文斯悖论，即效率提升之后，总消耗反而增加。

这一环的机会主要分布在降价与扩量曲线的两侧，而非简单“卖 token”。第一侧，是推理软硬件。当单纯出租 GPU 时长的毛利趋近于零，能够提供低延迟、高并发和低成本推理的公司，才可能获得溢价。晶圆级芯片公司 Cerebras 专注推理，2026 年 2 月融资后估值约 230 亿美元，并与 OpenAI 签下大额推理协议；Groq 与英伟达达成非独家授权；提供开源模型推理与 MaaS 服务的 Together AI，估值约 83 亿美元，年度签约额超过 11 亿美元。背后是一场结构性换挡：算力需求正在从“训练稀缺”转向“推理规模化”。

训练通常是一次性的、集中式的，并且高度依赖最先进芯片；推理则是高频、分散的，对成本和延迟更加敏感。2025 年 DeepSeek R1 发布后，需求迅速向推理侧转移，一批原本为训练而建的西部智算中

心随之闲置，这可以看作中国市场的一次预演。训练端的英伟达很难替代，但标准化、高频、成本敏感的推理，恰恰是云厂自研 ASIC 最可能切入的部分。第二侧，是 MaaS 及其配套服务。2025 年上半年，中国公有云大模型调用量达到 536.7 万亿 token，半年约增长四倍。火山引擎、阿里云和百度智能云三家合计占据九成以上份额，字节豆包的日均调用量一年增长超过十倍。

创业公司也在承接 DeepSeek 带来的推理需求。硅基流动旗下 SiliconCloud 曾连续数周位居 OpenRouter 日消耗榜首；无问芯穹累计融资近 15 亿元，管理着分布在 26 座城市的 53 个数据中心。但规模增长不等于盈利成立。潞晨科技曾测算：满血版 DeepSeek-R1 如果日产千亿 token，每月机器成本约 4.5 亿元，月亏损约 4 亿元。这一结果与 DeepSeek 自称的“成本利润率 545%”正面冲突。这场争议至少提示：满血大模型推理的市场定价可能过低，纯 MaaS 未必是一门好生意。

再往上一层，是 Agent 基础设施和数据评测。Agent 执行任务时，需要记忆、编排、评测和可观测等一整套工程能力。LangChain 融资后估值约 12.5 亿美元，其 LangSmith 平台服务超过 300 家企业客户。企业每年计划投入数千万乃至数亿美元用于生成式 AI，监控、调试和评测随之成为刚需。

数据侧则更靠近前沿。2025 年 6 月，Meta 以约 143 亿美元取得 Scale AI 约 49% 的股份，创始人 Alexandr Wang 加入 Meta。交易随后造成“中立性真空”：Google、OpenAI 因顾虑而削减合作，Mercor 和自举盈利的 Surge AI 因此受益。前沿实验室已经不再满足于低端标注。它们愿意付费购买的，是专家级、可评测、能够直接支撑模型改进的数据。这样的数据本身就难以复制。Bessemer 对这一环的判断很明确：企业把 AI 做起来的最大瓶颈不在模型，而在评估。能够提供持续、私有、可信评测的供应商，会掌握企业信任的入口。这一环的价值分布可以概括为：

智能本身在贬值；围绕智能形成的规模、专有数据和信任在升值。

利润通常不留在 token 本身，而留在 token 两侧。



图4 模型趋于廉价，价值向工作流与交付移动

数据注：中国公有云 MaaS 份额（IDC，2025 年上半年）为火山引擎 49.2%、阿里云百炼 27.0%、百度智能云千帆 17.0%；字节豆包 2025 年 12 月日均调用量突破 50 万亿 token；无问芯穹管理算力超过 2.5 万 P；LangSmith 平台累计处理超过 150 亿条 trace、逾百万亿 token；Scale AI 交易对应估值约 290 亿美元；Mercor 估值约 100 亿美元，管理三万余名承包商，日付薪超过 150 万美元；Surge AI 年化收入约 12 亿美元。

四、AI 原生软件：增长的形状

应用层离场景最近、进入门槛最低、创业公司股份最高，也最容易被误读。据 Menlo Ventures，2025 年企业在生成式 AI 上的支出达到 370 亿美元，同比增长 3.2 倍，其中应用层占 190 亿美元，超过一

半；创业公司拿走了 AI 应用市场约 63% 的份额。a16z 的判断是：模型能力越强，应用层并不会因此变薄，反而可能变得更有价值。关键在于把智能放进具体场景，而不在于占有多少原始智能。首先要区分两种产品。

传统 SaaS 加装 AI 功能，往往只是给原有 workflow 装上辅助件。用户仍然要进入系统、点击菜单、填写字段、导出报表，AI 只负责缩短其中几步。AI 原生软件则从一开始就把模型放在交互核心。用户不再逐项操作，而是直接说出目标，例如“找出可能流失的客户”或“生成一套报价方案”。系统随后自动调用数据、规则、模型、权限和 workflow，完成整个任务。这背后的商业变化，是责任的转移：传统软件卖工具，结果由客户负责；AI 原生软件卖结果，厂商需要对结果承担更多责任。红杉把这场变化概括为：云计算时代的企业出售软件、按席位收费；AI 时代的企业出售工作成果、按结果收费。

增长差异可以从三组指标观察，但这些数据来自不同研究口径，正式发表前仍需回源核实：

- ARR 增长率中位数：AI 原生企业约 100%，传统 SaaS 约 23%。
- 净收入留存率：AI 原生企业约 118%，传统 SaaS 约 95%。前者的老客户仍在扩容，后者的客户收入则在缓慢收缩。
- 人均 ARR：AI 原生企业约 23.8 万美元，传统 SaaS 约 7.1 万美元。

速度差距更加明显。AI 原生企业做到 10 万美元 ARR，平均约需 4.7 个月；传统 SaaS 通常需要 12—18 个月。做到 100 万美元 ARR 时，AI 原生企业的平均团队规模约为 4.2 人，传统 SaaS 则需要约 14 人。据 Bessemer，AI 企业达到 1 亿美元年化收入，平均只需约 5.7 年。速度来自验证链条被显著缩短。只要有一个边界清晰的场景和一个能运行的原型，客户往往可以当场看到效果，销售周期随之被压缩。

另一面是：AI 产品第一次演示很容易惊艳，首月流失率也可能更高。要形成留存，仍然需要连接数据、配置权限、改造流程和培训团队。AI 原生软件的留存优势，往往要到三个月之后才会显现，前提是产品已经进入客户流程。定价方式也在变化。按席位收费适合“企业扩张就增加账号”的时代，但 AI 会制造一个悖论：假设某客服系统把 50 人的工作压缩到 30 人，按席位收费的供应商反而会因为客户效率提高而损失收入。在这个约束下，AI 原生软件开始转向按用量或按结果计费。行业研究称，约 41% 的 AI 原生软件已经采用用量或结果导向定价；但贝恩对三十余家在位厂商的调研显示，65% 仍只是“席位价加用量费”，没有一家完全转向纯结果定价。趋势已经明确，但商业模式仍在过渡：

收入逻辑正在从“有多少人登录”，转向“系统替客户完成了多少工作”。

也就是从出售访问权，转向出售完成度。

五、护城河的位置：从套壳到交付

应用层增长很快，复制同样很快。判断谁能活下来，要少看产品演示的惊艳程度，多看护城河的位置。最典型的方向，是“服务即软件”：把过去依赖人工完成的专业服务，直接交付为可以计费的结果。

法律领域的 Harvey，在 2026 年 3 月融资后估值约 110 亿美元，覆盖 1300 余家客户和超过 10 万名律师，红杉连续领投三轮。客服领域的 Sierra 按“成功解决的会话”收费，语音渠道的使用量已经超过文本。医疗领域的 Abridge 把医患对话直接生成结构化病历，并深度集成 Epic；OpenEvidence 对医生免费，通过药企广告变现，据称已被约 65% 的美国医生使用。

这些公司的商业模式各不相同，价值来源却高度一致：模型本身并非重点，重点是深度嵌入专业场景之后积累的能力，包括对行业工作流的理解、与既有系统的衔接、进入专业人员日常决策的默认入口，以及“为结果负责”形成的责任壁垒。反面案例同样具有代表性。AI 销售公司 11x 对外宣称年化收入接近 1000 万美元，但 TechCrunch 核查后认为实际约为 300 万美元，客户流失率高达 70%—80%，部分所谓标杆客户甚至从未签约。评估这类公司，不能只听自报数字，还要核对可验证的客户留存。

再看 Agent 与工作流。这是当前最热门、估值也最陡峭的方向之一。Cursor 年化收入约 20 亿美元，估值约 293 亿美元；Cognition (Devin) 估值约 260 亿美元，约为收入的五十倍，并已收购 Windsurf。Replit、Lovable、Perplexity 等公司也分别达到不同量级。不过，这个赛道存在一个根本约束：复合错误率。

假设 Agent 每一步的准确率为 95%，连续执行 100 步之后，整体成功率只剩约 0.6%；即使单步准确率提高到 99%，整体成功率也只有约 37%。更可靠的做法，是把任务拆成多个阶段，并在每一段设置检查点，及时拦截错误，而不是让 Agent 一口气跑完长任务。这也是第三篇将展开的“驾驭工程”的核心机制。据此看，仅仅在大模型外面封装一层界面的 Agent 最为脆弱。底层模型来自 OpenAI 或 Anthropic，一旦模型厂商亲自进入编程、办公等应用场景，外层封装很容易失去价值。可靠性工程、工作流嵌入和专有数据积累，才有机会留下来。

几家头部投资机构的判断已经趋于一致：

- a16z 认为，模型正在商品化，代码可以快速重建，真正难以复制的是流程能力、专有数据和切换成本。
- Bessemer 提出，“行动系统正在取代记录系统”，AI 原生应用不只保存数据，还会直接对数据采取行动。
- 红杉则判断，“价值将在应用层累积”。

三句话共同指向一个位置：价值来自模型进入真实工作流之后凝结下来的部分，而非模型本身。竞争格局随之分层。在标准化的通用场景中，巨头可以依靠分销和捆绑守住市场，例如把 AI 直接嵌入办公套件、CRM 或 ERP。对于需要重构工作流、行业知识密集的场景，创业公司则可能凭借“为 AI 而生”的架构先取得优势，达到一定规模后再被巨头收购或整合。Cognition 收购 Windsurf，就是这种收编过程的一部分。对创业者而言，最危险的是进入巨头可以顺手捆绑的通用场景；更值得做的，是那些必须重新设计工作流的垂直领域。

在这一格局中，中国团队在全球移动 AI 榜单上的可见位置，主要来自面向海外的消费级应用，而不是国内企业工作流软件。a16z 2026 年 3 月的榜单显示，全球前 50 名移动 AI 应用中，约 22 个由中国团队开发，但其中只有 3 个主要面向国内市场。代表产品集中在视频、图像和陪伴类应用：快手可灵年

化收入约 2.4 亿美元，其中海外占七成；美图海外月活超过一亿；MiniMax 携 Talkie 上市，收入七成以上来自国际市场。

但规模不等于收入。2025 年上半年，全球生成式 AI 应用下载量接近 17 亿次，ChatGPT 一家却占据全球 AI 应用收入约 63%。中国出海产品虽然用户量大，单品收入普遍仍然偏低。地缘风险更加刚性。DeepSeek 在多个国家受到监管或下架，Talkie 一度被美国区应用商店下架；Agent 创业公司 Manus 年化收入约 1.25 亿美元，Meta 原计划以约 20 亿美元收购，但交易最终因中国技术出口审查而被否决并取消。最后还要看毛利。

传统 SaaS 毛利率通常可达 80%—90%，AI 原生软件则普遍只有 50%—60%，因为每一次调用都产生真实的推理成本。Bessemer 将 AI 公司分为两种典型：一种是“超新星”，第一年就冲到约 4000 万美元 ARR，但毛利率只有约 25%；另一种是“流星”，第四年达到约 1 亿美元 ARR，毛利率接近 60%。模型价格每年下降十倍，Agent 又把单个任务的 token 消耗提高 10—100 倍，两股力量相互抵消。结果是：可无限复制的通用接口可能每单都在亏钱，而难以复制、深度绑定客户的私有交付才有机会盈利。应用层的谜底，与算力层相似：

产品完成只是一半，价值要在客户流程中兑现。



图 5 从套壳到责任交付：护城河阶梯

数据注：服务即软件各家估值（多为公司自报或第三方估算）——Sierra 约 158 亿美元，Decagon 约 45 亿美元，Abridge 约 53 亿美元（覆盖 150 余家医疗系统），OpenEvidence 约 120 亿美元；Agent 赛道中，Replit 约有 50 万付费企业客户、估值约 90 亿美元，Lovable 上线八个月做到 1 亿美元年化收入、四个月翻倍、估值约 66 亿美元，Perplexity 年化收入约 5 亿美元、估值约 200 亿美元；出海阵营中，可灵 2025 年 12 月单月营收突破 2000 万美元、全球用户超过 6000 万，爱诗科技（PixVerse）融资约 3 亿美元、全球用户破亿，字节 CapCut 移动端月活约 7.36 亿。

六、垂直行业：参数不值钱，交付才值钱

把“出售结果”的逻辑推进到制造、能源、医疗等重场景，就进入了垂直行业大模型。这里也是检验“算力能否转化为产业价值”的主要场所。工业场景首先给这门生意设下三道约束。第一道约束是数据形态。工业数据主要是表格、时序和传感器信号，而这些恰恰不是语言大模型最擅长处理的对象。第二道约束是容错率。在通用场景中，一次幻觉可能只需要重写一段文本；在工业场景中，一次误判却可能演变为安全事故。理想 MEGA 的一次冷却液召回，损失超过 11 亿元。

第三道约束是数据主权。工艺参数和配方属于企业核心机密，“数据不出厂区”是常态；而最贴近生产现场的 OT 层数据，又往往积累不足、质量不高。三道约束叠加，决定了工业 AI 不可能复制消费互联网式的扩张速度。在工业场景中，所需的“大模型”未必需要特别大。官方测评显示，主流大模型在工业应用中的平均准确率不足 60 分，GPT-4 和文心一言也只有约 70 分。麦肯锡对 2025 年灯塔工厂的统计则显示，排名前五的用例中，77% 采用传统判别式 AI，生成式 AI 只占 9%。

判断工业 AI 案例时，还要警惕“名不副实”。宁德时代所谓“缺陷率降至十亿分之一”，核心技术是机器视觉；工业富联营收中的“AI”，很大一部分来自 AI 服务器代工。看案例不能只看名字里有没有“大模型”，而要看它到底解决了什么问题、用了什么技术、产生了多少可量化价值。落地效果较好的案例，往往具有鲜明的共同特征。中控技术的流程工业大模型被用于乙烯裂解装置，预测准确率达到 99.79%，单炉年净收益超过 300 万元。它没有追求通用智能，而是把模型嵌入一个边界清晰、结果可以量化的工艺环节。

华为盘古气象大模型登上《Nature》，在确定性指标上超过传统数值预报，运算速度提升约一百万倍，并已被香港天文台投入使用。两个案例虽然行业不同，却共享三项条件：场景边界清晰，结果可以量化，价值集中在难以复制的环节。三项同时成立，才更接近真实机会。这一环最大的难题更多在商业模式，而非技术。谜题的第三格也在这里解开。智谱招股书显示，本地私有化部署的毛利率约为 66%，云端 API 部署则为 -0.4%。能规模化的云端业务，每做一单都可能亏钱；能够赚钱的本地部署，又被困在项目制中，难以复制，增长存在上限。

这不是智谱一家公司的问题；面向工业客户时，这种结构性两难普遍存在。客户要求私有化、定制化和数据不出厂，天然会把业务推向“高毛利、难扩张”的一端。国务院研究室有关人士也曾提醒，要防止“私有化加项目制”造成市场碎片化。有没有办法缓解这组矛盾？可能的出路是分层，而不是二选一：将跨行业通用、可以标准化的部分，例如通用视觉检测和设备预测性维护，做成可规模化产品；将涉及各厂核心机密的部分，保留为本地定制交付。前者负责摊薄成本，后者负责保住毛利。有机会完成这种分层的，往往不只是纯 AI 创业公司，而是那些长期深耕工业软件和自动化、既懂行业又拥有交付队伍的既有厂商。

最终瓶颈仍然落在人和交付上。业内测算，同时具备机器视觉能力和产线经验的工程师，供需比只有约 0.2。比人才短缺更困难的，是把一次成功试点复制成标准化交付：同一个场景在不同组织中的落地周期，可能从几周拉长到几年；设备型号、数据格式、工艺参数和管理流程各不相同，再加上数据不能离开厂区，供应商很难用一套云端模型服务所有客户，只能逐家部署。这也解释了工业 AI 为什么长期“叫好不叫座”：最终胜出的，未必是模型最强的公司，而是行业理解最深、工程交付最扎实的公司。

政策正在推动这一方向。工信部等八部门发布的“人工智能+制造”专项行动提出，到 2027 年推动 3—5 个通用大模型深度应用，推出 1000 个工业智能体，推广 500 个典型场景，并鼓励地方发放算力券和模型券。不过，渗透率数据需要谨慎。网传“工业企业大模型使用率一年从 9.6% 升至 47.5%”，目前无法回溯到任何官方报告。可溯源的数据包括：IDC 估算 2024 年中国工业大模型应用市场约 12.1 亿元；第三方预计 2030 年渗透率约 50%；CNNIC 统计中国生成式 AI 用户达到 6.02 亿。无论采用哪一种口径，判断仍应回到一点：

渗透率上升，并不等于已经有人赚到了钱。

七、超级个体：一人公司的上限

产业链的末端，是个人。据 Carta 统计，美国由单人创立的公司占新设企业的比例，从 2019 年的 23.7% 上升到 2025 年的 36.3%，五十多年来首次超过三分之一。Anthropic CEO 判断，“一人独角兽”出现的概率约为七至八成，最可能诞生在自营交易和开发者工具领域。Altman 的说法更直接：未来的创业公司，可能是“一个人加一万块 GPU”。案例已经开始出现。

Base44 由一名创始人和八名员工组成，产品上线三周做到 100 万美元年化收入，成立六个月后被 Wix 以 8000 万美元现金收购，整个过程没有融资。类似样本还包括 Cal AI、Midjourney、ElevenLabs、Mercor 和 Bolt.new。它们的共同特征是：团队极小，人效极高，从零到规模的时间大幅缩短。

AI 对个体生产率的提升，也有较扎实的实证支持。GitHub 对近五千名开发者进行随机对照试验，结果显示，使用 Copilot 的开发者每周完成的代码合并数量约增加 26%，经验较浅者受益更明显。斯坦福与 MIT 对真实呼叫中心的研究覆盖三百万条对话和五千余名客服，结果显示，员工每小时解决的问题平均增加 14%，新手提升达到 34%。麦肯锡估算，生成式 AI 在 63 个用例中每年可创造 2.6 万亿至 4.4 万亿美元价值，其中约四分之三集中在客服、营销销售、软件工程和研发。

一个人究竟如何指挥一支“虚拟团队”，第三篇会从劳动形态出发展开“驾驭工程”：把个人的习惯、知识、价值和审美，外化为一套长期生长在自己工作流中的工程系统，再由它调度空白算力。这里先看一个量级案例。YC 现任 CEO Garry Tan 开源的个人工作流，由约两百行命令行外壳和二十多个 Markdown“角色技能”组成。据他自述，依靠这套配置，他在 60 天内完成数十万行生产代码，同时仍在全职运营 YC。但产业侧要追问的是：一套个人驾驭工程，究竟能够支撑多大的收费权？答案受三重约束；这些约束并不会随着生产门槛下降而消失。第一重，是判断。

生成越快，审核和把关就越值钱。据 Anthropic 2026 年的一项调查，开发者已经把约六成工作交给 AI，但完全不复核输出的人只占 0—20%。当人长期被动接受 AI 输出，对自己交付内容的理解会逐渐下降，业内将其称为“认知债”。第二重，是信任。企业与个人的差别，不只在团队人数，还在能否被审计、被起诉、被追责。减重药公司 Medvi 只有两名员工，临床工作全部外包，营收达到数亿美元；但在 2026 年初收到 FDA 警告信后，公司随即陷入集体诉讼和数据泄露争议。在医疗、金融、招聘等行业，牌照、合规和真人背书，是一人公司很难跨越的墙。第三重，是分销。

当生产成本快速下降，竞争壁垒会更多地转移到品牌、渠道、牌照和网络效应上。a16z 的判断是：代码本身通常不是价值所在，新进入者无法在短期内复制别人多年积累的信任和分销网络。还有一个延伸问题：AI 能否继续改进自己，让“人指挥 AI”进一步变成“AI 指挥 AI”？现有研究显示，AI 已能自行标注、自定奖励、自行出题、求解和评判。但目前主要结论仍有三点：第一，AI 以自定目标训练自身，前期可能有效，训练过久却会出现自我劣化；第二，以强训弱虽然可行，训练出来的模型仍弱于教师模型，并且可能出现“作弊”；第三，截至 2026 年年中，AI 尚未实现完全自我改进。Anthropic 一位联合创始人曾预测，到 2028 年底，AI 研发不再依赖人类的概率约为 60%。

人的位置由此更清楚：设定目标，校准方向。AI 自我改进的主要风险，不只是算力失控，而是目标错位。这又回到“驾驭工程”的定义：在习惯、知识、价值与审美四个成分中，价值方向最难通过训练直接获得。超级个体的上限可以概括为：

生产的门槛已经降到个人，收费权的门槛却没有。



图 6 超级个体的瓶颈：信任与分销

一个人能走多远，取决于他能积累多少难以被他人复制的东西：别人不具备的判断、专有数据、客户关系，以及可被追责的信誉。这就是“最后一层”的个人版本。

数据注：一人公司与小团队样本——Cal AI 青少年团队做到约 3000 万美元年收入后被收购；Midjourney 约 40 人、不融资、年收入数亿美元；ElevenLabs 以约二十个五至十人小队运作，人均年营收约 82.5 万美元，年化收入超过 3.3 亿美元；Mercor 做到 5000 万美元 ARR 时约有三十名全职员工；Bolt.new 上线五个月做到 4000 万美元 ARR，团队不足 40 人。

中部 理论：最后一层与三类收费权

八、最后一层：价值为什么向难以复制处集中

现在回到开头的三个毛利率：英伟达约 75%，服务器整机厂约 4.8%，智谱云端 API 约 -0.4%。前面的链条分析给出的解释相同：AI 正在把“生产”和“复制”的成本推向极低水平。凡是能够被系统化、无限复制，或者直接由模型生成的部分，其价值都在被迅速抹平；利润只会停在那些仍然抹不平的地方。这些抹不平的部分包括：判断、专有数据、行业交付、信任和分销。本系列将它们统称为最后一层。

这里需要说明：“最后一层”和“收费权”是本系列为方便讨论提出的概括，并不是现成的行业术语。但它们背后的判断有明确来源：a16z 强调流程能力与专有数据，Bessemer 提出“行动系统取代记录系统”，红杉认为“价值将在应用层累积”。三家机构从不同角度描述的是同一件事：模型正在变成类似水电的通用要素；能够持续定价的，是某个位置上他方难以获得的能力。模型为什么会持续商品化？前文已经反复出现三个原因：第一，开源模型快速追赶闭源模型，企业可以在不同模型之间低成本切换。第二，推理价格持续下降，“可负担”本身不再构成壁垒。

第三，最前沿的几类模型在通用能力上逐渐接近，“更聪明一点”对多数应用不再具有决定性意义。经济学早已提供了相应框架：当一种投入品变得充裕而同质，超额利润就会离开它，流向那些仍然稀缺、难以复制的互补环节。前面的章节虽然讨论不同产业位置，追问的都是：互补环节究竟在哪里？这个结构不仅存在于产业，也存在于个人工作流。一个人的习惯、知识、价值和审美，同样会凝结成难以复制的一层。第三篇将把它称为“驾驭工程”，并从劳动哲学层面展开。产业中的“最后一层”和个人的“驾驭工程”，是同一种结构在不同尺度上的显现。

九、三类收费权：资源、入口、流程

将“最后一层”落到产业链中的具体位置，可以进一步区分为三类收费权。



图 7 最后一层与三类收费权

这里所说的“收费权”并非法律许可，而是一种定价能力：企业占据他方难以绕开的位置，并把技术优势持续转化为收入、毛利和现金流。

1. 资源收费权

资源收费权来自对算力、系统交付和关键供应链瓶颈的控制。英伟达是典型案例。其约 2159 亿美元收入和 75% 毛利率，不止来自芯片设计，还来自对加速计算系统、CoWoS 先进封装和 HBM 等稀缺

环节的综合掌控。它的风险也非常明确：一旦瓶颈缓解，或者客户自研能力成熟，稀缺性就会下降。资源收费权能够维持多久，取决于稀缺本身能够维持多久。

2. 入口收费权

入口收费权来自对开发者和企业工作流入口的控制。OpenAI 与 Anthropic 争夺的，正是这个位置。竞争重点正在从“谁的模型更聪明”，转向“谁能更深地进入用户工作流”。据报道，Anthropic 年化收入从 2024 年初约 8700 万美元，增长到 2026 年约 300 亿美元量级，企业级份额提升到三至四成；OpenAI 的企业份额据称从约 50% 回落到 25%—27%，同时还背负巨额现金消耗和基础设施义务。这些数据来自媒体与二级市场口径，波动较快，并非审计数据。入口收费权最大的风险，是商品化。当模型差距缩小、切换成本下降、开源模型已经“够用”时，平台就可能退化成普通供应商。1980 年的盖茨已经演示过这场争夺的核心：入口收费权的要义是守住安装发动机的那道门，并且拒绝把它彻底卖断，而非造出最好的发动机。

3. 流程收费权

流程收费权控制的是模型进入企业数据、权限、行动和审计链条的方式。这是三类收费权中最难建立、也最难拆除的一类。Palantir 提供了一个完整案例。先看数字。2026 年第一季度，Palantir 总收入达到 16.33 亿美元，同比增长 85%，为公司历史最高增速；调整后的经营利润率达到 60%，“Rule of 40”约为 145%，远高于健康软件公司通常采用的 40% 门槛。数字之外，还要看它究竟在卖什么。

Palantir 出售的不是更强的模型，而是一套让模型“被允许在组织内部行动”的系统。其中的基础是 Ontology：将工厂、产线、设备、订单和客户抽象为可查询、可授权、可触发、可审计的业务对象；AIP、Foundry 和 Apollo 再以此为基础，把模型能力转化为组织可以执行的行动。企业应用 AI 的难点通常不在生成一段文本，而在让系统有资格行动：它必须读取正确的数据，识别权限，把建议写回工单和审批系统，并对每一次结果留痕、审计和追责。掌握这一层的企业，收入更像组织操作层的租金，而非 token 费。

模型可以替换，但企业的本体建模、数据权限和流程嵌入一旦完成，迁移成本会非常高。高毛利和高留存也随之产生。市场给予 Palantir 约 3150 亿美元市值、约 43 倍市销率，定价所指向的并非模型本身，而是它占据了一个“他方难以绕开的行动控制位置”。当然，边界也必须看清。43 倍市销率已经计入多年高增长预期，一旦增速回归，回撤空间同样巨大；政府收入占比较高，既构成壁垒，也带来政治和合规风险；高度定制的交付会限制复制速度，也可能受到云平台同类能力或客户自建系统的替代。两件事可以同时成立：价值确实落在行动控制层，高估值却未必一定能够兑现。

三类收费权讲完之后，还需要一把统一的尺。仅看毛利率并不足以判断收费权是真是假，更应该检验四个维度：

- **复制成本比**：他方重建一个替代位置所需的全部成本，与企业年收费额之间的比值。比值越高，收费权越强。英伟达的竞争对手需要重建的，不只是芯片设计，还包括 CoWoS 产能、HBM 供给和 CUDA 生态。
- **供给弹性**：价格上涨后，替代供给能够多快出现。纯算力租赁的经营利润率在一年内从 17% 跌到 1%，就是供给弹性过高的结果。
- **转换成本**：客户离开时，除了更换供应商，还要重写多少自身流程和数据。这一项应以真实迁移报价和第三方审计为准，而不是以厂商自述为准。
- **共同专用化**：客户是否围绕供应商做过只对该供应商有用的投资，例如本体建模、权限体系和员工培训。如果有，收费权就部分“长在客户身上”；如果没有，供应商随时可能被拔掉。

四个维度最后落到一个判断：

不要只看利润率有多高，要看利润能够维持多久。

高毛利只是一张照片，收费权则是一段录像。应观察的，不是企业此刻赚得多不多，而是被竞争者围攻十个季度之后，还能留下多少。这把尺适合衡量一家公司的存量；如果要判断一个新机会，则可以使用第十一章的五问框架。至于资本市场如何为三类收费权整体定价（例如 SpaceX、Anthropic 和 OpenAI 三家未上市公司的合计隐含市值约 3.63 万亿美元，约 54% 的基金经理认为 AI 已进入泡沫区域），这将是系列第二篇《泡沫诊断》的主题。本篇只保留一个收费权视角的判断：公开市场最终会奖励那些占据难以绕开的位置，并能把这一位置转化为可审计经营结果的公司。至于当前估值是否已经透支，需要留到第二篇对账。

数据注：Palantir 2026 年第一季度美国商业收入 5.95 亿美元，同比增长 133%；美国政府收入 6.87 亿美元，同比增长 84%；GAAP 净利润 8.71 亿美元；公司据此将全年收入指引上调至约 76.5 亿美元。

下部 中国与判别

十、中国的位置：政策降低进入成本，不替代收费权

中国主体在这条产业链上，多了一层政策与算力环境的支撑。它不会改变“价值集中于最后一层”的主线，但确实能够降低进入门槛。国家层面已经把“应用”直接写入政策目标。2025 年 8 月，国务院印发《关于深入实施“人工智能+”行动的意见》，提出到 2027 年，新一代智能终端、智能体等应用普及率超过 70%；到 2030 年超过 90%，并部署六大行动和八项基础支撑。

算力底座也在持续铺开。“东数西算”布局八个枢纽、十个集群，将东部需求引向西部绿电；国家超算互联网在 2026 年初用户突破百万。据国家数据局，截至 2025 年年中，全国在用标准机架超过千万架，智能算力达到数百 EFLOPS，累计建成高质量数据集超过 3.5 万个。这些数字的意义，不限于规模变大，还在于算力开始从“大厂专属”走向“公共可及”。

在地方层面，主要政策工具变成了各种“券”。横向比较，深圳的体系相对完整，算力券、训力券、模型券和语料券四类并设。训力券补贴比例可达 50%，初创企业可达 60%，单家企业每年最高 1000 万

元，全市资金盘约 45 亿元。杭州、北京和上海也设有相应补贴，但比例和单家上限相对较低。需求侧的底盘同样在扩大。据 CNNIC，截至 2025 年底，中国生成式 AI 用户达到 6.02 亿，普及率为 42.8%，一年增长 141.7%，九成以上用户优先使用国产大模型。这些红利能否直接转化为收费权？答案是否定的。

政策降低的是“进入成本”：它让算力更便宜，让训练和部署门槛更低，却不能替代价值捕获本身。算力券可以帮助一家公司把模型跑起来，但无法替它获得别人拿不到的行业数据、客户信任和工程交付能力。政策红利更合理的用法，是用它节省基础成本，再把省下来的资源投入“最后一层”。

政策可以帮助企业进场，却不能替企业建立收费权。



图 8 政策降低进入成本，但不替代收费权

数据注：各地补贴口径——杭州算力券补贴比例约 30%，单家企业最高 800 万元；北京约 20%，累计最高 200 万元；上海最高约 20%。

十一、五个问题：判别真假机会

全文可以压缩成一套可以直接使用的判断框架。无论是自己做项目、投资一家公司，还是替政府遴选 AI 项目，都可以先问五个问题。

第一问：它卖的是结果，还是工具？

出售结果并承担责任，本身就可能构成壁垒。Sierra 按成功解决的会话计费，中控按照工艺环节产生的收益交付，供应商必须为结果负责。只提供工具的产品，则随时可能被平台内置或被竞争者替代。

第二问：它在规模化时，毛利究竟是正还是负？

记住两组对照：智谱本地部署毛利率约 66%，云端 API 约 -0.4%；“超新星”型 AI 公司第一年可以做到约 4000 万美元 ARR，毛利率却只有约 25%。增长无法修复亏损结构，只会把亏损放大。

第三问：客户留存能否被第三方验证？

11x 自报年化收入接近 1000 万美元，外部核查后实际约为 300 万美元，客户流失率达到七至八成。凡是只能依靠公司自报的数据，都应当先打折。

第四问：壁垒来自哪里？

它依靠的是牌照、专有数据、工作流和信任，还是一层可以随时替换的封装？前者难以被巨头快速复制，后者可能被巨头顺手内置。

第五问：它能否运行在国产算力上？

这决定项目能否享受政策红利，也决定它在出口管制和供应波动中能否生存。五个问题最终指向同一个方向：

机会集中在难以复制的“最后一层”，风险集中在那些可以被迅速抹平的部分。



图 9 五道闸门：判别真假机会

结语

现在，开头的谜题可以合上了。同一条产业链上，约 75% 的毛利属于难以复制的制造，约 4.8% 属于可以标准化的组装，而 -0.4% 则属于可以被无限复制的接口。利润并不按照“离 AI 有多近”来分配，而是按照“有多难被复制”来分配。凡是能够被系统化、无限复制，或者直接由模型产出的部分，都在趋于贬值；价值则不断向“最后一层”迁移，并具体表现为资源、入口和流程三类收费权。对产业而言，它叫“最后一层”；对资本市场而言，它叫“收费位置”；落到个人身上，它还会有第三个名字，留待第三篇

展开。不过，在进入个人与劳动之前，还有一个更大的问题需要回答：涌入这条产业链的钱，本身是否合理？当市值越来越集中在少数公司，当产业链核心厂商互为客户，当万亿美元承诺押注的是 2030 年代的需求，价格是否已经跑在现金流前面？这将是系列第二篇《泡沫诊断》的任务。

参考文献

引子：IBM 与微软授权史

- Gary Kildall, 《Computer Connections》回忆录（家属公开节选，2016；对“飞行缺席”版本有异议）：<https://computerhistory.org/blog/computer-history-museum-license-agreement-for-the-kildall-manuscript/>
- Tim Paterson 关于 86-DOS/QDOS 的自述与公开访谈（约 7.5 万美元买断一节，回源核对具体金额口径）
- 计算机史通行记述见 Freiburger & Swaine, 《Fire in the Valley》（IBM-微软-DRI 一节；“授权而不卖断”为公开史实）

算力与芯片

- Epoch AI, 全球算力份额与保有量：<https://epoch.ai/data-insights/hyperscalers-control-most-compute>
- NVIDIA, FY2026 四季度及全年业绩：<https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-fourth-quarter-and-fiscal-2026>
- CoreWeave SEC 8-K（纯租赁经营利润率 17%→1%）：<https://www.sec.gov/Archives/edgar/data/1769628/000176962826000220/coreweave1q26earningspress.htm>
- SemiAnalysis, 华为 CloudMatrix 384：<https://newsletter.semianalysis.com/p/huawei-ai-cloudmatrix-384-chinas-answer-to-nvidia-gb200-nvl72>
- 科技日报, 智算中心利用率约三成：https://www.stdaily.com/web/gdxw/2025-07/17/content_371093.html
- 中际旭创 / 光模块全球份额（公开财报与行业统计）：<https://www.innolight.com/>

模型与推理基础设施

- a16z, LLMflation（推理成本每年降约十倍）：<https://a16z.com/llmflation-llm-inference-cost/>
- Epoch AI, 推理价格趋势：<https://epoch.ai/data-insights/llm-inference-price-trends>
- IDC / 量子位, 2025H1 大模型调用量 536.7 万亿、火山引擎 49.2% 第一：<https://www.qbitai.com/2025/09/334030.html>
- Meta 投资 Scale AI（约 143 亿美元 / 约 49%）：<https://www.cnbc.com/2025/06/13/meta-scale-ai-alexandr-wang.html>
- CNBC, Mercor 估值达 100 亿美元：<https://www.cnbc.com/2025/10/27/ai-hiring-startup-mercior-funding.html>

AI 原生软件

- Menlo Ventures, 2025 企业生成式 AI 现状：<https://menlovc.com/perspective/2025-the-state-of-generative-ai-in-the-enterprise/>
- a16z, AI Will Eat Application Software：<https://a16z.com/good-news-ai-will-eat-application-software/>
- 红杉《Generative AI's Act o1》：<https://www.sequoiacap.com/article/generative-ais-act-o1/>
- Bessemer《State of AI 2025》：<https://www.bvp.com/atlas/the-state-of-ai-2025>
- ICONIQ Growth, 2026 State of AI（增长/留存/定价，行业口径）：<https://www.iconiq.com/growth/reports/2026-state-of-ai-bi-annual-snapshot>
- CNBC, Harvey 融资至 110 亿美元估值：<https://www.cnbc.com/2026/03/25/legal-ai-startup-harvey-raises-200-million-at-11-billion-valuation.html>
- CNBC, Sierra 融资至 158 亿美元估值：<https://www.cnbc.com/2026/05/04/bret-taylor-sierra-fundraise-openai.html>
- CNBC, OpenEvidence 估值翻倍至 120 亿美元：<https://www.cnbc.com/2026/01/21/openevidence-chatgpt-for-doctors-doubles-valuation-to-12-billion.html>
- TechCrunch, 11x 夸大客户与收入的争议：<https://techcrunch.com/2025/03/24/a16z-and-benchmark-backed-11x-has-been-claiming-customers-it-doesnt-have/>
- CNBC, Anysphere (Cursor) 估值约 293 亿美元：<https://www.cnbc.com/2025/11/13/anysphere-cursor-funding.html>

- TechCrunch, Cognition 融资至约 260 亿美元估值: <https://techcrunch.com/2026/05/27/ai-coding-startup-cognition-raises-1b-at-25b-pre-money-valuation/>
- a16z, Top 100 Gen AI Consumer Apps (第 6 版): <https://a16z.com/100-gen-ai-apps-6/>
- 快手 IR, 可灵 AI 年化收入达 2.4 亿美元: <https://ir.kuaishou.com/news-releases/news-release-details/kling-ai-annualized-revenue-run-rate-hits-usd240-million/>
- TechCrunch, 中国否决 Meta 收购 Manus: <https://techcrunch.com/2026/04/27/china-vetoes-metas-2b-manus-deal-after-months-long-probe/>

垂直行业大模型

- 证券时报, 中控技术 TPT (乙烯裂解 99.79%): <https://www.stcn.com/article/detail/3312909.html>
- Nature, 盘古气象大模型: <https://www.nature.com/articles/s41586-023-06185-3>
- 中国经营报, 智谱招股书毛利 (本地 66% / 云端 -0.4%): <https://finance.sina.com.cn/jjxw/2025-12-22/doc-inhcsexh9214350.shtml>
- 微软亚洲研究院, Industrial Foundation Models: <https://www.microsoft.com/en-us/research/articles/towards-industrial-foundation-models-integrating-large-language-models-with-industrial-data-intelligence/>

超级个体

- Carta, 单人创始报告 (36.3%): <https://carta.com/data/solo-founders-report/>
- Inc., Amodi 预测一人独角兽: <https://www.inc.com/ben-sherry/anthropic-ceo-dario-amodei-predicts-the-first-billion-dollar-solopreneur-by-2026/91193609>
- TechCrunch, Base44 六个月被 Wix 以 8000 万美元收购: <https://techcrunch.com/2025/06/18/6-month-old-solo-owned-vibe-coder-base44-sells-to-wix-for-80m-cash/>
- gstack (GitHub): <https://github.com/garrytan/gstack>
- Anthropic 《2026 Agentic Coding Trends》: <https://resources.anthropic.com/hubfs/2026%20Agentic%20Coding%20Trends%20Report.pdf>
- NBER, Generative AI at Work (呼叫中心 +14%、新手 +34%): <https://www.nber.org/papers/w31161>
- Microsoft Research, GitHub Copilot 生产率实验: <https://www.microsoft.com/en-us/research/publication/the-impact-of-ai-on-developer-productivity-evidence-from-github-copilot/>
- FDA 对 Medvi 的警告信: <https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/warning-letters/medvi-llc-dba-medvi-721455-02202026>

价值捕获 · 收费权 · 政策

- Palantir, 2026 年一季度业绩: <https://www.sec.gov/Archives/edgar/data/0001321655/000132165526000026/a2026q1ex991pressrelease.htm>
- Palantir, 2025 年四季度及全年业绩: <https://www.sec.gov/Archives/edgar/data/1321655/000132165526000004/a2025q4ex991earningsrelease.htm>
- 国务院《关于深入实施“人工智能+”行动的意见》: https://www.gov.cn/zhengce/content/202508/content_7037861.htm
- 工信部等八部门《“人工智能+制造”专项行动实施意见》: https://www.nda.gov.cn/sjj/zwgk/zcfb/0112/20260107214358696030895_pc.html
- 深圳市加快打造人工智能先锋城市行动计划 (训力券、45 亿资金盘): http://gxj.sz.gov.cn/gkmlpt/content/12/12645/post_12645602.html
- CNNIC 第 57 次报告, 生成式 AI 用户 6.02 亿、普及率 42.8%: <https://www.news.cn/tech/20260302/66c4ab06b6f34f8d806b416b3acc9f0b/c.html>

理论框架与配套论文

- David Teece, 《Profiting from Technological Innovation》, Research Policy 15(6), 1986 (互补资产与共同专用化投资, 第九章四测度的理论来源之一)

数据口径与使用说明：本篇为公开信息的整理与分析，不构成投资建议。各公司 ARR、估值多为公司自报或第三方估算、未经审计，2025 – 2026 年变动较快，文中已尽量标注时点与口径；英伟达 71% – 75%、智谱本地 66%、服务器整机 4.8% 为毛利率口径，CoreWeave 17%→1% 为经营利润率口径，比较时注意口径差异；第四章的增长率、净收入留存、人均 ARR 等指标整理自行业研究报告（口径不一），正式发表前需回源核实；11x 自报数据存在已披露争议；Manus 交易已被否决并取消；"工业大模型渗透率 9.6%→47.5%"未能回源官方，本篇改用可回源口径。正式引用请回溯原始来源。

02

泡沫诊断

价格是否跑在现金流之前

技术革命可以成立，资产价格仍可能失败

质观工作室 · "AI 时代的价值分布"三部曲之二 (2026)

摘要：从金融意义上界定泡沫：判断重点不在技术是否有用，而在资产价格是否超出未来现金流能够解释的范围。

文章先用历史建立参照。郁金香狂热只有价格上涨，没有能力进步、规模采用和现金流，是纯粹的价格泡沫；石油、铁路、电力与互联网则提供相反样本：技术最终都改变了世界，但早期资产价格和融资结构仍然失败。用同一框架观察 AI：能力在进步，采用在扩大，部分企业已经产生真实现金流；价格却在提前透支多年后的需求。由此看，AI 并非一个整体泡沫，更接近一场真实技术革命中的资产泡沫。风险主要集中在五类资产：低利用率算力、缺少收费权的模型估值、没有进入工作流的应用、被过度包装的 Agent，以及无法产生可衡量回报的企业 AI 项目。

文章还讨论泡沫的系统性结构：市值向少数 AI 公司集中，产业链核心厂商又通过投资、采购和长期承诺彼此连接，单点错配正在变成“铁索连环”。附论以郁金香与比特币作对照，说明没有现金流的资产如何定价，以及承诺现金流的 AI 资产为什么最终必须接受现金流检验。

关键词：人工智能；资产泡沫；现金流；收费权；市值集中；循环交易；郁金香狂热；比特币；资本开支；Agent

本篇术语说明

- **token（词元）：**大模型处理和生成文本的最小计费单位；模型每读入或产出文字，都按 token 计数与计费。
- **推理（inference）：**模型训练完成后实际运行、对外给出回答的过程；每次调用都会产生“推理成本”。
- **Agent（智能体）：**能自主拆解目标、调用工具、连续执行多步任务的 AI 程序，而非只做一问一答。
- **PoC（概念验证）：**在正式采购或部署前，用小规模试点验证技术可行性的项目阶段。
- **SaaS（软件即服务）：**通过网络订阅使用、按期付费的软件。
- **毛利率：** $(\text{收入} - \text{直接成本}) \div \text{收入}$ ，衡量单位业务的盈利空间。
- **自由现金流：**经营现金流扣除资本开支后的余额，衡量企业真实的造血能力。
- **NDR / NRR（净收入留存）：**同一批老客户一年后贡献收入的比例，高于 100% 表示老客户仍在扩容。
- **backlog（未交付订单）：**已签约、尚未确认为收入的合同金额。
- **收费权：**本系列提出的概括性概念，指在产业链中占据他方难以绕开、并能持续定价的位置；其完整分类（资源、入口、流程三类）已在系列第一篇界定，本篇沿用。

一、先把问题说清：技术有效，不等于资产值钱

讨论 AI 泡沫时，容易把“泡沫”等同于“技术没有用”。这混淆了两个完全不同的问题。金融意义上的泡沫，是资产价格超过了未来现金流能够解释的范围。旧金山联储对资产泡沫的经典表述也指向这一点：当价格显著脱离经济基本面，尤其超出未来现金流折现能够支撑的水平，泡沫就已经形成。所以，本篇所说的泡沫，指的是“价格与现金流脱节”，不是对技术有效性的否定。全文从这里展开。现金流来自本系列所说的“收费权”：在产业链上占据别人难以绕开、自己能够持续定价的位置。判断一项 AI 资产是否处于泡沫中，应当追问：它所在的位置能不能长期收费，而不能只问技术有没有用。

技术有效性与资产收费能力，是两个独立命题。把二者混为一谈，正是泡沫定价常见的认知基础。系列第一篇讨论价值停在哪些位置；这一篇则继续追问：资本为这些位置开出的价格，是否合理。一定要区分这两件事，是因为历史上多数重大泡沫并非围绕伪技术发生，而是围绕真实技术发生。石油、铁路、电力和互联网都改变了现代文明，却都在早期出现过严重的资产错配。投资者往往看对了技术方向，却看错了三件事：谁能赚钱、什么时候赚钱、能赚多久。在这四个案例之前，历史还提供过一个更纯粹的样本：一场与任何技术无关的泡沫。两类样本合在一起，构成本文的参照系。

二、历史参照：技术可以成功，投资仍然可以失败

（一）对照组：郁金香，酒馆里的风的交易

1630 年代的荷兰是当时人均最富裕的经济体。阿姆斯特丹已建立世界最早的证券交易所，东印度公司股票在其中规范交易。但郁金香交易不走交易所，而是在哈勒姆、阿姆斯特丹、阿尔克马尔的酒馆里进行。交易者结成所谓“学院”（college），在烟雾与酒气中竞价球茎，买方按惯例支付一笔“酒钱”（wijnkoopsgeld），成交即记入私人账目。交割方式更能解释这场交易的形态：交易最狂热的冬季，球茎尚埋在土中无法起挖，买卖的实际是来年春季的交割凭证。同一颗球茎的凭证可在起挖之前转手十余次，时人称之为“风的交易”（windhandel）。

由于球茎不可见，交易依据的是画师绘制的郁金香图册，功能上接近招股说明书。整条链条建立在私人信用之上，没有保证金，没有清算机构，法院此后亦拒绝将此类合约作为可强制执行的债务。最高价格还依赖一个今天看来颇具讽刺意味的生物学事实。最昂贵的品种是带火焰状条纹的“碎色”郁金香，以 *Semper Augustus* 为最：其时报价数千至上万荷兰盾一颗。作为参照，阿姆斯特丹运河边一栋体面宅邸约值一万盾，熟练工匠年收入约三百盾。

二十世纪的植物学研究表明，碎色条纹由郁金香碎色病毒（一种花叶病毒）感染所致：病毒制造了条纹，也削弱了球茎、使其难以稳定繁殖。支撑最高价格的稀缺性，在生物学上是一种病。至 1636 年末，上涨已不限于珍稀品种，按磅计价的普通球茎亦成倍上涨；行情蔓延到毫无稀缺性的品类时，通常

已近末端。1637 年 2 月初，哈勒姆的一场拍卖无人应价（其时该城正流行瘟疫），恐慌沿借据链在数日至数周内传遍各城。善后方式同样能说明这场泡沫的性质：各城当局先行冻结合约，哈勒姆最终裁定买方可按合约价约 3.5% 了结。

现代档案研究（Goldgar, 2007）指出，真实破产者寥寥，荷兰宏观经济几乎未受损伤。这场狂热主要在商人与工匠的社交网络内展开，其文化冲击（讽刺小册子、猴子炒作郁金香的油画）远大于经济损失。另一个常被忽略的事实是：交易崩溃后，花卉贸易没有随之消失。荷兰此后数百年始终是全球花卉贸易的中心，至今仍是最大的花卉出口国。

郁金香为本文提供了一个最纯粹的对照样本。按照下一章的四线框架衡量：

- 没有能力线：球茎本身不会持续进步；
- 没有采用线：用途始终只是观赏；
- 没有现金流线：持有球茎不会产生收入；
- 只有价格线在独自上涨。

它是"没有技术的泡沫"的极端样本。随后四个案例则相反：技术全部成立，定价全部失败。AI 属于哪一类，需要用这套框架检验。

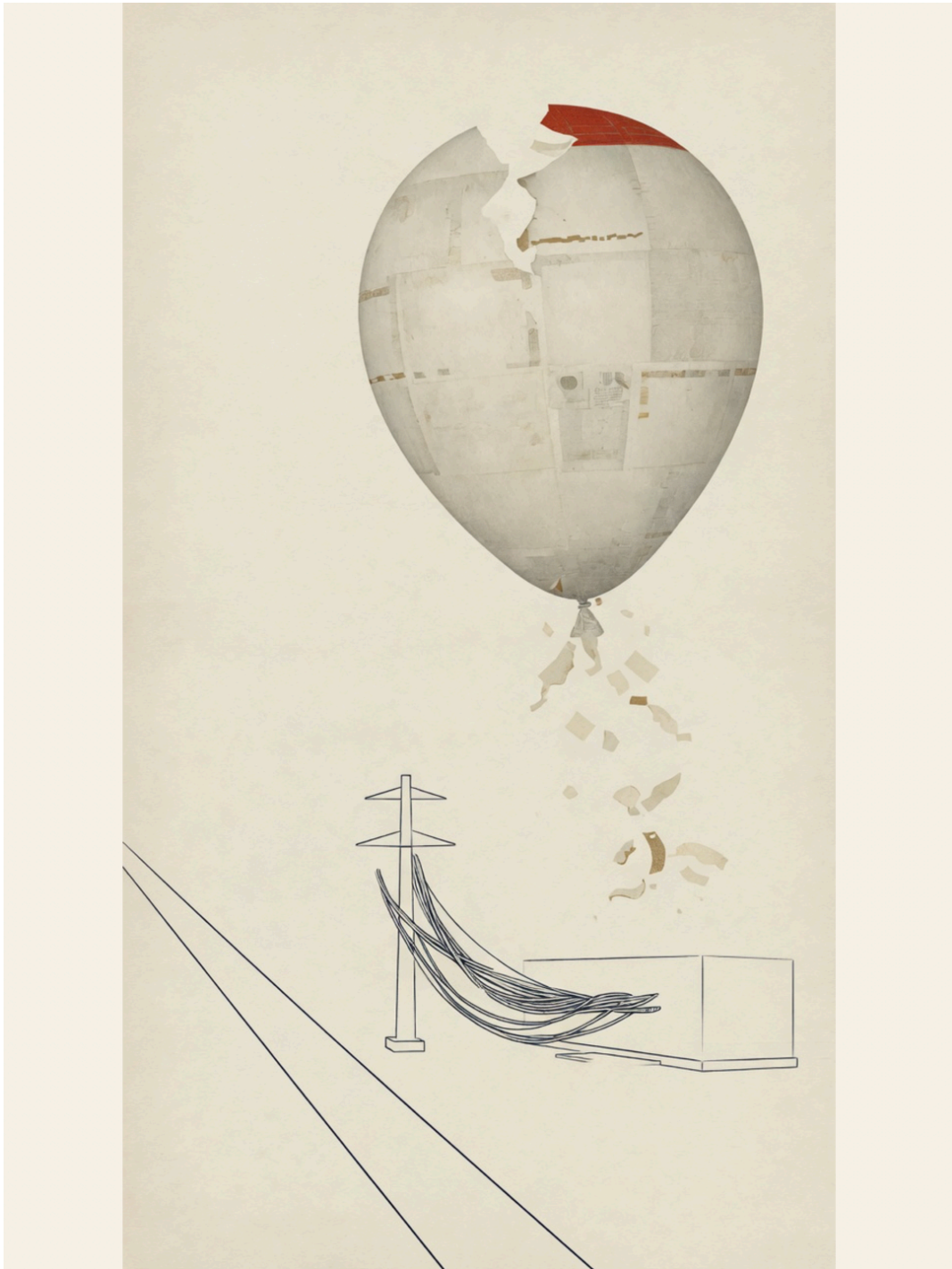


图 1 技术革命可以成立，价格泡沫仍会破裂

（二）四次被错误定价的技术

石油时代的早期油潮提供了第一个案例。宾夕法尼亚的 Pithole 在 1865 年成为美国早期石油狂热的中心之一，当年秋天人口一度膨胀至约 1.5 万人，日产油量超过 5000 桶。油价则从每桶约 8 美元跌至约 2.5 美元。到 1866 年 12 月，当地人口已不足 2000 人。1870 年秋天，只剩下 43 户居民和一家旅馆。这个小镇的资产故事迅速破灭，但石油运输和基础设施能力并未消失。

Van Syckel 输油管道于 1865 年投入使用，每小时可输送约 80 桶油，相当于 10 小时内替代 300 辆马车运输。油镇投机被泡沫清算，石油工业所需的基础设施却保留下来。再看铁路。这场泡沫更接近现代资本市场的形态：英国 1840 年代铁路热中，低利率、铁路收入增长、建设成本下降与中产投资者入场共同推高铁路股价。

根据纽约联储对 Railway Mania 的回顾，1843 至 1845 年间上市铁路证券价格大幅上涨，部分口径下接近三倍。剑桥大学相关研究则指出，1843 年至 1845 年秋，铁路股价超过翻倍，并有 562 个新铁路项目被提交至议会。大量项目依赖分期缴款股份与杠杆，部分线路在尚未获得批准前便已交易。泡沫最终破裂，但英国铁路网络并未因此失去意义。一批过度融资、过度承诺的线路被清算之后，铁路才成为工业社会的基础设施。

电力的教训集中于金融结构。1920 年代美国公用事业公司大量通过复杂控股结构扩张，形成高度杠杆化、层层嵌套、信息不透明的金融体系，Samuel Insull 的公用事业帝国是其中代表。大萧条之后，这类结构暴露出过度资本化、证券发行基础不实、关联交易与监管规避等问题。1935 年，美国通过《Public Utility Holding Company Act》，目标直指拆解和简化公用事业控股公司体系。但电力本身没有失败。EIA 数据显示，1932 年美国农场只有约 11% 通电，到 1942 年约一半通电，到 1950 年美国人几乎都已用上电。

清算落在金融结构上，电气化继续推进。互联网泡沫最为人熟知。Goldman Sachs 对 dot-com 泡沫的回顾显示，纳斯达克综合指数 1999 年上涨 86%，2000 年 3 月 10 日收于 5048 点附近，此后至 2002 年 10 月下跌约 77%。大量公司没有收入、没有利润，却被按未来网络效应定价。但互联网本身没有消失。Our World in Data 数据显示，1997 年全球只有约 2% 人口使用互联网，2019 年已超过一半，2023 年达到 67%。

美国电商占零售总额的比例，到 2026 年一季度已达 16.9%。市场否定了一批公司，没有否定互联网本身。四段历史指向同一个结论：泡沫破裂否定的是资产价格、融资结构和错误的公司，不是技术方向本身。Carlota Perez 在技术革命与金融资本周期研究中提出过相应框架。新技术进入安装期时，金融资本往往先于生产资本过度扩张，既推动基础设施建设，也制造资产泡沫。泡沫破裂后，若制度、企业组织和产业应用跟上，技术才进入更长的部署期。讨论 AI 泡沫，不应从“AI 有没有用”开始。石油、铁路、电力和互联网都极其有用，但“有用”并不能自动证明“价格合理”。要问的是：今天被高价买入的公司、园区、模型、应用和故事，未来能否形成真实现金流，能否守住收费权。泡沫破裂后能留下来的，是基础设施、制度、流程，以及重新划分后的价值归属。

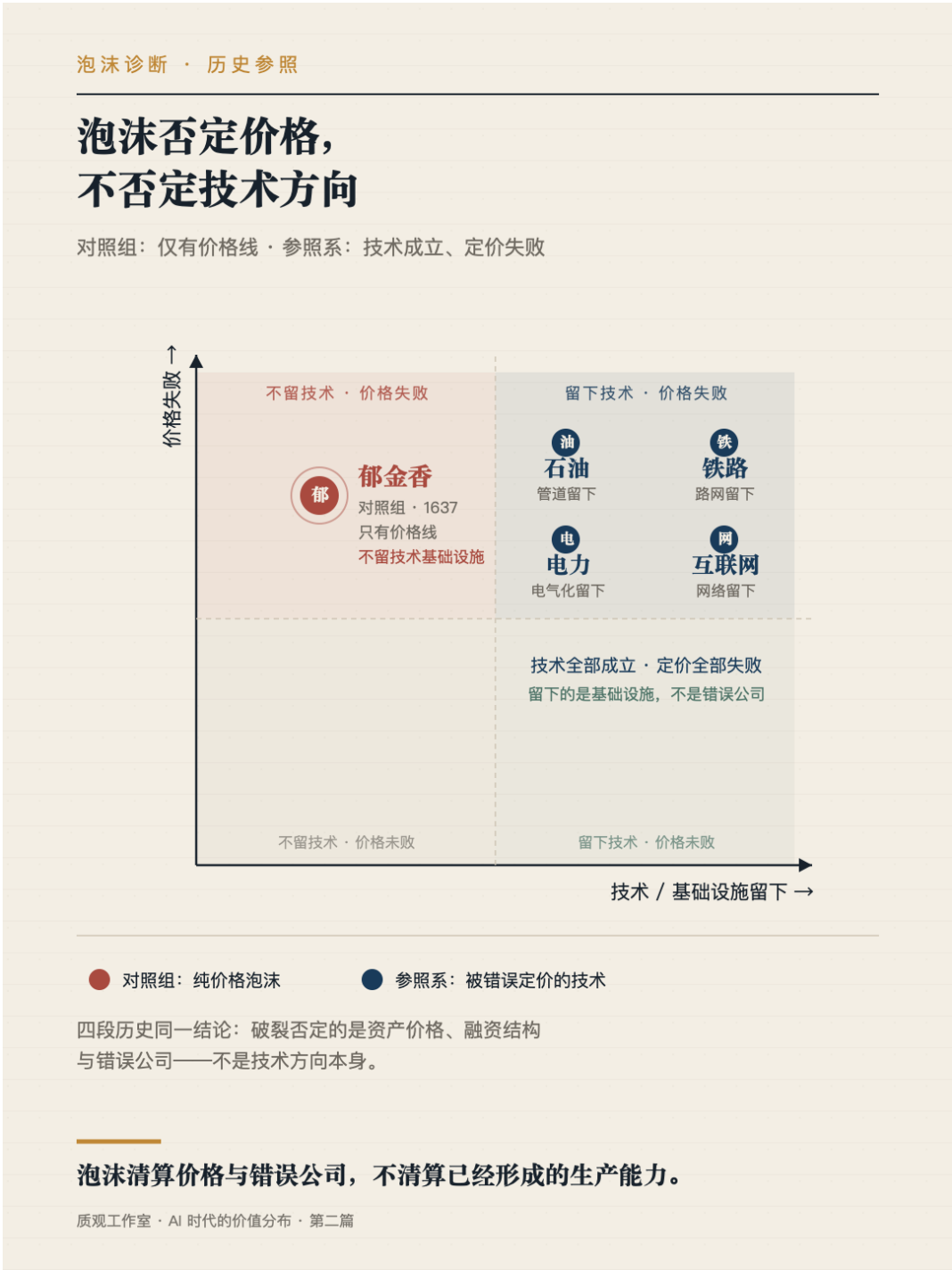


图 2 历史泡沫：价格失败，技术是否留下

三、四线诊断：技术走了多远，价格又跑了多快

判断 AI 是否存在泡沫，既不能只看模型演示，也不能只看融资金额。更稳妥的方法，是同时观察四条线：能力、采用、现金流和价格。最危险的情况，往往发生在技术持续进步、价格线却远快于现金流时。这一框架也可回头校验第二章的两类样本：郁金香四条线中仅价格线运行，属纯价格泡沫；四次技术泡沫则是四条线严重不同步。

（一）能力线：技术真的在进步吗？

能力线回答的是：技术有没有发生真实进步。这条线无法支持“AI 全是泡沫”的判断。OpenAI 官方价格页列示，GPT-5.5 高端档的标准短上下文价格为每百万 token 输入 5 美元、输出 30 美元；Batch 与 Flex 可降至输入 2.5 美元、输出 15 美元。更便宜的 GPT-5.4、GPT-5.4 mini 和 nano，则形成了多个价格档位。高端智能仍然昂贵；但同一能力体系下已经出现明显的价格分层。价格分层更像供给成熟后的现象，而不是单纯叙事。

比单个模型价格更能说明问题的，是固定能力水平下的推理成本下降。Epoch AI 对六类 benchmark 的研究显示，达到 GPT-4 水平的博士级科学问答任务，其推理价格约以每年 40 倍的速度下降。不同任务与阈值的下降速度差异很大，区间可达每年 9 倍至 900 倍，中位数约为每年 50 倍。研究者同时提醒，最快的降价集中于最近一年，未必长期持续。即便按最保守口径处理，也可以得到一个判断：智能的生产成本正在快速下降。

Stanford AI Index 2026 对能力线的描述同样清楚：AI 能力没有停滞，模型在博士级科学问题、多模态推理、竞赛数学和代码 benchmark 上继续提升。SWE-bench Verified 一类编码任务的表现，一年内从 60% 提升到接近 100%。但同一报告也指出，模型能力呈现明显的“锯齿边界”：AI 可以赢得国际数学奥林匹克金牌，却未必能稳定读出指针时钟。OSWorld 一类真实计算机任务中，agent 成功率虽明显提升，仍约有三分之一的尝试失败。由此看，能力进步是真实的，但并不均匀。AI 在某些高难任务上已接近顶尖水平，在另一些看似简单的现实任务中却仍会失败。这种“锯齿状边界”，决定了商业化时不能只看最强样例。

（二）采用线：AI 真的进入生产了吗？

采用线回答的是：AI 是否已经从尝鲜走进真实生产。McKinsey 2025 年全球调查显示，88% 的受访组织已经在至少一个业务职能中规律使用 AI，高于前一年的 78%。但开始规模化部署的只有约三分之一，多数组织仍停留在实验或试点阶段。Agent 的采用呈现同一结构：23% 的组织称正在某处规模化 agentic AI，另有 39% 处于实验阶段。但在任何单一业务职能中，规模化比例都不超过 10%。

大厂财报提供了更硬的采用证据。Microsoft FY26 Q3 披露，其 AI business 年化收入运行率超过 370 亿美元，同比增长 123%；Microsoft Cloud 收入超过 540 亿美元，同比增长 29%。Palantir 2026 年一季度收入 16.3 亿美元，同比增长 85%，美国商业收入同比增长 133%。调整后自由现金流 9.25 亿美元，自由现金流率 57%，NDR 达 150%。这类数据表明，AI 已不只是消费级聊天工具，它已进入云、办公、开发和企业流程。

不过，采用线也有警告的一面。Gartner 2024 年预测，至少 30% 的生成式 AI 项目会在 2025 年底前止步于概念验证。2025 年又预测，超过 40% 的 agentic AI 项目会在 2027 年底前取消，原因包括成本上升、商业价值不清和风险控制不足。“开始使用 AI”，与“业务流程已被 AI 重写”，并不是同一件事。泡

沫最容易发生在这里：把试用率当成留存率，把 PoC 当成收入，把一次漂亮的演示当成已经完成的组织改造。

（三）现金流线：收入能覆盖投入吗？

现金流线回答的是：收入、毛利和自由现金流，能否跟上快速增长的资本开支。这条线已经出现明显分化。一边是上游。NVIDIA 2027 财年一季度收入 816 亿美元，同比增长 85%；Data Center 收入 752 亿美元，同比增长 92%，GAAP 毛利率 74.9%。Broadcom 2026 财年二季度 AI 半导体收入 108 亿美元，同比增长 143%，自由现金流 102.62 亿美元，占收入 46%。芯片、网络 and 定制加速器已经形成巨大现金流，AI 基础设施并非虚假需求。

另一侧是云厂商与重资产算力平台。Amazon 2026 年一季度披露，AWS 销售额 376 亿美元，同比增长 28%。但过去 12 个月自由现金流从 259 亿美元降至 12 亿美元，主要因为物业与设备采购增加 593 亿美元，公司明确表示这主要反映 AI 投资。Meta 将全年资本开支指引上调至 1250 亿至 1450 亿美元；CoreWeave 同季收入 20.78 亿美元、backlog 994 亿美元，但净亏损 7.4 亿美元、利息费用 5.36 亿美元。需求可以真实，回本周期仍要单独计算。

（四）价格线：资本提前买了多少未来？

价格线回答的是：融资、估值和市场预期，是否跑得太快。OECD 统计显示，2025 年 AI 公司获得了全球 VC 投资的 61%，金额达 2587 亿美元，占全部 VC 投资 4271 亿美元的一半以上。相比 2022 年的 30%，这一比例已经翻倍。资本还高度集中于大额交易：1 亿美元以上的 mega deals 占 2025 年 AI VC 投资价值约 73%，10 亿美元以上交易约占总价值近一半。IT infrastructure and hosting 类 AI 公司吸引 1093 亿美元，超过其他行业合计的三分之二。以上为一级市场口径。价格线在二级市场与产业链内部的两种系统性形态（市值集中与循环交易），见第五章。

（五）把四条线放在一起

把四条线放在一起看，结论比简单的“有泡沫”或“没泡沫”更清楚：

- 能力线是真实的；
- 采用线正在扩大；
- 一部分现金流已经兑现；
- 价格线却在提前资本化未来多年的需求。

因此，AI 并非整体泡沫。泡沫集中在那些把“AI 有用”直接等同于“我的资产能持续收费”的地方。问题可以概括为：把技术有效性，误当成了收费权。

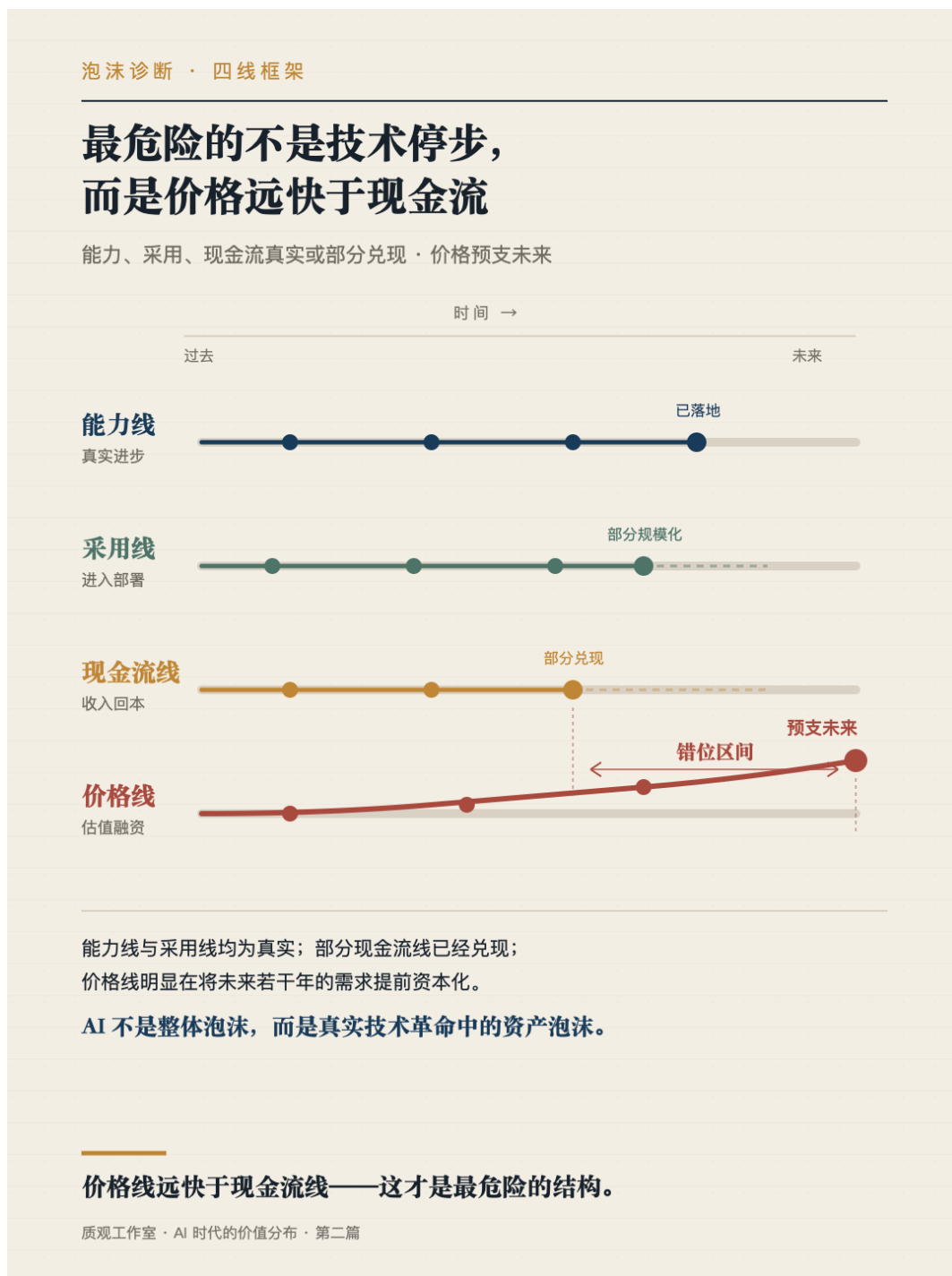


图 3 四线诊断：价格线为何跑在现金流之前

四、泡沫藏在哪里：五类高风险资产

AI 泡沫并不均匀。风险主要集中在五类资产：低利用率算力、缺少收费权的模型估值、没有进入流程的 AI 应用、被过度包装的 Agent 项目，以及无法产生可衡量回报的企业 AI 预算。



图 4 五类高风险资产及共同病灶

（一）算力囤积：有 GPU，不等于有生意

从 NVIDIA 与 Broadcom 的数据看，上游芯片、网络和先进封装不是纯粹叙事，主要压力在下游。下游算力要形成收费能力，至少需要四个条件：长期客户、电力保障、高利用率和较低的融资成本。缺少这些条件，数据中心建得越大，风险反而越大。Amazon、Meta、Alphabet、Microsoft 的资本开支均在显著上升。这些大厂自身拥有流量、云客户、模型和应用场景，也需要承受自由现金流压力。不具备这些入口的重资产算力平台，风险更高。

能源约束会放大这一风险。IEA 预计，全球数据中心用电量到 2030 年将达到约 945 TWh，接近当前水平的两倍，占全球用电约 3%。其中 AI 驱动的 accelerated servers 用电增长更快，年增速约 30%。EPRI 对美国的估算更为激进，认为数据中心到 2030 年可能占美国发电量的 9% 至 17%。算力扩建因此不只是购置 GPU，还须锁定电力、变电容量、液冷、园区许可和网络。只具备融资故事、拿不出电力合同和客户合同的算力项目，是这一轮热潮中最接近泡沫的部分。

（二）模型估值：能力在上升，收费权在收缩

模型能力不断进步，模型本身的收费权却在被压缩。两条曲线方向相反。OpenAI 官方价格页显示，旗舰模型和 mini/nano 模型已经形成多档价格；Epoch AI 从固定能力水平测算出推理价格快速下降。Stanford AI Index 2026 还显示，中美顶级模型能力差距已大幅收窄，自 2025 年初以来双方多次交换领先地位。模型层并非没有价值，问题是价值持续受到降价、开源、多模型路由、云厂补贴和企业议价的挤压。只有模型本身、没有入口、生态、企业关系和责任承担时，估值便容易跑在收费权前面；泡沫的定义在这里可以直接适用。

（三）AI Wrapper：封装不是问题，没有壁垒才是

对“套壳”不宜一概嘲笑。许多重要软件最初也只是把底层能力封装成可用产品。分界在于：应用是否拥有专有数据、工作流入口、客户关系、行业交付和责任承担。若只是调用通用模型、包一层界面、把原有任务改名为 AI 任务，那么模型降价、平台内置或大厂复制中的任何一项都会迅速侵蚀其价值。Cursor、Replit 等开发工具在定价调整中反复提示了一个约束：复杂 AI 任务的 token 与推理成本差异极大。AI 应用并不天然享有传统 SaaS 的毛利。能否把成本转嫁为按结果收费，是这类应用能否穿越泡沫的关键。

（四）Agent 叙事：能走下一步，不等于能接管全流程

Agent 是真实方向，但大量项目把“能够演示下一步”提前包装成“能够托管整条业务链”。Gartner 对 agentic AI 项目取消率的预测，反映的就是这道落差：成本、商业价值和风险控制还没有匹配。可收费的 Agent 必须被限制在有边界、有权限、有日志、有回滚、有责任人的流程中。没有这些约束，Agent 只会把单步错误累积为复合错误，把演示中的表现变成生产中的事故。

（五）企业 AI ROI：买到模型，不等于买到回报

McKinsey 的 88% 采用率显示，企业已经普遍开始使用 AI；但同一调查也显示，多数企业仍未把 AI 深度嵌入流程。Gartner 对 PoC 失败率的判断也指向同一个问题：企业买到了模型，不等于买到了回报。回报主要来自模型之外：清理数据、重写流程、调整权限、划定人机责任边界，并建立可以衡量的

收益指标。把使用率当成利润率，是企业 AI 预算中最常见的泡沫。这五类泡沫有一个共同病灶：收费权错配。

算力资源没有利用率，模型入口没有留存，应用没有进入流程，Agent 缺少可靠性，企业项目没有可衡量回报。反过来，拥有电力与供应链、模型分发入口、行业流程、专有数据、客户信任和责任承担的位置，即使经历估值回调，也更可能留下来。五类资产描述的是单点错配。但 2025 年以来，价格线上还出现了另一种形态：错配的单点正在被市值集中与产业链内部的相互交易连成整体。下一章单独处理这一结构。

五、铁索连环：当局部风险连成系统风险

（一）AI 已经不只是行业，而是指数本身

第三章的价格线主要观察一级市场。到了二级市场，风险变得更直观：如果 AI 定价发生错误，影响范围将从一个行业扩展到整个指数。集中度数据可作三层观察。第一层是指数结构：标普 500 前十大成分股权重已升至约 40%（2025 年一度录得 40.7% 的历史峰值）。2000 年互联网泡沫顶点时，这一比例约为 23% - 27%。当前集中度已超过 dot-com 时代。

第二层是 AI 权重。据市场统计，AI 相关股票合计约占标普 500 总市值的 45%，自 2022 年 11 月 ChatGPT 发布以来上升约 20 个百分点。摩根大通 2025 年 9 月的测算称，此后标普 500 约 75% 的涨幅与约 80% 的盈利增长来自 AI 相关公司。第三层是单一公司：英伟达 2025 年 10 月成为历史上第一家市值突破 5 万亿美元的公司，2026 年 5 月一度触及 5.5 万亿美元。其指数权重约 7%，超过标普 500 中能源、公用事业等整个行业板块，也超过美国与中国之外任何单一国家的 GDP。

宏观侧的对应数据，极端程度不相上下。据对美国经济分析局数据的测算，2026 年一季度美国 GDP 增长中约四分之三由 AI 相关资本开支贡献。Pantheon Macroeconomics 的分析师则指出，剔除 AI 支出后，美国企业资本开支为负。再看绝对规模：2026 年超大规模云厂商合计资本开支预计超过 6000 亿美元，同比增长约 36%。这些数字本身还不能直接判定泡沫，因为盈利增长确实存在。它们改变的是赌注的性质：当一个技术叙事同时占据指数近半市值，并成为宏观增长的重要引擎，价格线一旦出错，就很难再被限制在局部。

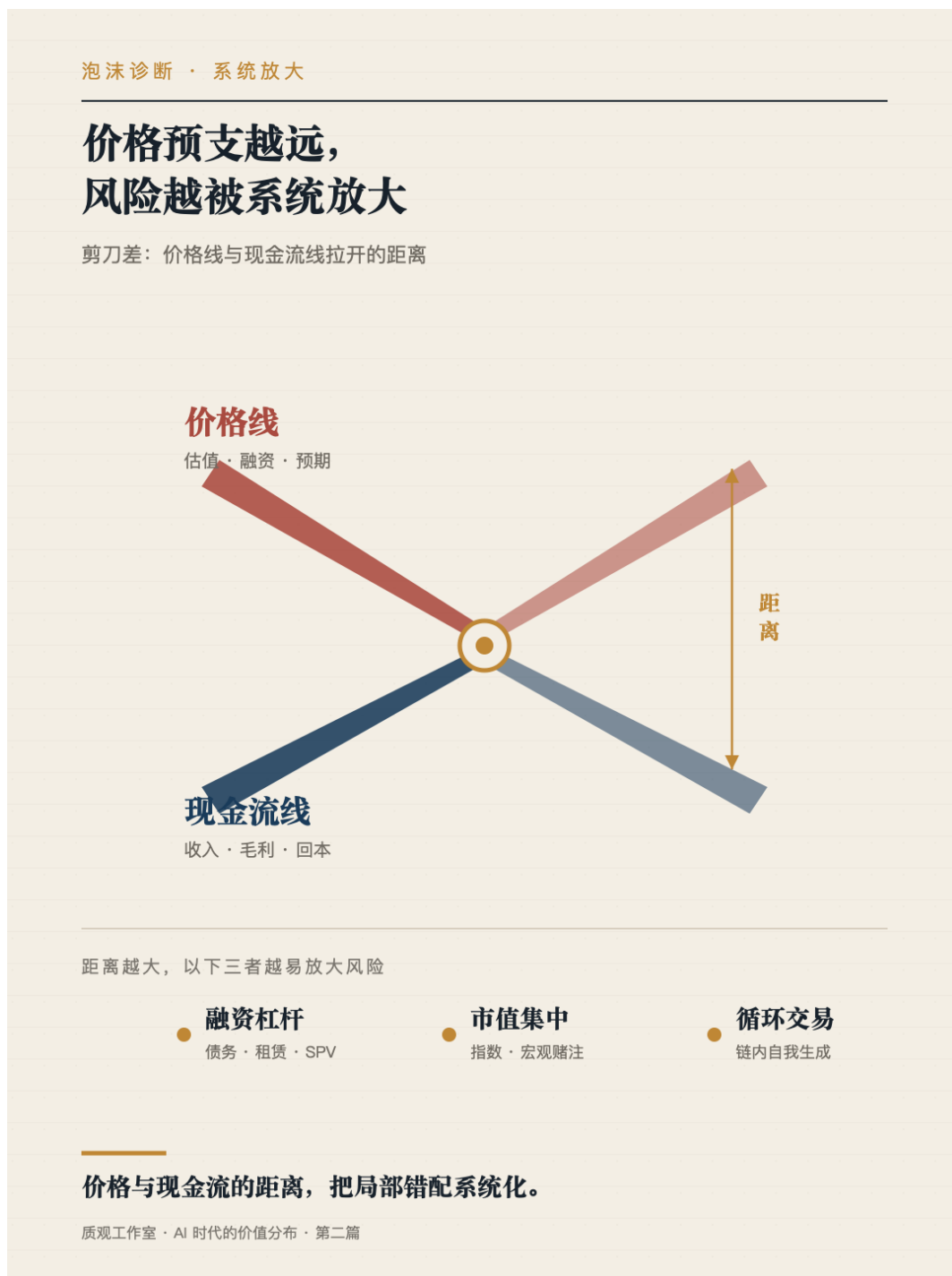


图 5 价格—现金流剪刀差如何放大风险

据此，四线框架还要补上一问：价格线不仅要和现金流线比较速度，还要判断这场赌注是否已经系统化。

（二）循环交易：同一美元被计算了几次？

价格线的另一重失真来自产业链内部。核心厂商之间的投资、采购和长期承诺，正在形成一条首尾相接的闭合链。2026 年以来，多家机构与媒体开始系统梳理这类安排，彭博将其概括为“微软、OpenAI

与英伟达如何互相付钱”。有分析将各类循环安排的累计规模估计在 8000 亿美元以上（口径不一，发表前建议回源）。

链条的基本走向可以简化为一圈：英伟达向 OpenAI 承诺最高 1000 亿美元的投资。OpenAI 将这笔资金连同其他融资，转化为对云厂商的算力采购承诺。据媒体统计，OpenAI 与七家供应商的基础设施承诺累计约 1.15 万亿美元（2025 – 2035，七家明细见章末数据注）。甲骨文等云厂商为履行合同，再向英伟达采购 GPU，货币回到起点。链条上还有股权纽带：AMD 向 OpenAI 授出最多约一成股份的认股权证，绑定其算力采购。英伟达持有 CoreWeave 约 7% 股份，并曾为其算力订单提供采购兜底。

这一结构的危险在于，同一笔资金会在不同环节呈现为不同的“好消息”：

- 在供应商处，它被计为投资；
- 在客户处，它变成资本开支；
- 在云厂商处，它又被确认为收入或未履约余额（backlog）。

backlog 支撑估值，估值支持再融资，再融资又进入下一轮采购。甲骨文因单一客户合同而急剧膨胀的剩余履约义务，是这一机制的公开样本。链条中的每一笔交易单独看都可能合规；合在一起观察，需求信号就可能在产业链内部自我生成。

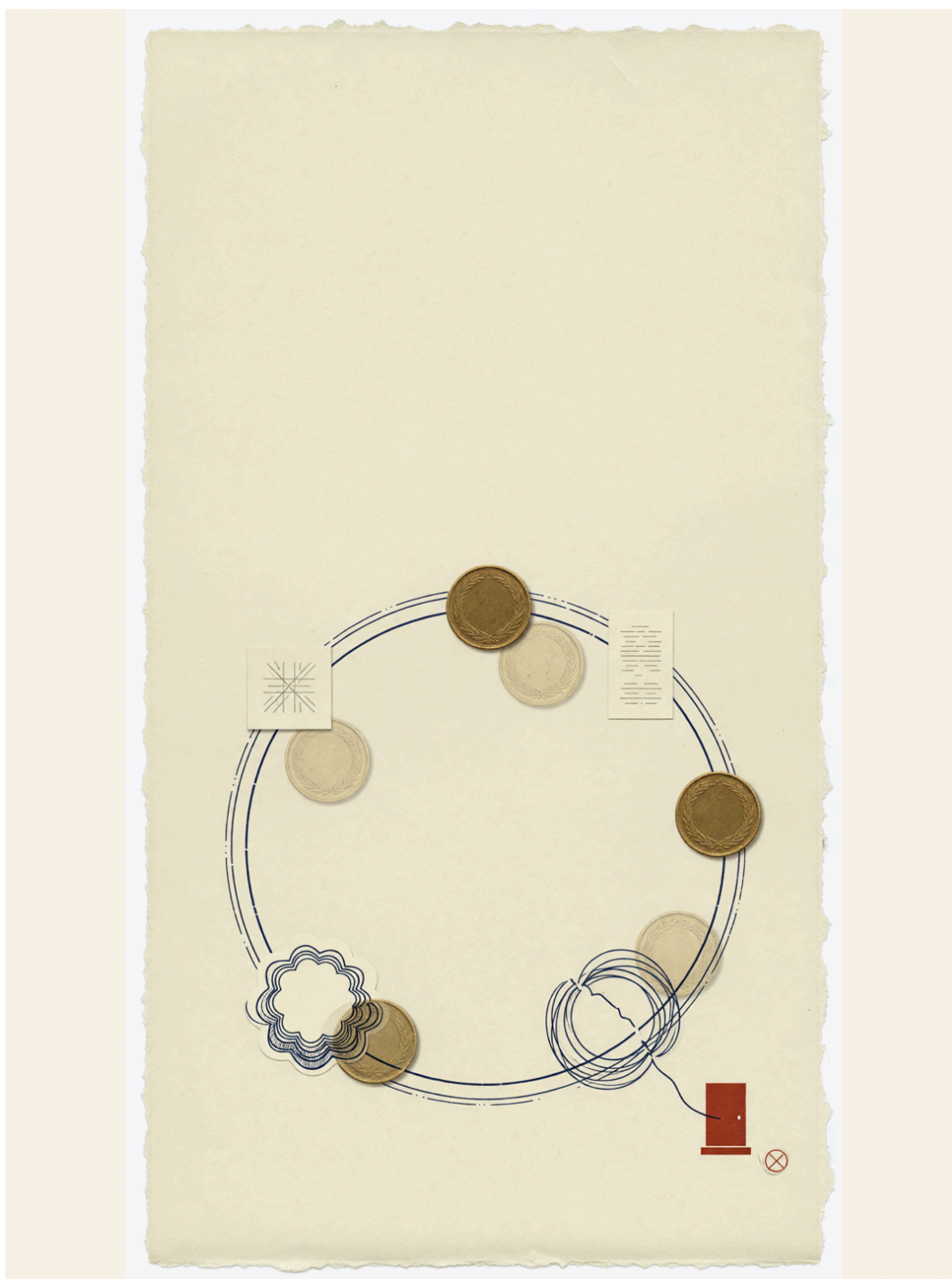


图 6 铁索连环：同一美元如何在链内循环

分析现金流线时，必须把收入拆成两部分：来自真实第三方客户的链外收入，以及产业链内部相互投资和采购形成的链内循环。

（三）未来是怎样被提前搬进今天的？

循环交易之所以构成“透支”，是因为三件事正在同时发生。其一，合同标的是远期需求。上述万亿美元级承诺多以 2030 年代的算力消耗为标的，而支付能力依赖尚未兑现的收入。据报道，OpenAI 2026 年亏损将扩大至百亿美元量级，其收入目标押在 2029 年前后（报道口径，需回源核验）。其二，融资

工具在加杠杆。数据中心建设大量使用债务、租赁与特殊目的载体，重资产平台的利息负担已在财报中显形（第三章 CoreWeave 单季 5.36 亿美元利息即为一例）。

其三，会计政策在拉长时间。多家算力厂商将 GPU 按六年折旧，而工程口径的有效寿命通常为三至四年，部分分析认为仅两至三年。有测算称，折旧年限的拉长将使行业到 2028 年少计费用逾千亿美元（第三方测算，需回源核验）。三者方向一致：未来需求和收入被提前计入今天的估值，未来成本则被尽量推迟确认。历史上可以找到一个近似对照。1996 至 2001 年，电信设备商为催谷需求向自己的客户放贷，朗讯、北电与思科都作出了供应商融资承诺（三家金额见章末数据注）。泡沫破裂后，这些贷款的回收率介于 20% 至 67% 之间，朗讯与北电此后再未回到巅峰。

不过，今天的差异必须说明：英伟达与超大规模云厂商自身拥有真实的高额现金流，资产负债表远厚于当年的电信商。但结构上的相似处同样清楚：供应商的收入与客户的资本开支互为镜像，链条整体的乐观被同时计入每一家的估值。“铁索连环”适合描述这种结构：局部波动可能被链条吸收，较大的风险也会沿链条更快传播。循环交易不必然是骗局，平台生态中的相互投资在产业史上并不罕见。它的真实代价，是让泡沫诊断的两条线同时失真：现金流线混入了链内自我循环的成分，价格线则把链条的系统性风险定价为零。对单个公司的检验方法，仍要回到收费权：链外收入（真实第三方付费）的占比与增速，是衡量每一环成色的第一指标。

数据注：OpenAI 与七家供应商的基础设施承诺明细（媒体汇总口径，非审计数据，2025 - 2035）——博通约 3500 亿美元、甲骨文约 3000 亿、微软约 2500 亿、英伟达约 1000 亿、AMD 约 900 亿、AWS 约 380 亿、CoreWeave 约 220 亿。1996 至 2001 年电信设备商的供应商融资承诺——朗讯约 81 亿美元、北电约 31 亿、思科约 24 亿。

六、为什么 AI 不是整体泡沫

拆清楚泡沫的位置和结构后，还必须看到另一面：AI 并不是一个整体泡沫。真实收入、真实生产率和真实的基础设施迁移，已经同时出现。

（一）基础设施已经产生真实现金流

芯片、网络、内存和电力已经对应现金流，并非只有叙事。NVIDIA 单季收入 816 亿美元，其中 Data Center 收入 752 亿美元；Broadcom AI 半导体收入 108 亿美元，自由现金流率达到 46%。这一部分类似互联网时代的光纤、服务器和路由器，也类似电气化时代的电网和变压器。个别公司可能被高估，个别产能可能过剩，但计算基础设施的迁移是真实的。

（二）云资本开支有风险，云需求却不是虚构

Amazon AWS 单季销售额 376 亿美元，同比增长 28%；Microsoft AI business 年化收入运行率超过 370 亿美元。Alphabet 披露，Google Cloud 收入超过 200 亿美元，同比增长 63%，并将增长归因于 AI 产

品和基础设施需求。但此处必须保留一个区分：需求真实只能证明产业方向成立，不能证明每一美元资本开支都合理。把二者混同，正是泡沫定价的组成部分。

（三）进入流程的 AI 已经开始收费

Palantir 的意义不止是个案。它展示了一种商业模式：收入高速增长、自由现金流率较高，NDR 达到 150%。它出售的不是聊天界面，而是一套把现实对象、数据、分析、决策和执行连接起来的系统。ServiceNow、Microsoft、Google Cloud 的企业 AI 进展亦表明这一点。AI 进入组织后，会被接入 IT、客服、运维、审批、工单和审计系统；一旦进入权限、数据、流程和责任分配，它就会从模型调用转为流程资产。这是流程收费权的雏形，系列第一篇已以专章展开。

（四）生产率提升已经可以被测量

Noy 与 Zhang 在 Science 发表的实验显示，ChatGPT 使中等专业写作任务的用时下降 40%、质量提升 18%。Brynjolfsson、Li、Raymond 对客服场景的研究则显示，生成式 AI 助手使每小时解决问题数平均提升 14%，新手和低技能员工的提升更明显。

GitHub Copilot 的控制实验中，开发者完成指定 JavaScript 任务速度提升 55.8%。HBS 与 BCG 的 "Jagged Frontier" 实验显示，在 AI 能力边界内，顾问完成任务数量提升 12.2%、速度提升 25.1%。但在边界外的任务上，错误上升。这些研究不能支持 "AI 会平均提高所有任务"，但足以反驳 "AI 完全没有生产率价值"。更准确的判断是：AI 先提高可文本化、可模板化、可快速校验的任务。越接近真实世界、组织责任和价值判断，越需要人类和制度兜底。

（五）科学应用已经产生长期价值

AlphaFold DB 已开放超过 2 亿个蛋白质结构预测。根据 EMBL-EBI 与 Google DeepMind 的公告，该数据库已被 190 多个国家的 300 多万人使用。2024 年诺贝尔化学奖将 AlphaFold2 列为关键突破之一，官方表述是它解决了蛋白质结构预测这一悬置 50 年的难题。若只观察聊天应用，容易把 AI 视为新一轮消费互联网热潮。但在蛋白结构、药物发现、材料设计、代码生成和工程仿真中，AI 更接近一种新的实验设备：它不替代科学本身，而是改变试错成本。目前，AI 中已经被真实证据支撑的部分包括：算力基础设施、云端智能供给、企业流程软件、可验证任务上的生产率提升，以及科学工具化应用。

风险更大的，是一连串概念偷换：

- 因为 AI 会改变世界，所以任何 AI 公司都值钱；
- 因为算力需求真实，所以任何数据中心都能回本；
- 因为模型有能力，所以任何封装应用都有护城河。

技术革命是真实的，资产定价未必是真实的。本篇论证围绕这两个命题展开。

七、泡沫退去后，什么会留下

泡沫破裂清算的是错误定价，不是已经形成的生产能力。AI 泡沫如果出清，被淘汰的将是高估值、低利用率、没有收费权的资产；留下来的，大致有四类。

（一）更便宜的智能

模型公司可能被重新估值，模型 API 的利润率可能继续下降，前沿模型之间的竞争也会更加激烈。但智能成本持续下降的方向，不会轻易逆转。当模型能力接近、开源权重可用、多模型路由成熟、云厂补贴加深时，智能会像带宽、云服务器、开源软件一样，成为多数企业和个人可负担的生产资料。这会压低单纯“生产答案”的价值，同时扩大“基于答案改造流程”的机会；系列第一篇的价值分布分析，正建立在这一不对称之上。

（二）被重新定价的基础设施

算力、电力和数据中心会出现过剩，也会出现烂尾。但已经建成的电力接入、光纤、液冷、园区和集群，只要能够被重新定价，就可能成为下一轮应用的低成本底座。互联网泡沫后的光纤过剩最终降低了宽带成本。AI 泡沫后的算力过剩，也可能降低推理和行业部署的门槛。但这一类比须加以限定：AI 数据中心受电力与本地容量约束，不能简单照搬互联网光纤的结局。

（三）真正进入组织流程的 AI

企业不会因为一些失败的 PoC 就停止使用 AI，而会变得更加挑剔：少买漂亮演示，多买能够进入工单、审批、审计、生产、销售和售后流程的系统。AI 从工具变成流程之时，才开始产生长期收费权。这一变化将价值从“谁能生成答案”迁移至“谁能让答案进入现实行动”。

（四）更清晰的价值分配规则

泡沫期，市场往往奖励“看起来像 AI 公司”的公司。泡沫退去后，市场会更准确地识别拥有收费权的位置：

- 资源收费权来自电力、芯片、网络、云和供应链；
- 入口收费权来自用户、办公、开发和行业入口；
- 流程收费权来自数据、权限、组织行动和责任承担。

仅能调用模型的公司会被压价，能把模型转化为组织行动的公司更可能留存。三类收费权的完整界定、代表企业与各自的风险敞口，见系列第一篇，此处不再重复。



图 7 泡沫清算价格，不清算智能

八、附论：没有现金流的资产，究竟靠什么定价

财经写作经常把郁金香与比特币并置，并用前者证明后者必然归零。但这种简单类比，已被比特币此后的市场历史多次否定。这一节不重复“比特币必然归零”的断言，而是借这组对照回答一个更窄的问题：没有现金流的资产靠什么定价？这个问题也能帮助我们划清 AI 资产的估值边界。相似之处确实成立。两者都不产生现金流：球茎不分红，比特币不付息，持有者的全部回报预期都寄托于下一位买家。

两者的稀缺性都由某种外部机制维护：郁金香靠病毒制造的条纹（且该病毒同时在削弱球茎），比特币靠协议写定的 2100 万枚上限。

两者都伴随交易基础设施的狂热：一边是酒馆学院、酒钱与借据链，一边是交易所、杠杆与衍生品。两者都吸引了大量此前从未参与投机的新钱，也都招致了同时代道德家的批判。但两处结构性差异不容忽略。第一是存续记录。郁金香只经历了一轮生死：1637 年 2 月之后，作为投机资产的郁金香再未回来。

比特币自 2009 年运行以来，至少经历四轮 75% 以上的回撤。2011 年约下跌 93%（约 32 美元至 2 美元），2013 至 2015 年约下跌 85%。2017 至 2018 年约下跌 83%（约 19,700 美元至约 3,200 美元）。2021 至 2022 年约下跌 77%（约 69,000 美元至约 15,500 美元）。每一轮之后，价格均回到并超越前高。2024 年 1 月美国批准现货比特币 ETF，同年 12 月价格首次突破 10 万美元。（以上为公开市场历史数据，精确数字发表前建议回源；本篇不引用 2025 年之后的行情。）

每轮熊市之后，其基础设施（交易所、托管、合规框架、ETF）都在增厚。而郁金香的酒馆借据链，在信任断裂的当周即告消失。第二是供给与清算结构。球茎会繁殖，稀缺性靠病毒勉强维持，供给最终压垮了叙事。比特币的供给上限写入协议，清算依托的是协议本身而非任何单一对手方的私人信用。

泡沫诊断 · 附论对照

没有现金流线的资产， 如何定价——以及边界

郁金香与比特币：结构对照，不是类比判决



图 8 郁金香与比特币：无现金流资产的结构差异

对照可以得到两个核心结论。第一，二者都没有现金流线。比特币的价格无法像股票那样通过未来现金流折现来收敛；它主要由共识定价，在资产分类上更接近黄金。对它的核心追问也随之改变：问题不再是“现金流能否兑现”，而是“共识能够维持多久”。这一定义既不是对价格的辩护，也不是对价格的否定。第二，两个样本回答的是不同问题。郁金香检验的是“没有现金流、没有持久基础设施、没有可靠清算制度”的资产会怎样；比特币检验的则是“没有现金流，但拥有基础设施和清算制度”的资产能够走多远。

这组对照还划出了一条边界：AI 资产不能用比特币式的“共识定价”为自己辩护。芯片、云、模型和应用向市场承诺的，都是未来现金流。承诺现金流的资产，最终要按现金流对账。一项 AI 资产若开始以稀缺叙事或生态共识替代现金流路径来支撑估值，它便正在从“被错误定价的技术”滑向“郁金香”。四线框架可以据此使用：资产缺少哪条线，就以哪条线的标准审视它。比特币坦承自己没有现金流线，多数 AI 资产则声称有。声称有而长期拿不出，才是本篇定义下的泡沫。

结论：泡沫会清算价格，但不会清算智能

AI 泡沫的危险，在于市场把“AI 有用”直接等同于“这项资产一定能赚钱”。能力进步、企业采用和部分现金流，已经足以证明 AI 的技术方向成立。估值、资本开支和叙事的速度，也足以证明资产泡沫确实存在。还需要警惕的是，市值集中与循环交易正在把这场泡沫的赌注，从单个公司推向指数和宏观经济层面。因此，结论既不是“AI 是泡沫”，也不是“AI 没有泡沫”，而是：

AI 是一场真实技术革命中的资产泡沫。

泡沫如果破裂，被清算的将是低利用率项目、伪需求和收费权薄弱的应用；留下来的，是算力基础设施、工具链、数据结构、行业流程，以及能够进入现实世界的应用能力。石油、铁路、电力和互联网的历史，都以“清算错误定价、保留生产能力”收场。郁金香则展示了另一种结局：没有现金流、没有持久基础设施、没有可靠清算制度。AI 整体并不符合这三个条件，但 AI 内部确有一部分资产正在向它们靠近。识别这些资产，是当前投资判断的重点。至此，三部曲的前两个问题已经问完：第一篇问“钱停在哪里”，这一篇问“价格是不是真的”。还剩最后一个问题：在价值分布与资产定价之间，人还剩下什么位置？这是第三篇《人的位置》要回答的问题。三个答案最终会指向同一个地方：当智能越来越便宜，能够留住价值的，是难以复制的位置，以及无法外包的责任。

附录 泡沫史闻九则

以下轶事按可信度分为三类：档案可证，指有原始文献或档案支持；时人轶闻，指出自同时代记载、但细节难以完全核实；后世传说，指主要经麦凯 1841 年《大癫狂》等后世著作流传，现代研究尚未找到档案支持。

标注“后世传说”的故事不宜作为史实引用。它们被收录，是因为故事的流传本身也是泡沫心理的史料；每个时代，都需要一些关于上一场疯狂的笑话。

一、被当成洋葱吃掉的球茎。

1630 年代的荷兰正值黄金时代。海上贸易带来的财富，催生了历史上第一场有完整记录的投机狂热。1636 年冬至 1637 年初，郁金香球茎论颗计价，最珍稀品种的报价足以买下阿姆斯特丹运河边的一栋宅邸。交易则全部在酒馆中以私人借据完成。流传最广的故事也发生在这个冬天：一名水手误把商人柜台上一颗价值连城的 Viceroy 球茎当成洋葱，就着腌鲱鱼吃了下去。按照当时报价，这顿早餐足以供养一整条船的船员一年；故事的完整版还说，水手因此入狱数月。但现代研究者从未在档案中找到这起案件，而且郁金香球茎带有轻微毒性，味道也与洋葱相去甚远。（出处：Mackay《大癫狂》；可信度：后世传说。）

二、一颗球茎的购物清单。

1637 年 2 月泡沫破裂后，荷兰印刷市场立即涌出大批道德劝诫小册子，嘲讽刚刚过去的疯狂。泡沫清算，往往伴随着出版业的繁荣。其中一份小册子列出了一颗 Viceroy 球茎的“等价物”：2 拉斯特小麦、4 拉斯特黑麦、4 头肥牛、8 头猪、12 只羊；另加 2 大桶葡萄酒、4 桶啤酒、2 吨黄油、1000 磅奶酪，以及一张床、一套衣服和一只银杯，合计约 2500 荷兰盾。当时熟练工匠的年收入约为 300 盾。这份清单宣传色彩明显，数字很可能被夸大，但它真实反映了时人面对价格失衡时的震惊。（出处：时代小册子，经 Mackay 转述；可信度：时人轶闻，数字宜作修辞看待。）

三、画册即招股书。

郁金香交易最狂热的月份在冬季，但球茎埋在土中，要到次年春天才能起挖。买家从头到尾看不到商品，这也是这场投机被时人称为“风的交易”的原因。交易只能依据画师绘制的郁金香图册（tulip books）进行：一页一个品种，标注名称与重量，功能很像今天的招股说明书。画师也因此成为泡沫产业链中最稳定的受益者；无论球茎涨跌，画册总要有画。如今，存世的郁金香图册已经成为博物馆藏品和拍卖会上的高价艺术品。泡沫中最耐久的资产，反而是那本为泡沫做广告的册子。（出处：存世图册与荷兰档案；可信度：档案可证。）

四、“经营何业，无可奉告”。

1720 年的伦敦上演了英国历史上最大的股票泡沫。南海公司以承接英国国债为条件，换得南美贸易垄断权；在一套精心设计的换股方案推动下，股价半年内上涨约八倍，从贵族到女仆都涌入市场。当时证券交易还没有固定场所，交易巷（Exchange Alley）的几家咖啡馆就是交易所。热潮中，上百家蹭行情的“泡沫公司”涌现，招股名目包括从海水中提取白银、制造永动机等。流传最广的一家公司在招股书中写道：“经营一项具有巨大优势的事业，但无人知晓其为何业。”据说发起人上午募得 2000 英镑，当晚便离开英国。

这家公司很可能只是时人讽刺作品的虚构，但有据可查的真实招股项目，荒诞程度并不逊色。另一层讽刺则有档案支持：1720 年《泡沫法案》的通过得到南海公司推动，目的之一是取缔与它争夺资金的

其他泡沫公司；由泡沫推动立法，禁止其他泡沫。（出处：Mackay 及同时代出版物；可信度：该公司故事介于时人轶闻与后世传说之间，《泡沫法案》背景为档案可证。）

五、牛顿的两次进出。

同一场南海泡沫，也套住了当时公认最聪明的人。艾萨克·牛顿时任皇家造币厂总监，他在南海股票上的操作有账目可查：早期入场，获利了结；随后眼看股价继续翻倍、身边的人不断暴富，又在高位重新重仓买入。泡沫破裂后，牛顿亏损约 2 万英镑，相当于他多年的薪俸。至于那句“我能计算天体的运行，却无法计算人类的疯狂”，出自后人的轶闻集，并未见于牛顿本人文字。（出处：牛顿账目研究；可信度：亏损档案可证，名言为后世传说。）

六、出租后背的驼背人。

与南海泡沫几乎同时，巴黎正在上演规模更大的密西西比泡沫。苏格兰人约翰·劳主持法国财政，一边发行纸币，一边由密西西比公司垄断法属北美贸易。公司股价一年内上涨约二十倍。全城资金涌向坎康普瓦街（rue Quincampoix），这条窄街成了露天交易所，沿街房租涨到平日的十余倍。据说，一名驼背者甚至出租自己的后背，供交易者垫着签约收费。这轮泡沫还为法语留下了一个新词：millionnaire（百万富翁）。在约翰·劳的纸币体系制造出第一批百万富翁之前，这个词甚至没有存在的必要。（出处：同时代巴黎记载；可信度：街景与词源档案可证，驼背人为时人轶闻。）

七、勃朗特小姐的铁路股。

1840 年代中期的英国铁路狂热，是中产阶级第一次大规模参与股票投机。1845 年前后，有 562 个新铁路项目提交议会；报纸整版刊登招股广告，教师、牧师和寡妇都成了股东。股东名册中还有夏洛蒂·勃朗特。她和两个妹妹持有约克—北米德兰铁路股票。泡沫破裂后，她在书信中写道，损失已成定局，所幸投入的是“可以承受损失的钱”，并拒绝责怪任何人。同一场泡沫的另一端，是“铁路大王”乔治·哈德森。他一度控制全国近三分之一的铁路里程，并用新募集的资本冒充股息派发，以维持股价和声望，最终在崩盘后身败名裂。一场泡沫，两种股东。（出处：勃朗特书信、议会对哈德森的调查；可信度：档案可证。）

八、灰泥与飓风。

1920 年代的美国，汽车开始普及，佛罗里达被宣传为人人都能抵达的热带天堂。迈阿密地价在 1925 年前后一年翻了数倍，买地的定金收据（binder）本身也变成投机标的。土地认购权收据一天之内可以在人行道上转手数次，倒卖收据的“宾德小子”（binder boys）甚至成为一种职业。出狱后的查尔斯·庞兹（“庞氏骗局”一词的本尊）也来到佛州卖地。他在广告中声称地块“临近杰克逊维尔”，实际位置却在 65 英里外的沼泽中。1926 年 9 月，迈阿密大飓风为这场行情收尾。马克斯兄弟 1929 年的电影《椰子》以此为题材。格劳乔的台词流传至今：“想要什么样的房子都行，甚至可以要灰泥的（stucco）——

哦，你会被粘住的（stuck-oh）。”（出处：佛州地产史、庞兹案法庭记录、影片对白；可信度：档案可证，庞兹广告措辞为时人轶闻。）

九、袜子木偶与改名的冰茶。

1999 年至 2000 年初，是 dot-com 泡沫的顶点。公司名称只要沾上“.com”，似乎就能上市融资，盈利与否反而不重要。宠物用品电商 Pets.com 的袜子木偶，几乎成了这场泡沫的官方吉祥物：2000 年 1 月登上超级碗广告，2 月 IPO 募资 8250 万美元，11 月便宣布清算。从全国明星到关门，只用了九个月。

学术界为这种现象留下了严谨注脚。Cooper、Dimitrov 与 Rau（2001）发现，1998 至 1999 年间，公司名称加上“.com”，就能在公告期获得显著超额收益；主营业务是否与互联网相关，并不影响这一效应。论文口径的超额收益约为七成，精确数字仍需回源核验。2017 年，这一现象再次出现。比特币牛市正盛，“区块链”成为新的魔法词。长岛冰茶公司（Long Island Iced Tea Corp.）更名为“长区块链公司”（Long Blockchain Corp.），股价盘中一度上涨近三倍。公司后来被交易所摘牌，美国证监会也在 2021 年就相关内幕交易提起指控。泡沫的标的会更新，命名对价格的刺激却反复出现。（出处：当年报道、Journal of Finance 论文、SEC 执法公告；可信度：档案可证。）

参考文献

泡沫的界定与历史参照

- Federal Reserve Bank of San Francisco, Asset Price Bubbles: <https://www.frbsf.org/research-and-insights/publications/economic-letter/2007/10/asset-price-bubbles/>
- Charles Mackay, 《Extraordinary Popular Delusions and the Madness of Crowds》（1841），Project Gutenberg: <https://www.gutenberg.org/ebooks/24518>
- Anne Goldgar, 《Tulipmania: Money, Honor, and Knowledge in the Dutch Golden Age》，University of Chicago Press, 2007: <https://press.uchicago.edu/ucp/books/book/chicago/T/bo5414939.html>
- Peter M. Garber, 《Famous First Bubbles: The Fundamentals of Early Manias》，MIT Press, 2000: <https://mitpress.mit.edu/9780262571531/famous-first-bubbles/>
- AOGHS, Pithole Creek 与早期石油管道: <https://aoghs.org/petroleum-pioneers/pithole-creek/>
- Drake Well Museum, Pithole 历史: <https://www.drakewell.org/plan-your-visit/what-to-see-and-do/pithole>
- New York Fed, Railway Mania 回顾: <https://libertystreeteconomics.newyorkfed.org/2015/06/crisis-chronicles-railway-mania-the-hungry-forties-and-the-commercial-crisis-of-1847/>
- Cambridge University Press, Railway Mania 研究: <https://www.cambridge.org/core/books/abs/boom-and-bust/democratising-speculation-the-great-railway-mania/F5A16DC376684C14020C78D86B4A4B48>
- Yale Energy History, 1920s – 1930s 美国公用事业控股公司: <https://energyhistory.yale.edu/electricity-and-the-public-good-private-public-power-debates-in-the-1920s-30s/>
- SEC, Public Utility Holding Company Act of 1935: <https://www.sec.gov/about/laws/puhca35.pdf>
- EIA, 美国农村电气化历史: <https://www.eia.gov/todayinenergy/detail.php?id=67824>
- Goldman Sachs, dot-com 泡沫回顾: <https://www.goldmansachs.com/our-firm/history/moments/2000-dot-com-bubble>
- Our World in Data, 全球互联网使用率: <https://ourworldindata.org/data-insights/internet-use-became-the-norm-for-humanity-only-very-recently>

- FRED / U.S. Census, 电商占美国零售总额比例: <https://fred.stlouisfed.org/series/ECOMPCTSA>
- Carlota Perez, 技术革命与金融泡沫: <https://carlotaperez.org/wp-content/downloads/media/articles-and-blogs/PEREZTechnologyandbubblesforEngelsbergseminar.pdf>

四线诊断

- OpenAI API Pricing (2026 年 7 月 8 日可见口径, 模型价格与命名可能变化, 正式发表前建议回看原页面): <https://developers.openai.com/api/docs/pricing>
- Epoch AI, LLM 推理价格下降趋势: <https://epoch.ai/data-insights/llm-inference-price-trends>
- Stanford HAI, 2026 AI Index Report: <https://hai.stanford.edu/ai-index/2026-ai-index-report>
- McKinsey, The State of AI: Global Survey 2025: <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>
- Microsoft FY26 Q3 Earnings: <https://www.microsoft.com/en-us/investor/events/fy-2026/earnings-fy-2026-q3>
- Palantir Q1 2026 Business Update: <https://investors.palantir.com/files/Palantir%20-%20Q1%202026%20Business%20Update.pdf>
- Gartner, 30% 生成式 AI 项目或止步 PoC: <https://www.gartner.com/en/newsroom/press-releases/2024-07-29-gartner-predicts-30-percent-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025>
- Gartner, 超过 40% Agentic AI 项目或于 2027 年底前取消: <https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027>
- NVIDIA Q1 FY2027 财报: <https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-first-quarter-fiscal-2027>
- Broadcom Q2 FY2026 财报: <https://investors.broadcom.com/news-releases/news-release-details/broadcom-inc-announces-second-quarter-fiscal-year-2026-financial>
- Amazon Q1 2026 财报: <https://ir.aboutamazon.com/news-release/news-release-details/2026/Amazon-com-Announces-First-Quarter-Results/default.aspx>
- Meta Q1 2026 财报: <https://investor.atmeta.com/investor-news/press-release-details/2026/Meta-Reports-First-Quarter-2026-Results/default.aspx>
- CoreWeave Q1 2026 SEC 披露: <https://www.sec.gov/Archives/edgar/data/1769628/000176962826000220/coreweave1q26earningspress.htm>
- OECD, AI 风投融资至 2025 年 (Preqin 数据库口径, 按已披露交易统计, 可能低估未披露交易): https://www.oecd.org/en/publications/venture-capital-investments-in-artificial-intelligence-through-2025_a13752f5-en/full-report.html
- IEA, Energy demand from AI: <https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai>
- EPRI, Powering Intelligence 2026: <https://powering-intelligence.epri.com/load-growth.html>

生产率与科学证据

- Science, Noy & Zhang, 生成式 AI 对写作任务生产率的影响: <https://www.science.org/doi/10.1126/science.adh2586>
- NBER, Generative AI at Work: <https://www.nber.org/papers/w31161>
- Microsoft Research, GitHub Copilot 生产率实验: <https://www.microsoft.com/en-us/research/publication/the-impact-of-ai-on-developer-productivity-evidence-from-github-copilot/>
- Harvard Business School, BCG / Jagged Frontier 实验: <https://www.hbs.edu/faculty/Pages/item.aspx?num=64700>
- AlphaFold Protein Structure Database: <https://alphafold.ebi.ac.uk/>
- EMBL-EBI, AlphaFold Database 使用与续约公告: <https://www.embl.org/news/science-technology/google-deepmind-partnership-renewal/>
- Nobel Prize, 2024 年化学奖新闻稿: <https://www.nobelprize.org/prizes/chemistry/2024/press-release/>

市值集中与循环交易

- J.P. Morgan Asset Management, How extreme is market concentration?: <https://am.jpmorgan.com/us/en/asset-management/liq/insights/market-insights/market-updates/on-the-minds-of-investors/how-extreme-is-market-concentration/>
- Pensions & Investments, Top 10 stocks account for nearly 40% of the S&P 500: <https://www.pionline.com/data-rankings/chart-of-the-day/pi-sp500-index-concentration/>
- Yahoo Finance, AI 相关股票占标普 500 市值约 45%: <https://finance.yahoo.com/markets/stocks/articles/ai-swallows-wall-street-stocks-094052523.html>
- Econofact / Gigafact (引摩根大通 2025 年 9 月测算: ChatGPT 发布以来 AI 相关公司贡献约 75% 涨幅与 80% 盈利增长): <https://econofact.org/factbrief/fact-check-have-ai-related-companies-driven-stock-gains-since-the-launch-of-chatgpt>

- CNBC, 英伟达市值收盘突破 5 万亿美元: <https://www.cnbc.com/2026/04/24/nvidia-stock-closes-at-record-pushing-market-cap-past-5-trillion.html>
- Forbes, 英伟达市值触及 5.5 万亿美元: <https://www.forbes.com/sites/antoniopequenoiv/2026/05/13/nvidia-hits-record-55-trillion-value-first-company-to-ever-reach-mark/>
- Bloomberg, AI Circular Deals: How Microsoft, OpenAI and Nvidia Keep Paying Each Other: <https://www.bloomberg.com/graphics/2026-ai-circular-deals/>
- Advisor Perspectives, Nvidia Deals: Vendor Financing or Round-Tripping?: <https://www.advisorperspectives.com/articles/2025/11/10/nvidia-deals-vendor-financing-round-tripping>
- Tomasz Tunguz, Circular Financing: 英伟达与北电供应商融资对照: https://tomtunguz.com/nvidia_nortel_vendor_financing_comparison/
- Fortune, Pantheon: 剔除 AI 后美国企业资本开支为负: <https://fortune.com/2026/02/23/ai-capex-us-gdp-negative-pantheon/>
- CNN Business, OpenAI 关联交易与 dot-com 时代担忧: <https://www.cnn.com/2025/10/07/business/openai-nvidia-bubble-nightcap>
- Quinn Emanuel, AI 数据中心融资与诉讼风险 (含 GPU 折旧年限讨论): <https://www.quinnemanuel.com/the-firm/publications/client-alert-emerging-litigation-risks-in-financing-ai-data-centers-boom/>

附论与附录

- Cooper, Dimitrov & Rau, 《A Rose.com by Any Other Name》, Journal of Finance, 2001: <https://onlinelibrary.wiley.com/doi/10.1111/0022-1082.00408>
- SEC, 现货比特币 ETP 批准相关声明 (2024 年 1 月, sec.gov, 正式引用前请检索原文)
- SEC, Long Blockchain 相关执法公告 (2021, sec.gov, 正式引用前请检索原文)
- 夏洛蒂·勃朗特书信 (关于铁路股票投资, 收录于其书信集, 正式引用前请核对版本)

数据口径与使用说明：本篇为公开信息的整理与分析，不构成投资建议。历史轶闻部分的可信度分级见附录题注，标注“后世传说”者请勿作为史实引用。比特币价格为公开市场历史数据，本篇不引用 2025 年之后的行情，正式发表前请回源补核精确数字；Cooper 等 (2001) 的超额收益口径、SEC 两项公告的准确链接亦请发表前回源。第五章的市值集中度、AI 市值占比、循环交易总规模、OpenAI 供应商承诺明细、GPU 折旧测算与 OpenAI 亏损预测，均为 2025 - 2026 年媒体报道与第三方分析口径，变动快、非审计数据，正式发表前必须逐项回源核验。其余数据口径说明同系列总稿：各公司财务数据以其财报与 SEC 披露为准，估值与份额多为媒体报道或第三方估算，正式引用请回溯原始来源。

03

人的位置

AI、劳动与不可外包的责任

当生成趋近免费，人的位置转向判断、选择与责任

质观工作室 · “AI 时代的价值分布”三部曲之三（2026）

摘要：本文从作品、劳动和任务三个层面讨论 AI 时代人的位置。

上部讨论作者性。工具史表明，工具越复杂，作者并不会消失，作者性只是从直接执行转移到目的设定、材料选择、过程控制和结果责任。各国关于 AI 作品的立法与判例也在沿着同一条边界推进：美国版权局认为，仅凭提示词通常不足以构成人类作者身份；北京互联网法院则承认，持续迭代提示、调整参数和筛选结果可能体现独创性投入。各法域争论的焦点，是“人的投入达到什么程度才算创作”；至于“AI 能否成为作者”，目前没有任何法域承认 AI 本身具有作者身份。

中部讨论劳动。本文以马克思的对象化与异化理论为起点，并借助阿伦特、黑格尔、布雷弗曼、斯蒂格勒以及《庄子》中的庖丁与轮扁，回答两个问题：劳动在什么条件下仍然属于人，劳动又在什么条件下沦为外部系统的附属物。《长安的荔枝》中的李善德提供了一个具体样本：异化不必等到 AI 出现；人的主体性，至少取决于四件事：目的是否由人确定，过程是否由人理解，结果是否由人选择，责任是否由人承担。

下部讨论任务和未来劳动形态。现有研究表明，AI 首先替代的是任务，而不是整份职业。最先被压价的，是可文本化、可模板化、容易验证、责任成本较低的工作；冲击尤其集中于年轻人的入门任务。据此提出，人的位置不在“比机器更会生成”，而在成为责任性的生成者：定义问题、整合流程、校验结果，并对现实后果负责。

最后，本文以 2026 年兴起的 harness 工程为样本，将其译为“驾驭工程”。token 是按量计价、无差别供应的空白智能；把它转化为具体劳动的，是使用者长期积累的习惯、知识、价值方向和审美标准。驾驭工程因此既是一种工作流，也是一种外在于身体、却属于个人的人力资本。未来劳动很可能由此上移：直接执行交给算力，人负责建造、维护和拥有自己的驾驭工程。三部曲至此收官：第一篇讨论钱停在哪里，第二篇讨论价格是否真实，第三篇回到分布与价格背后的人。

关键词：人工智能；劳动；任务替代；异化；对象化；作者性；工具史；默会知识；版权；驾驭工程（harness）；责任

本篇术语说明

- **生成式 AI / 大语言模型（LLM）：**以海量数据训练、能够生成文本、图像、代码等内容的 AI 系统；大语言模型是其中处理语言的一类。
- **任务暴露度：**某项职业的工作任务中，可能受到 AI 显著影响的比例。暴露不等于被替代，更不等于失业。
- **自动化与增强：**技术影响任务的两种方向。自动化是由系统直接代替人完成；增强是帮助人更快、更好地完成。
- **对象化：**劳动哲学概念，指人通过劳动，把自己的能力、意图和判断凝结到外部世界的产品之中。
- **异化：**在特定生产关系下，劳动产品、劳动过程和社会关系反过来成为支配劳动者的异己力量。

- 作者性：一件作品在何种意义上属于某个人。它不取决于体力投入多少，而取决于目的、选择、控制与责任。
- 扩展心智：心灵哲学观点，认为身体之外的工具和外部结构，也可以成为认知过程的一部分。
- 技术具身：技术哲学观点，指熟练掌握的工具会退入背景，成为人感知和行动世界的方式，例如眼镜之于视力。
- 默会知识：波兰尼提出的概念，指“能做却难以完整说清”的知识，如骑车、辨味和手艺。这类知识难以彻底文本化，因而也更难被模型直接复制。
- 去技能化：布雷弗曼对泰勒制的诊断。构想与执行分离后，劳动者的知识与技能被系统性剥离。
- harness（驾驭工程）：包裹在 AI 模型之外，负责任务拆解、工具调用、结果验证、权限控制与过程留痕的工程层。行业常用公式是 $\text{Agent} = \text{Model} + \text{Harness}$ 。harness 字面义为马具或缰轡，本文译作“驾驭工程”，强调它将使用者的习惯、知识、价值和审美转化为可运行的工作系统。

上部 作品与作者性

一、九千年前的骨笛

河南舞阳贾湖遗址出土过三十余支骨笛。它们以丹顶鹤尺骨制成，多为七孔，距今约七千八百年至九千年。九千年前，有人拿起一根鹤骨，打磨、钻孔、试音，再反复调整，直到它发出自己想要的声音。1999 年，《Nature》发表了关于这批骨笛的研究：保存最完整的一支至今仍可吹奏，并能发出接近七声音阶的音列。它是目前已知最早的可演奏乐器之一。九千年后，一个人在软件里输入一句话，几秒钟便得到一段旋律。于是，一个古老的指控换上了新外衣：这不算音乐。这个人也不算音乐人，因为实际干活的是机器。

这种指控并不新鲜。工具史上，每当一种新工具进入创作，它都会重新出现一次。但如果把它的逻辑推到底，结论会变得荒谬：工具越先进，人的创作越不算数。照此推论，骨笛吹奏者似乎最“纯粹”，钢琴家要打折，使用合成器的乐手更可疑，在录音棚里工作的音乐人几乎不配署名。文明史的经验恰好相反。人的能力，本来就是一部工具史。没有弓、犁、船、印刷机、显微镜、照相机、录音棚、合成器、数控机床、CAD、编译器和搜索引擎，所谓“人的能力”会小得多。

乐器的历史把这一点说明得很清楚。乐器越复杂，音乐人的控制方式也越复杂。钢琴扩展了手指，管风琴扩展了气息和空间，录音棚扩展了声音的可编辑性。MIDI 让不同品牌的电子乐器、合成器、鼓机、音序器与电脑互相通信；DAW 则把剪辑、编曲、混音、采样和自动化纳入同一个创作界面。Ableton 对“录音棚作为乐器”的历史梳理说明，录音棚早已不只是记录声音的场所，它本身就是作曲工具。音乐也不再只是现场演奏的瞬间，而包括对声音材料的切割、重组和重新赋义。工具没有取消作者，只是改变了作者出现的位置。

摄影的作者性，不在于人是否亲手逐点画出图像，而在取景、等待、曝光、选择、冲洗、裁剪、叙事和发表。电子音乐的作者性，也不取决于每个声音是否来自人声或木制乐器，而取决于声音设计、节奏组织、采样选择、结构控制和审美判断。程序员是否亲手敲出每一行字符，同样不是要害。要看的是需求理解、架构设计、边界处理、调试、取舍和维护责任。木匠用电锯做桌子，作品仍然属于木匠。判断的焦点不在机器是否参与，而在人是否仍在做五件事：设定目的，选择材料，控制过程，修正结果，并承担作品进入世界之后的后果。AI 应当被放回这条工具史中理解。它不是第一种介入创作和劳动的机器，却第一次把语言、图像、代码、分析和决策建议，也纳入了可自动生成的范围。

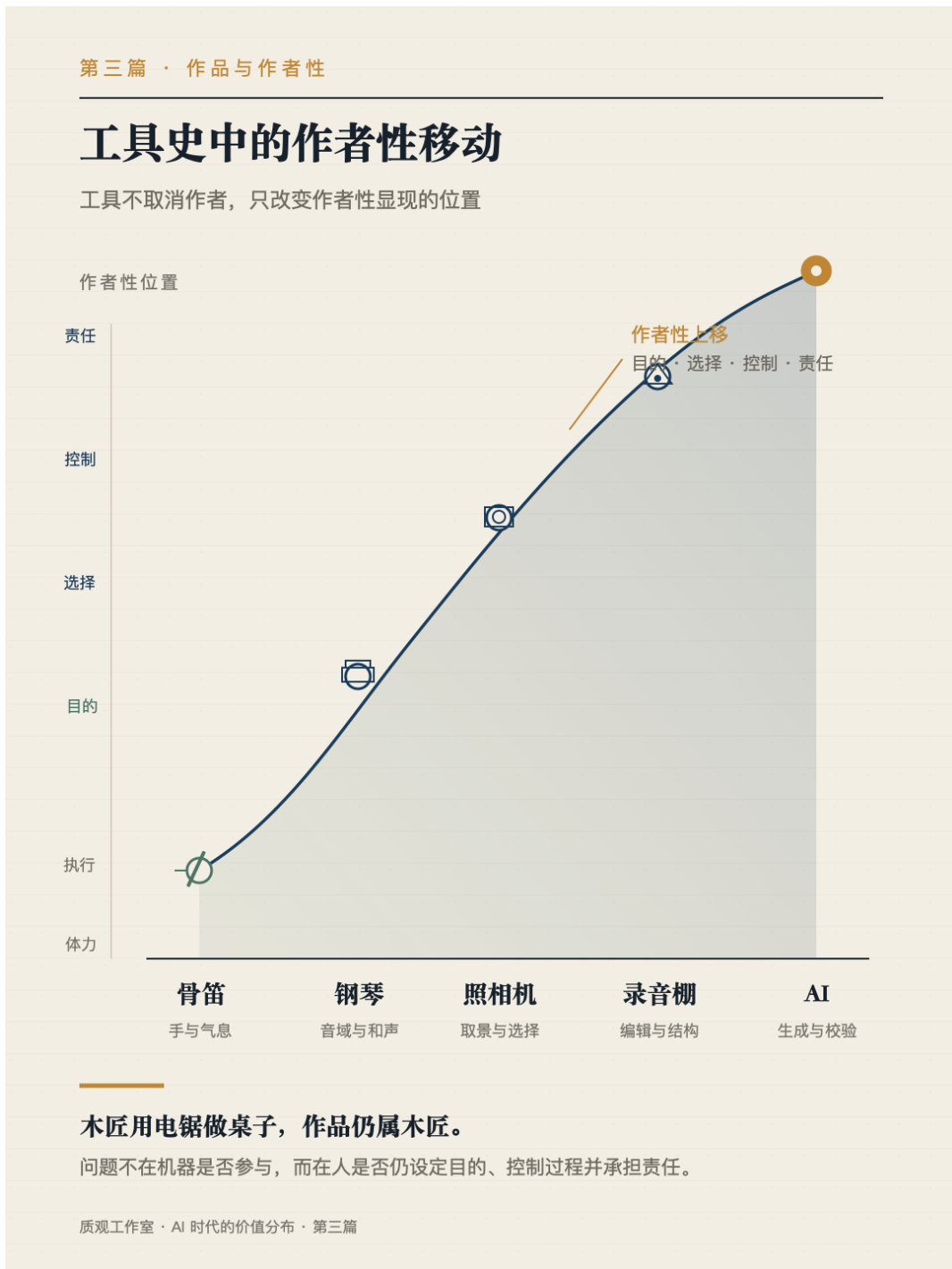


图 1 工具史中作者性的位置移动

本文借用一个来自价格数据的前提：智能的生产成本正在快速下降。据 Epoch AI 测算，达到 GPT-4 水平的推理价格，约以每年 40 倍的速度下降。当“生成”越来越便宜，单问“人会不会被取代”已经不够。还需要追问：

人的哪一部分，无法被生成？

本文分三步回答。先看作品：机器参与之后，什么仍然算人的作品。再看劳动：劳动为什么对人有意义，又怎样失去意义。最后回到任务与位置：AI 最先拿走什么，人应当站在哪里，又靠什么手艺站稳。

二、全世界的法庭遇到同一道题

“机器参与之后，什么才算人的作品？”这个九千年前便已存在的问题，如今成了一道全球立法与司法题。各法域几乎同时遇到它，答卷并不完全相同。

美国的答案较为收紧。

美国版权局 2025 年关于生成式 AI 可版权性的报告认为，仅提供提示词（即使提示词非常详细、经过多轮修改）通常仍不足以构成人类作者身份。不过，人类若对 AI 输出进行了创造性修改、选择和编排，或把生成材料纳入更大的人工创作，作品并不会因此自动失去保护。同年，哥伦比亚特区巡回上诉法院在 *Thaler v. Perlmutter* 案中维持了“作者必须是人”的底线。

中国法院给出的答案更宽。

北京互联网法院在 2023 年 AI 绘图第一案中认定，原告通过提示词设计、参数调整、多轮迭代和结果筛选，体现了独创性的智力投入，涉案图片可以获得版权保护。常熟法院 2025 年进一步确认，若使用者对 AI 输出进行了充分思考，并实施了创造性控制，便可能具备作者资格。

英国采取的是第三条路径。

英国《版权、外观设计与专利法》第 9(3) 条早在 1988 年便为“计算机生成作品”设置了拟制作者：作者是“为创作进行必要安排之人”。英国政府目前正在评估，这一条款在生成式 AI 时代是否仍然适用。欧盟 AI Act 第 50 条主要要求对 AI 生成内容作透明度标识，版权讨论则逐渐聚焦于“实质性人类投入”。日本文化厅 2024 年发布的指引也以“人的创造性贡献”为边界：没有人类创造性参与的自动生成物，不构成著作物。表面上看，中美给出了相反答案：同样是提示词，美国版权局认为通常不够，北京互联网法院认为在特定条件下可以够。

分歧之下，两边实际上在使用同一套检验：

- 人是否设定了明确目的；
- 人是否进行了有意义的选择；
- 人是否对过程和结果实施了创造性控制；

- 人是否能够解释并承担作品的后果。

各法域争论的是“多少人类投入才算够”，不是“AI 能不能成为作者”。目前，没有任何法域承认 AI 本身可以成为作者。这条正在全球成形的边界，也是本文后文“责任链”的法律镜像。法律所谓“独创性智力投入”，正是在为目的、选择、控制和责任称重。若一个人只是输入一句提示词，无法说明为什么需要这个结果，也无法判断结果好坏，不进行修改和筛选，更不承担后果，那么这个输出很难称为完整的人类作品。反过来，AI 的参与本身并不会贬低作品。

人的目的、选择和责任一旦全部缺席，作品才会被贬低。



图 2 生成可以外包，选择与责任不能

中部 劳动：意义、异化与主体性

三、劳动的三种形态

作品的问题解决了一半，接下来转向劳动这一侧。马克思讨论劳动异化，不是在说“劳动很辛苦”。他关心的是：劳动原本应当是人把自身能力对象化到世界中的过程；但在特定生产关系下，劳动产品、劳动过程、人的类本质以及人与人的关系，都会反过来成为支配人的异己力量。这为 AI 时代提供了一条不同于“会不会失业”的判断尺度。沿着这条尺度，可以把劳动分成三种形态。

1. 附属型劳动：人被压缩成流程零件

最低的一层，是人把身体和时间交给外部系统，却不拥有目的、过程和结果。这样的人，即使尚未被 AI 替代，也已经只是流程中的可替换件。仓库计时器、外包平台评分、内容工厂的日更指标、客服脚本和自动绩效看板，都可能把人变成机器系统的附属物。

阿伦特在《人的境况》中，把人的活动分成三类：

- 劳动 (labor) 维持生命，消耗之后不留下持久之物；
- 制作 (work) 创造相对持久的对象，把意图凝固在世界中；
- 行动 (action) 是在他人面前开启新事物，并为后果作出承诺。

附属型劳动，便是人被压缩到纯粹 labor 的状态：重复、耗尽、没有作品留下，也很难在结果中认出自己。布雷弗曼在《劳动与垄断资本》中指出了这种压缩的机制：泰勒制的要点，是把“构想”与“执行”分开。构想被集中到管理层，执行被留给流水线，劳动者的技能由此被系统性剥离。这一诊断放到今天仍然成立。若构想集中到模型和平台，人只负责点击、审核、补漏和承受考核，这就是 AI 版本的去技能化。但同一种技术也可以走向另一方向：把重复执行交给机器，把构想和判断收回给人。此时，AI 才是增强，而不是替代。

2. 技能型劳动：人理解对象，而不只是重复动作

更高的一层，是技能。木匠知道木材如何变形，厨师知道火候与食材怎样相遇，工程师知道系统为什么会在边界条件下崩溃，医生知道同一组指标放在不同病人身上，不意味着同一件事。技能的要害，是人对对象的理解。重复动作只是外壳。中文传统里，有两个很贴切的样本。庖丁解牛说：“臣之所好者道也，进乎技矣。”十九年后，刀刃仍像刚磨过一样锋利，靠的是对牛体结构的把握，即他所说的“依乎天理”。轮扁斲轮则说：“得之于手而应于心，口不能言，有数存焉于其间。”这位七十岁的轮匠甚至无法把手艺完整传给儿子。波兰尼后来把这种“能做却说不清”的知识称为默会知识。它与 AI 的关系很直接：越容易被写成文本、流程和示范的部分，越容易被模型学习；留在身体、现场、经验与判断之间的部分，则更难被完整生成。

技能的护城河，就画在轮扁说不出的那一段。

3. 生成型劳动：人在产品中认出自己

技能再往上，是生成性的劳动。世界给出材料、问题、阻力和节奏，人则以身体、经验、判断和工具回应。最终的作品不是“人脑中一个纯粹想法”的直接复制，也不是机器自动吐出的结果；它来自人与世界的相互作用。好的劳动让人能够在产品中认出自己。桌子、菜肴、舞蹈、文章、代码、诊断和制度设计，都是某种被对象化的生命活动。黑格尔在主奴辩证法中为这种劳动给出了哲学根据。奴隶通过劳动塑造事物，也在对象中看见自己的力量，由此形成自我意识。主人不劳动、只享用成果，看似自由，却在依赖中变得空洞。这个结构对 AI 时代构成一个倒转的警告：

把全部劳动外包给机器的人，即便坐上主人的位置，也可能继承主人的命运。

判断力不是凭空出现的天赋，而是劳动留下的沉积物。若跳过劳动，判断便无从形成。这不是在浪漫化劳动。马克思的异化理论说明，机器本身不是问题。问题在于：机器由谁拥有，为谁服务，又怎样组织劳动过程。如果 AI 把人从重复工作中释放出来，让人回到目的、判断、照料和创造，它更接近工具。如果 AI 被用于压低人的议价权、剥夺训练机会、吞掉作品归属，并让人只负责修补系统错误，它就会成为新的异化装置。

斯蒂格勒把技术的这种两面性称为“药毒同体”（pharmakon）。文字外置了记忆，机器外置了手艺，AI 又外置了表达与分析。每一次把能力交给技术，都可能带来知识从人身上的流失；除非这种外置同时伴随着更高层次的再内化。文字没有毁掉思想，因为人类后来发明了与文字相匹配的阅读和写作训练。AI 是否会毁掉判断，也取决于我们能否发明出与它相匹配的训练。人与 AI 的分界，不在“人高贵、机器低级”，也不在“机器先进、人已经过时”。可检验的分界只有一条：

劳动中，人的主体性是否仍然存在。

把主体性说完整，可以归结为四件事：1. 目的是否由人确定；2. 过程是否由人理解；3. 结果是否由人选择；4. 责任是否由人承担。这四件事是全文的轴。后文将把它们简称为“主体性四要素”。



图 3 劳动的三种形态与主体性四要素

四、李善德：一台异化机器的全功率运转

主体性四要素不是抽象清单。文学给过我们一个具体样本，主角是唐朝的一名九品小吏。马伯庸《长安的荔枝》中，天宝年间的上林署监事李善德被同僚设局，接下一桩几乎必死的差事：赶在贵妃生辰之前，把岭南鲜荔枝运到长安。白居易《荔枝图序》写得清楚：荔枝离枝后，“一日而色变，二日而香变，三日而味变”。而岭南距长安五千余里。

用上一章的语言说，这是一台异化机器的全功率运转：

- 目的不由李善德确定——贵妃要吃；
- 过程几乎不由他选择——敕牒已经下达；
- 结果没有选择余地——办不成便是死罪；
- 责任却全部落在他身上。

韦伯所谓家产官僚制的要害，正在这里：它有科层制的外表，有层级、公文和分工；但骨子里，一人的私欲可以成为最高命令，所有规则在“上头要”三个字面前都会变软。链条上的每个人既被碾压，也在碾压别人。替李善德垫资奔走的胡商被过河拆桥，献出荔枝树的峒女眼看果园被砍。吴思把这种制度零件称为“合法伤害权”。机器最深的一道机关，是让人亲自完成最后的驯化。李善德一度不断自责：“是我拖累了大家。”但他没有先问，那道题本身是否就是无解的。尼采在《道德的谱系》中追溯过这种内疚的来源：向外的力量被堵死后，便转而向内撕咬。福柯所说的规训，其最终形态也是人成为自己的狱卒。

小说的动人之处在于，李善德后来用两次行动，把主体性一点点夺了回来。第一次，他重新拾起算学，把“不可能”拆成保鲜时限、驿道脚程和分段接驳。对过程的理解，被他拿了回来。他又说：“就算失败，我也想知道，自己倒在距离终点多远的地方。”评价标准不再完全由机器规定。事情是否成功仍不由他决定，但如何完成、如何理解这一程，重新归他所有。第二次，任务完成后，他向杨国忠说出了那句多余的实话，并接受流放。这个结果是他自己选择的，后果也是他自己承担的。李善德没有战胜那台机器，也没有能力改变整个制度。但他证明了主体性的最小形态：

人可以被抛进死局，被要求交出时间、体力，甚至性命，却未必要把判断、尊严和自我一并交出去。这个唐代样本对 AI 时代的意义在于：异化不必等到 AI 出现，AI 也不会自动制造异化。驿路、敕牒、KPI、自动绩效看板和平台派单算法，只是同一台机器在不同时代的零件。决定人的位置的，始终是主体性四要素掌握在谁手里。下一章把它进一步拆成一条可以逐环检验的责任链。

五、责任链：工具如何成为身体的一部分

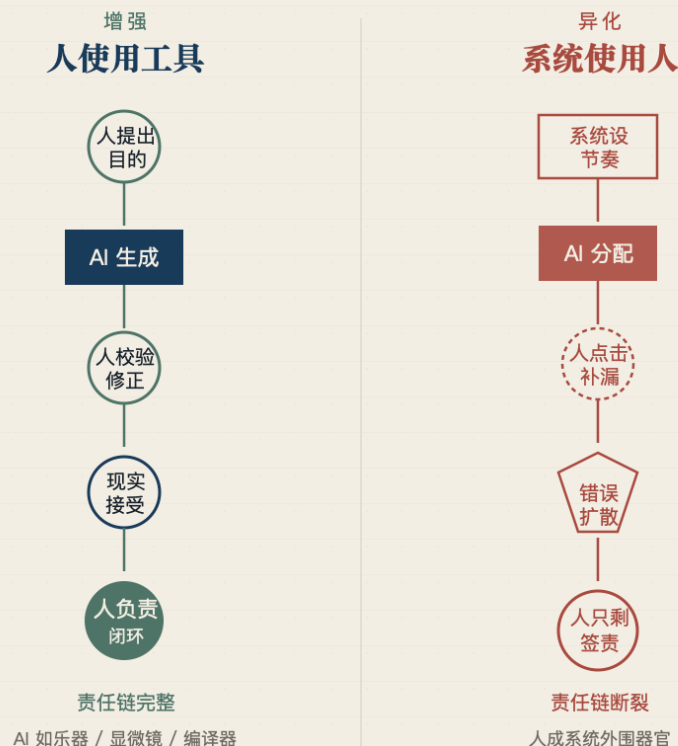
为了避免把人机关系写成空泛的人文宣言，可以把它拆成一条责任链：

人提出目的 → 工具扩展能力 → AI 执行局部生成 → 现实结果接受检验 → 责任回到人和组织。

第三篇 · 责任链

责任链的两种命运

同一工具，两种位置——主体或附属物



AI 重新划分：人是主体，还是附属物

位置不由“有没有 AI”决定，而由那四件事掌握在谁手里决定

链不断，人仍是使用者与责任人。

质观工作室 · AI 时代的价值分布 · 第三篇

图 4 责任链的两种命运

只要这条链没有断，人仍然是系统的使用者、校验者和责任人。一旦责任链断裂，人便可能只剩下点击、审核、背锅，或被系统调度。即使他仍然坐在岗位上，也已经被降格为系统的外围器官。技术哲学为这条链提供了更准确的语言。Clark 与 Chalmers 的“扩展心智”理论认为，人在解决问题时，身体之外的外部结构可以成为认知过程的一部分。Don Ihde 对“技术具身”的分析也指出，熟练掌握的工具会退入背景，成为人感知和行动世界的方式。眼镜、汽车、显微镜、机械臂、乐器、相机和软件界面，都可能被人纳入身体与认知。

熟练驾驶者会把汽车的边界感纳入身体；摄影师会通过镜头观看世界；音乐人会把控制器、采样器和软件变成声音器官；优秀工程师也会把调试器、日志系统和 AI 代码助手纳入自己的判断链。工具不只是身体外部的死物。它也可以成为人的行动结构的一部分。但“扩展”不自动意味着进步。Ihde 同时提醒，技术总会放大一些能力，也会缩减另一些能力。AI 放大了生成、搜索、改写、模拟和草拟的能力，却可能削弱人的耐心、记忆、基础训练、现场观察和责任感。一个人如果只通过 AI 看世界，很快就会把 AI 的默认分类误认为世界本身，把模型的流畅回答误认为自己的判断。于是，人机关系分成两个相反方向。

工具性增强

人设定目的，AI 完成局部任务，人校验和修正结果，最后把结果带入现实世界。此时，AI 类似更复杂的乐器、显微镜或编译器，是人的能力延伸。

系统性异化

平台、资本或组织设定目的，AI 分配任务、规定节奏和生成标准，人只负责追随、补漏与承受后果。此时，使用机器的主体不再是人，而是系统。马克思在机器大工业中担心的，正是人从工具使用者变成机器系统的附属物。AI 不会原样复制十九世纪工厂，却可能在知识劳动中制造类似结构：人不再拥有问题，只被要求按照系统节拍生成、审核、标注、纠错并承担 KPI。李善德就是责任链断裂后的完整记录：目的在贵妃，节奏在驿路，责任在小吏。沿着这条责任链，可以得到一个判断：

AI 不会简单消灭人，但会重新划分人是主体，还是系统的附属物。

人的位置，不取决于“有没有 AI”，而取决于主体性四要素掌握在谁手里。

下部 任务、位置与手艺

六、把职业拆开：AI 最先拿走哪些任务

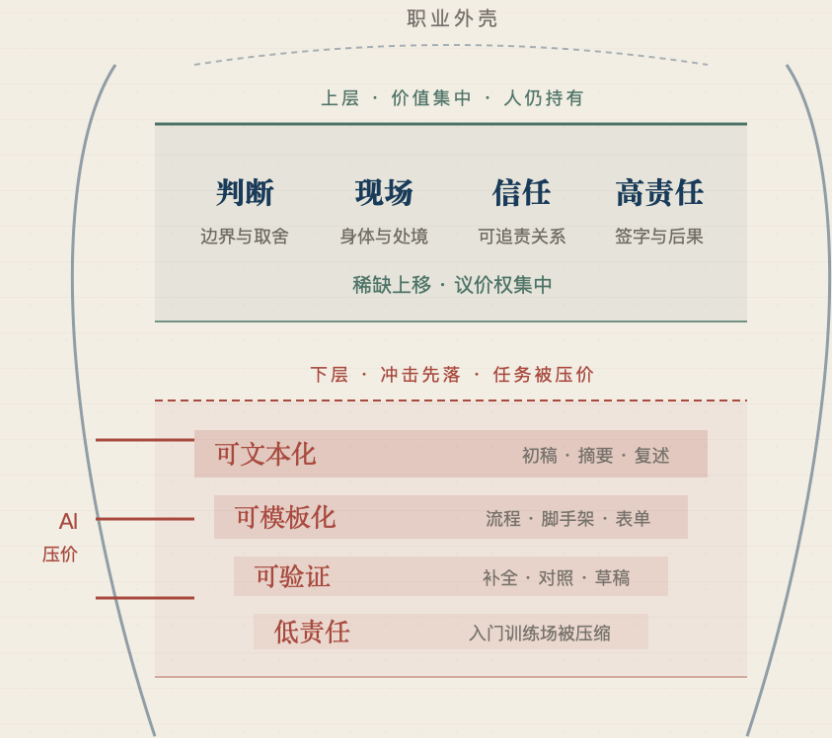
“人会不会被 AI 取代？”这个问题的首要偏差，是分析单位选错了。职业并非一块完整的铁板，而是一组任务的组合。人也不是一台输出机器，而是目的、身体、经验、关系和责任的结合。需要追问的，不是职业是否整体消失，而是职业内部的任务将如何重组。现有研究大体指向同一结论：

AI 首先重组任务，而不是整齐地替换整个职业。

第三篇 · 任务与职业

AI 替代任务，而非职业

职业是任务的组合；冲击先落在可编码、低责任的部分



AI 重组任务，而非整齐替换职业。

暴露不等于失业；任务变快也不等于职业消失。

砍掉入门任务，当期效率上升，下一代专家供给收缩。

分析单位选错，才会问“整份职业会不会消失”。

质观工作室 · AI 时代的价值分布 · 第三篇

图 5 AI 替代任务，而非整个职业

OpenAI、OpenResearch 与宾夕法尼亚大学 2023 年发布的《GPTs are GPTs》从任务暴露度出发估算：约 80% 的美国劳动力，至少有 10% 的工作任务可能受到 GPT 影响；约 19% 的劳动者，至少有一半工作任务可能受到影响。若把 LLM 之外的软件工具也计算在内，受影响任务比例还会更高。但论文作者同时强调：这只是暴露度测算，不是采用时间表，更不是失业预测。任务变快，不等于职业消失。

Autor 在《为什么仍有这么多工作》中提供了一个更稳定的解释框架：自动化会替代部分劳动，也会补充劳动、扩大产出，并提高机器难以完成的任务的价值。Acemoglu 与 Autor 的任务模型进一步说明，技术冲击应落到任务层面观察。机器替代一部分任务之后，会改变不同劳动者的比较优势和工资结

构，却不会按职业名称整齐地一排排消灭岗位。按照这一框架，最先被替代的并不一定是社会通常所谓的“低端工作”。更先受到影响的，是那些高度可编码、容易检查、责任成本较低的任务。

1. 标准化文字生产首先被压价

Noy 与 Zhang 在 Science 发表的实验显示，ChatGPT 可使中等难度的专业写作任务耗时下降 40%，质量提高 18%。邮件、摘要、会议纪要、宣传文案、报告初稿和常规说明书，都会被重新定价。AI 未必理解最终事实，却能大幅降低“先写出一份像样初稿”的成本。

2. 信息搜集、整理和复述被快速重组

Microsoft Research 2025 年分析 20 万条匿名 Bing Copilot 对话后发现，用户最常请 AI 协助的活动是信息搜集和写作；AI 最常执行的活动则是提供信息、写作、教学和建议。AI 适用度较高的职业群体，集中在计算机与数学、办公室行政支持和销售等需要大量信息沟通的岗位。这类冲击通常不表现为整个职业消失，而是“查、抄、整、复述、初步归纳”这些环节不再稀缺。

3. 受约束的代码任务被压缩

GitHub Copilot 的控制实验显示，开发者完成指定 JavaScript 任务的速度提高了 55.8%。AI 对代码补全、脚手架、接口调用、简单 bug 修复、单元测试草稿和文档生成非常有效。但软件开发并不等于写出代码。需求拆解、系统边界、上线责任、长期维护成本和团队协作，仍然无法被一次代码生成取代。

4. 标准客服与流程问答更快自动化

Brynjolfsson、Li、Raymond 对客服场景的研究显示，生成式 AI 助手使客服每小时解决的问题数平均提高 14%，而新手和低技能员工的提升更明显。在高度标准化的场景中，人工坐席需求会受到压缩。但问题一旦涉及情绪、争议、授权、赔偿与责任，仍需要人介入。

5. 模板化专业服务被改造

HBS 与 BCG 的“Jagged Frontier”实验显示，在 AI 能力边界内，顾问完成的任务数量提高 12.2%，速度提高 25.1%，质量也有所提升；但一旦任务落在能力边界之外，AI 反而会诱导错误。因此，专业的变化不是“顾问消失”，而是资料搜集、框架初稿、案例对照、表格生成和风格统一被自动化；判断边界、理解客户情境、签署责任和给出最终建议，则会变得更贵。

6. 年轻人的入门任务受到最直接冲击

Stanford Digital Economy Lab 2025 年研究发现，生成式 AI 广泛采用后，22 至 25 岁、处于高 AI 暴露职业中的早期职业者，就业出现约 16% 的相对下降。同一职业中，经验更丰富的劳动者则保持稳定

或继续增长。调整主要发生在就业数量，而非工资，并且集中在 AI 更偏向“自动化”而不是“增强”的岗位中。另一个后果是，AI 最先压缩的，可能正是过去用来训练新人的入门任务。回想黑格尔的警告：判断力是劳动的沉积物。专家不是直接从课堂里毕业出来的，而是在千百次低风险错误中成长起来的。入门任务，是人合法犯错、积累判断的场所。砍掉这些任务，当期效率会立刻提高；但下一代专家的供给，也可能在几年后悄悄收缩。

七、不可外包的六样东西

AI 会替代一部分任务，剩下的任务也不会天然安全。难以外包的工作，价值通常来自现实世界的约束，而非一次输出。这样的东西至少有六样。

1. 问题定义

AI 可以回答问题，但高价值的问题，往往并没有被事先定义好。企业最严重的错误，常出在问错问题、优化错指标，或把资源投入到并不重要的环节。两个企业都可以调用同一个模型。差距在于能否识别瓶颈、选择接入流程，并判断哪些结果值得下注。

2. 责任承担

在高风险场景中，答案只是中间产出，客户最终购买的是责任承担。监管文本已经把这一点写进制度。欧盟 AI Act 第 14 条要求高风险 AI 系统具备人类监督，以降低对健康、安全和基本权利的风险。FDA 关于临床决策支持软件的指南，以医疗专业人员能否独立审查建议依据来划定监管边界。NHTSA 对自动驾驶等级的说明中，L2 仍要求驾驶者保持投入并负责监控，L3 也要求人在必要时接管。各项规定口径不同，但方向一致：

风险越高，越需要一个可追责的人类位置。

3. 信任关系

客户购买律师、医生、顾问、投资经理、教师、教练和销售，购买的不只是知识输出，还包括一个能够理解处境、接受质询并承担责任的人。AI 可以生成合同草稿，但客户仍会问律师是否愿意负责；AI 可以提供诊断建议，患者仍需要医生解释风险；AI 可以生成投资分析，委托人仍然需要知道，谁在承担判断后果。

4. 跨组织协调

许多工作的难点，不是“知道应该怎么做”，而是“让组织真的去做”。流程改造涉及预算、权限、部门利益、员工抵触、合规要求、系统迁移和责任划分。AI 可以生成方案，却无法自动让销售、财务、法务、IT 与生产部门接受同一套行动。组织协调需要谈判、妥协、时机判断与现场权威。

5. 身体性与现场处置

生成式 AI 最擅长数字环境中的文本、图像、代码和结构化信息。但真实世界的工作，还包含噪音、设备差异、环境变化、临场安全和身体动作。电工、护士、厨师、维修工程师、建筑工人、物流调度和现场销售都会被 AI 改变，但主要价值并不完全存在于屏幕上。机器人会持续进步，但成本、可靠性、环境适配和安全责任，会使这类替代远慢于文本任务。

6. 价值判断

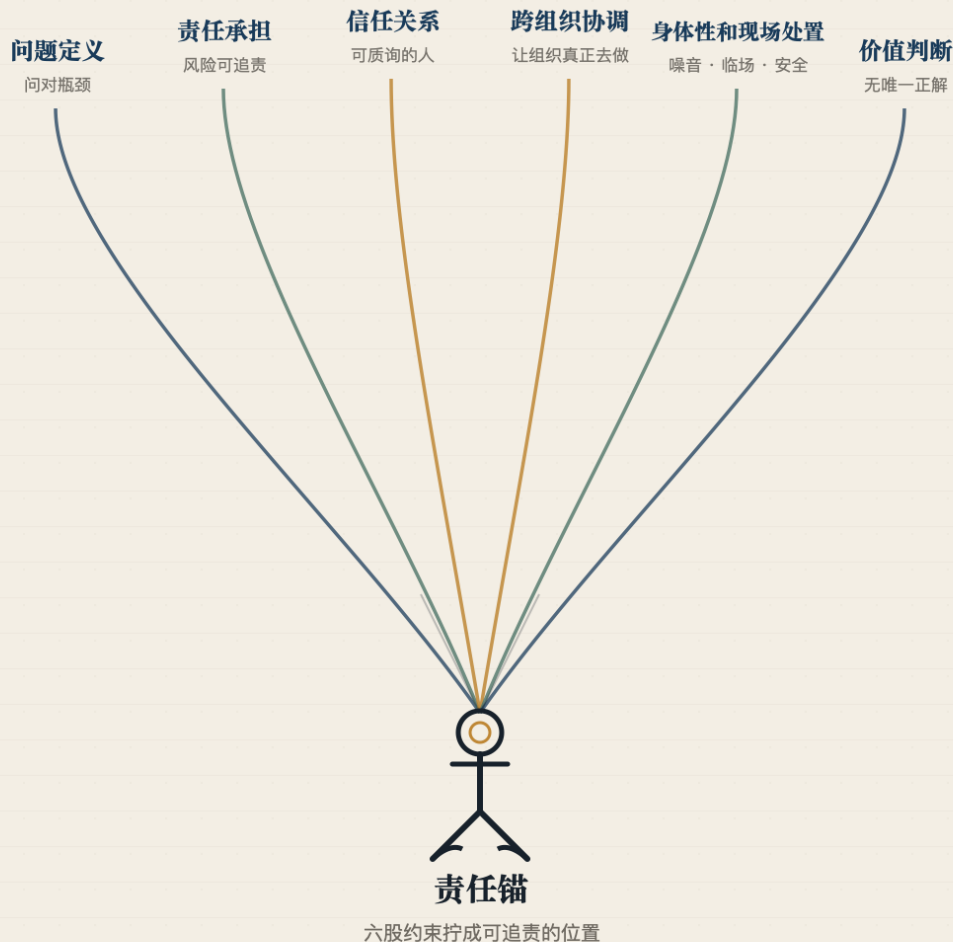
很多问题没有唯一正确答案：裁员还是培训，效率还是公平，增长还是安全，商业利益还是用户伤害，短期利润还是长期信誉。AI 可以罗列方案与权衡，却不能替社会承担最终选择。OECD AI Principles 强调 AI 应尊重法治、人权、民主价值和以人为中心的原则，指向的也是这个约束：

社会不能把价值选择完全交给一个概率系统。

第三篇 · 不可外包

六种不可外包

难外包因真实世界约束，非机器能力上限



价值来自真实世界约束，而非单次输出。

质观工作室 · AI 时代的价值分布 · 第三篇

图 6 六种不可外包能力汇成责任锚

八、人的位置：三种能力，四个层面

把劳动哲学与任务经济学合在一起，可以得到一个比“人不会被取代”更严格的判断：AI 会降低人在标准化生产中的稀缺性，却会提高人在目的、判断、整合、信任和责任中的价值。过去，很多人的价值来自熟练生产：会写、会算、会查、会整理、会画图、会写代码、会做 PPT、会生成报告。AI 正在直接压低这部分能力的市场价格。如果一个人只把“我能产出内容”当作核心价值，就会很快遇到一个更便

宜、更快、更稳定的系统。这不是努力不足的问题，而是单纯生产本身不再稀缺。未来更稀缺的是三种能力。

1. 判断能力

判断不等于“比 AI 知道得更多”。它要求人知道何时该使用 AI、何时不该使用；知道答案是否可信，也知道问题是否值得解决。判断来自领域经验、失败记忆、利益关系的理解，以及对不可量化风险的敏感。它不与 AI 竞争，而是站在使用 AI 的上一层。

2. 整合能力

AI 可以完成许多局部任务，但真实业务不是局部最优的简单相加。整合，是把模型接入流程，把数据接入权限，把输出接入行动，再把行动接入反馈。一个人或团队若能让 AI 从“生成答案”进入“改变流程”，其价值就不再限于劳动力，而更接近对流程本身的定价能力。本系列第一篇将其称为“流程收费权”。

3. 责任能力

责任不是一句抽象口号。它意味着能够被客户、公司、监管者、市场和团队追问：谁签字？谁解释？谁处理失败？谁面对客户？谁在不确定时作出选择？AI 可以降低决策成本，却不能取代责任位置。责任越昂贵，人越不会消失。这三种能力落到四个层面，含义不同。

对个人

危险不在于“自己的职业被 AI 点名”，而是工作内容几乎全部由可替代任务构成。只写通用稿的写作者、只搬运资料的分析师、只写模板代码的程序员、只做重复发布的运营，都会被压价。相反，能够提出选题、建立资料体系、定义判断框架、承担客户关系，并把结果转化为行动的人，更容易从 AI 中受益。

对企业

“少雇几个人”是 AI 带来的表层变化，进一步的变化是重新设计岗位。

企业需要区分三类任务：

- 可以完全自动化的任务；
- 适合人机协同增强的任务；
- 必须由人承担责任的任務。

不作区分，企业会犯两种相反的错误：要么把人长期困在低价值重复劳动里，要么把 AI 放进不该由系统托管的责任位置。

对教育与训练体系

最大的风险，是入门任务消失。如果 AI 接管整理、模仿、草稿、改写、简单编码和基础检索，年轻人便可能失去通过低风险工作积累经验的路径。未来的训练应更早引入真实问题、项目责任、跨学科协作和 AI 校验能力。教育重点不应只放在“更会写答案”，还应转向“更会判断、追问和使用答案”。

对社会

AI 不会自动带来更好的分工。World Economic Forum 2025 年报告预计，到 2030 年，22% 的工作将受到结构性扰动，近 40% 的岗位技能会发生变化。报告预计岗位净增约 7800 万个：新增 1.70 亿个岗位，同时替代 9200 万个岗位。Goldman Sachs 2026 年对美国劳动市场的基准估算是：若 AI 在约十年内广泛采用，约 6% 至 7% 的劳动力可能在转型期间被替代；不同假设下，区间为 3% 至 14%。技术创造新岗位，并不意味着每个被替代的人都能无成本转移。人不会因为 AI 自动获得安全，也不会因为 AI 自动消失。人的位置，取决于主体性四要素是否还掌握在自己手里。AI 越强，越会压低单纯生产的价格，也越会抬高目的、判断、控制和责任的位置。

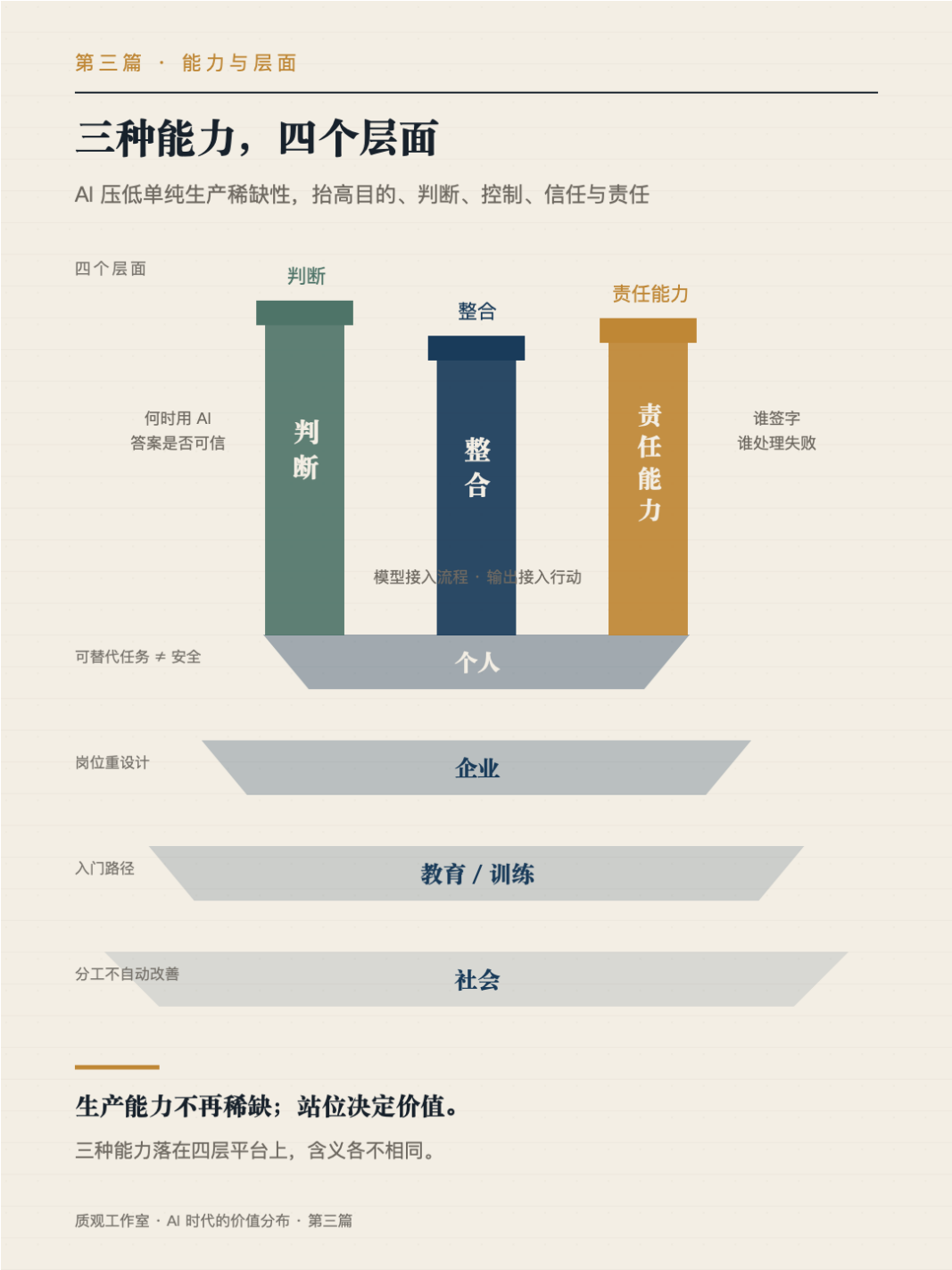


图 7 三种能力，四个层面

九、驾驭工程：未来劳动的形态

开篇讨论作者性会随工具移动位置。中部讨论劳动的意义在于对象化。这两条线在 2026 年出现了一个具体交汇点。工程界称它为 harness。这个词的字面义是马具、缰辔，行业常用的定义也带着驭手色彩：

Agent = Model + Harness。

本文不采用“缰绳”式的字面翻译，而将其译为驾驭工程。

它早已超出单件工具，成为一套完整实践：

- 任务如何拆解；
- 工具何时调用；
- 输出怎样验证；
- 权限止于何处；
- 失败留下什么记录；
- 同一错误如何避免再次发生。

Mitchell Hashimoto 在 2026 年 2 月给出的命名定义，几乎就是一部驭手手册：每当 Agent 犯错，工程师就建立一套机制，使这一错误不再重复；这些机制的总和，就是 harness。几天后，OpenAI Codex 团队独立提出了同一公式。其背景是，该团队在五个月内交付了约一百万行由 AI 生成、几乎无需人工逐行编写的生产代码。卡内基梅隆、耶鲁等机构二十位作者的综述论文，把 harness 定义为“包裹语言模型、管理长程任务执行的基础设施层”。本章行业数据转引自 2026 年 6 月的行业研究报告，一手出处见文末参考文献，正式发表前仍需回源。

1. 模型相同，驾驭工程不同，结果可以相差一个数量级

证据先来自基准测试。约束 Agent 可靠性的，未必是模型能力本身。对照实验显示，在完全不修改模型参数的情况下，仅改变编辑工具输出格式，同一模型在编码基准上的成绩就可能提升至十倍。其中一个模型的成绩从 6.7% 跃升至 68.3%。LangChain 团队只增加了一项“自验证钩子”，便使固定模型在 Terminal-Bench 上的得分提高了 13.7 个百分点。同一模型，经过不同驾驭方式，可能获得数量级不同的结果。

2. 工具接入不值钱，长期可复用的驾驭方式才值钱

劳动市场数据提供了同一结论的另一面。在实验中，AI 确实提高了效率：客服坐席平均每小时多解决 14% 的问题，中等难度的专业写作节省约四成时间。但丹麦研究者使用覆盖两万五千名劳动者的全员行政数据，测量 AI 聊天工具的真实劳动市场影响，结果显示：采用两年后，工资和工时的平均变化都接近零。收益差异首先来自测量对象。实验测的是接入工具那一刻，某项任务能快多少；工资测的是市场愿意长期为一个人多付多少钱。当人人都能打开同一个对话框时，“打开对话框”本身便不再值钱。经济学家研究电力和信息技术时，早已观察到类似时间差：通用技术的收益，要等互补资本建成后才会显现。在此之前，生产率账面可能长时间保持平静。因此，零效应不能证明 AI 无用。它更可能说明：工具接入已经普及，但个人与组织的驾驭工程尚未建成。

3. token 是空白算力，驾驭工程把它变成具体劳动

实验与劳动市场数据之间的落差，导向一个哲学判断。token 是按量计价的智能：每百万 token 一个价格，没有个性，没有方向，也没有偏好。可以把它理解为一种空白算力。马克思分析商品时，区分了具体劳动与抽象劳动。具体劳动各有质地，木匠劳动和裁缝劳动不能简单互换；抽象劳动则把差异抹平，化为可计量的无差别劳动时间。token 很像机器智能的“抽象劳动”形态：被抹平、可计量、随处可租。这里是类比，而不是等式。token 并不形成马克思意义上的价值。但这个类比足以说明：抽象的智能不会自动变成任何具体的社会劳动。把空白算力变成“这一个人的工作”的，是驾驭工程。

其中每一部分，都带着使用者的指纹：

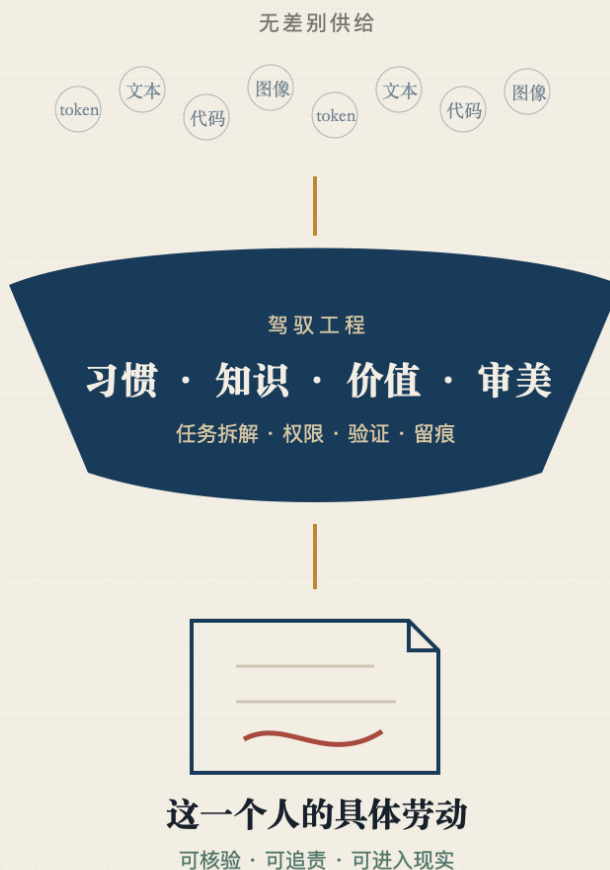
- 习惯：工作的节奏，任务的拆法，检查点设置在哪里；
- 知识储备：领域经验、语料、案例和失败记录；
- 价值方向：什么该做，什么不能做，红线画在哪里；
- 审美方向：什么算好，什么算俗，修改到什么程度才停止。

同一个模型，接上不同的驾驭工程，便成为不同的劳动者。

第三篇 · 驾驭工程

模型可租，稀缺的是自己的驾驭工程

空白算力只有经过习惯、知识、价值与审美的组织，才成为具体劳动。



AI 提供生成能力，人把它组织成可承担后果的工作。

质观工作室 · AI 时代的价值分布 · 第三篇

图 8 空白算力与驾驭工程：具体劳动的新构成

4. 驾驭工程是一种“二阶对象化”

中部的对象化理论，在这里还可以向前推进一步。过去，人把能力直接对象化到产品中：桌子、文章、诊断、代码。现在，中间多了一层：

人先把习惯、知识、价值与审美对象化到驾驭工程中，再由驾驭工程调度空白算力，把它们对象化到产品中。

这是一种二阶对象化。驾驭工程本身就是一件作品，一件以人格、经验和失败记录为材料的作品。未来的劳动，未必会被 AI 整体替代。更可能发生的是劳动形态上移：直接执行交给算力，人则负责建造、维护和修订自己的驾驭工程。社会分工的单位也可能随之改变。斯密时代，分工的单位是工序；涂尔干所谓有机团结，则建立在千百种职业的相互依赖之上。当空白算力人人可租，人与人之间的差异便不再只由岗位名称承载，而会越来越多地由一套套不同的驾驭工程承载。一个人在社会分工中的位置，将越来越接近其驾驭工程的独特性。

5. 它填补了资本谱系中的一块空地

还需要把这样一套工程放回资本谱系中观察。

经济学已经命名了两类资本：

- 技能与判断存在于身体中，随人移动、随人衰老，贝克尔称之为人力资本；
- 流程与制度存在于组织中，属于公司而非个人，研究者称之为组织资本。

如果把“体内 / 体外”与“个人 / 集体”交叉组合，会得到四个格子。其中有一格长期处于空白状态：

外在于身体，却属于个人的资本。

过去，这一格中只有笔记、模板和顺手的工具箱，规模太小，不值得单独命名。驾驭工程改变了这一点。外化物不再只是辅助劳动，而开始直接执行劳动。习惯、规则和语料不再只是备忘录，而成为一台能够替你工作的机器的操作系统。人力资本因此第一次拥有了一种可以“拿在手里”的形态：它会随使用不断增厚，可以迁移，也可能被平台锁住。



图9 资本的四象限：驾驭工程填补的空格

6. 驾驭工程生产的不是答案，而是证据

“人的位置等于其驾驭工程的独特性”，这一判断也有信任数据支撑。约六成知识工作者已经使用 AI，但完全放手委托的比例只有 0 至 20%；八成以上开发者日常使用 AI，对输出保持高度信任者却只有约 3%。业内把这一现象称为“委托差距”。根源通常不是模型是否足够聪明，而是缺少可验证的决策证据。生成越来越便宜，验证反而成为瓶颈。驾驭工程的经济学位置就在这里：它生产答案，也生产答案为什么可信、过程如何复现、失败如何追责的证据。

证据，是责任能力的工程形态。

工程界归纳过五项设计原则：1. 显式声明资源与目的； 2. 全程可追踪； 3. 用确定性证据验证结果，而不是相信模型自称“已完成”； 4. 失败后先归因，再修复； 5. 警惕维护负担不断累积。换成劳动主体的语言，这五条仍然指向主体性四要素。Anthropic 的研究还发现：Agent 无法稳定地评估自己的工作，验证必须由独立环节承担。生成与验证的分离，恰好是布雷弗曼所批判的“构想与执行分离”的反向操作。这一次，被交出去的是执行，被人重新收回的是判断。

7. 三个必须写足的风险

驾驭工程并非天然具有解放性。它至少带来三类风险。

第一，认知债务。

如果组织跳过驾驭工程，直接把任务放手交给 AI，AI 辅助代码在第二年的维护成本可能达到传统开发的四倍。斯蒂格勒所说的“无产阶级化”，在这里表现为一项真实的成本。只会复制他人模板的驾驭工程，承载的是别人的经验，而不是你的具体性。

第二，安全风险。

2026 年上半年 Agent 生态中出现的供应链投毒和漏洞事件说明，缰绳本身也需要约束。单一开源 harness 曾在一个月内披露九个以上安全漏洞（CVE），上千个恶意技能包也曾渗透生态。

第三，也是最重要的，所有权。

驾驭工程是一个人习惯、知识、价值和审美的外化物。如果它的记忆、语料和技能文件被锁在平台中，无法导出、无法迁移，那么异化只是在新的介质上重新发生：人格的对象化物成为别人的资产，人开始对着自己的镜像打工。驾驭工程的可迁移性与所有权，将成为 AI 时代劳动政治的新条款，就像上个世纪的工时与工伤。由此看，“怎样成为超级个体”的答案并不神秘。模型人人可租，token 只是空白算力。稀缺的是一套从自己的工作、经验和失败中长出来的驾驭工程。外壳可以简，技能必须厚。每一行约束，最好都对应一次真实经历过的失败。

它是一门手艺，与庖丁的刀、轮扁的轮属于同一谱系：说得出的部分可以开源共享，说不出的部分仍然存在于手与心之间，存在于自己的失败记录中。行业预计，harness 工程师可能在一年半内成为独立职业。Gartner 同时预测，超过四成 agentic AI 项目会在 2027 年底前被取消。这两条预测并不矛盾。总量会增长，淘汰也会加速；谁能留下来，取决于驾驭工程的质量。至于超级个体在判断、信任和分销三重约束下的上限，系列第一篇已经在产业层面给出答案。

结论：可被生成的部分贬值，责任的位置升值

本文的两条线索，在这里汇合。经济层面的证据显示，AI 重组的是任务。可文本化、可模板化、容易验证、责任较低的部分会最先被压价；入门任务的消失，也让早期职业者首当其冲。哲学讨论则表明，机器参与不会自动取消作者性和劳动意义。从贾湖骨笛，到录音棚，再到 AI，作者性一直在移动：从直接执行，移向目的、选择、控制和责任。取消劳动意义的，不是机器做了多少，而是人退出了这些位置。驾驭工程的出现表明，这种位置移动已经开始变成一门可以学习的手艺。责任链正在被工程化，劳动对象化也获得了新的介质：

人先把自己写进驾驭工程，再由它驾驭空白算力。

两条线索可以合成一句话：

当智能本身越来越便宜，价值不再属于单纯生产智能的人，而属于能够决定智能如何进入现实世界，并承担结果责任的人。

李善德在驿路上证明过的事情，今天在工作流中依然成立。第一篇讨论钱停在哪里，并提出资源、入口和流程三类收费权；第二篇检验资本为这些位置开出的价格是否真实；第三篇回到人本身，回答人应站在哪里。三部曲至此合拢：

钱的位置，价格的真伪，人的位置。

参考文献

工具史与作者性

- Juzhong Zhang et al., Oldest playable musical instruments found at Jiahu early Neolithic site in China, Nature (1999) : <https://www.nature.com/articles/23261>
- MIDI Association, The History of MIDI Collection: <https://midi.org/the-history-of-midi>
- Ableton, A Brief History of the Studio as an Instrument: <https://www.ableton.com/en/blog/studio-as-an-instrument-part-1/>
- U.S. Copyright Office, «Copyright and Artificial Intelligence, Part 2: Copyrightability»: <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf>
- U.S. Copyright Office, AI copyrightability 新闻稿: <https://www.copyright.gov/newsnet/2025/1060.html>
- Thaler v. Perlmutter, 美国哥伦比亚特区巡回上诉法院判决 (2025, 维持人类作者要求; 判决书请自法院网站回源)
- DLA Piper, UK/EU/China AI 生成内容版权保护现状 (2026 Q2) : <https://www.dlapiper.com/en-us/insights/publications/2026/04/ai-and-copyright-major-issues-and-direction-of-travel-in-the-uk-eu-and-china-in-q2-2026/can-ai-generated-content-be-protected-by-copyright>
- Quinn Emanuel, 中美 AI 生成内容可版权性比较: <https://www.quinnemanuel.com/the-firm/publications/noted-with-interest-copyrightability-of-ai-generated-content-in-china-and-the-us/>
- Law.asia, 中国法院对 AI 生成物版权的裁判 (北京互联网法院案) : <https://law.asia/copyright-ai-generated-works-china-judicial-ruling/>
- UK CDPA 1988 第 9(3) 条 (计算机生成作品) : <https://www.legislation.gov.uk/ukpga/1988/48/section/9>

- EU AI Act 第 50 条（AI 生成内容透明度义务）：<https://artificialintelligenceact.eu/article/50/>
- 日本文化厅《AI 与著作権に関する考え方について》（2024, bunka.go.jp, 回源）
- Epoch AI, LLM 推理价格下降趋势：<https://epoch.ai/data-insights/llm-inference-price-trends>

劳动哲学与技术哲学

- Marx, 《1844 年经济学哲学手稿·异化劳动》：<https://www.marxists.org/archive/marx/works/1844/manuscripts/labour.htm>
- Marx, 《资本论》第一卷第十五章, 机器和大工业：<https://www.marxists.org/archive/marx/works/1867-c1/ch15.htm>
- Andy Clark / David Chalmers, 《The Extended Mind》：<https://www.alice.id.tue.nl/references/clark-chalmers-1998.pdf>
- Don Ihde, 《A Phenomenology of Technics》：https://bhaven.org/uploads/3/4/0/3/34038663/a_phenomenology_of_technics_idhe.pdf
- Hannah Arendt, 《The Human Condition》, University of Chicago Press, 1958（劳动/制作/行动三分）
- Hegel, 《精神现象学》, "自我意识的独立与依赖：主人与奴隶"一节
- Harry Braverman, 《Labor and Monopoly Capital》, Monthly Review Press, 1974（去技能化、构想与执行分离）
- Bernard Stiegler, 《Technics and Time》及其"无产阶级化" (proletarianization) 相关论述
- Michael Polanyi, 《The Tacit Dimension》, 1966（默会知识）
- 《庄子·养生主》庖丁解牛、《庄子·天道》轮扁斲轮：<https://ctext.org/zhuangzi/zhs>
- 亚当·斯密, 《国富论》第一篇第一章（论分工）
- 涂尔干, 《社会分工论》（有机团结与分工的差异化）
- 马伯庸, 《长安的荔枝》, 湖南文艺出版社, 2022; 白居易《荔枝图序》
- 马克斯·韦伯, 《中国的宗教：儒教与道教》（家产官僚制）
- 福柯, 《规训与惩罚》（规训与全景敞视）; 尼采, 《道德的谱系》（内疚与债）
- 吴思, 《潜规则：中国历史中的真实游戏》（合法伤害权）

任务替代与劳动市场

- David Autor, 《Why Are There Still So Many Jobs?》：https://economics.mit.edu/sites/default/files/inline-files/Why%20Are%20there%20Still%20So%20Many%20Jobs_0.pdf
- Daron Acemoglu / David Autor, 《Skills, Tasks and Technologies》：<https://economics.mit.edu/sites/default/files/publications/Skills%2C%20Tasks%20and%20Technologies%20-%20Implications%20for%20.pdf>
- OpenAI / OpenResearch / University of Pennsylvania, 《GPTs are GPTs》：<https://openai.com/index/gpts-are-gpts/>
- arXiv, 《GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models》：<https://arxiv.org/abs/2303.10130>
- Science, Noy & Zhang, 生成式 AI 对写作任务生产率的影响：<https://www.science.org/doi/10.1126/science.adh2586>
- NBER, Generative AI at Work：<https://www.nber.org/papers/w31161>
- Microsoft Research, GitHub Copilot 生产率实验：<https://www.microsoft.com/en-us/research/publication/the-impact-of-ai-on-developer-productivity-evidence-from-github-copilot/>
- Microsoft Research, Working with AI: Measuring the Occupational Implications of Generative AI：<https://www.microsoft.com/en-us/research/publication/working-with-ai-measuring-the-occupational-implications-of-generative-ai/>
- Harvard Business School, BCG / Jagged Frontier 实验：<https://www.hbs.edu/faculty/Pages/item.aspx?num=64700>
- Stanford Digital Economy Lab, Canaries in the Coal Mine：<https://digitaleconomy.stanford.edu/publication/canaries-in-the-coal-mine-six-facts-about-the-recent-employment-effects-of-artificial-intelligence/>
- World Economic Forum, 《Future of Jobs Report 2025》：<https://www.weforum.org/publications/the-future-of-jobs-report-2025/>
- World Economic Forum, 2025 就业报告新闻稿：<https://www.weforum.org/press/2025/01/future-of-jobs-report-2025-78-million-new-job-opportunities-by-2030-but-urgent-upskilling-needed-to-prepare-workforces/>
- Goldman Sachs, How Will AI Affect the U.S. Labor Market?: <https://www.goldmansachs.com/insights/articles/how-will-ai-affect-the-us-labor-market>
- Goldman Sachs, How Will AI Affect the Global Workforce?: <https://www.goldmansachs.com/insights/articles/how-will-ai-affect-the-global-workforce>
- IMF, AI Will Transform the Global Economy: <https://www.imf.org/en/blogs/articles/2024/01/14/ai-will-transform-the-global-economy-lets-make-sure-it-benefits-humanity>

- ILO, Generative AI and Jobs: A Refined Global Index of Occupational Exposure: <https://www.ilo.org/publications/generative-ai-and-jobs-refined-global-index-occupational-exposure>
- OECD Employment Outlook 2025: https://www.oecd.org/en/publications/2025/07/oecd-employment-outlook-2025_5345f034.html
- U.S. Census Bureau, Large Firms With at Least 20 Employees Biggest AI Users: <https://www.census.gov/library/stories/2026/05/ai-use-businesses.html>
- Federal Reserve, Monitoring AI Adoption in the U.S. Economy: <https://www.federalreserve.gov/econres/notes/feds-notes/monitoring-ai-adoption-in-the-u-s-economy-20260403.html>
- Federal Reserve, AI Adoption and Firms' Job-Posting Behavior: <https://www.federalreserve.gov/econres/notes/feds-notes/ai-adoption-and-firms-job-posting-behavior-20260327.html>
- Philadelphia Fed, Has Generative AI Adoption Impacted Labor Demand: <https://www.philadelphiafed.org/community-development/workforce-and-economic-development/has-generative-artificial-intelligence-adoption-impacted-labor-demand-at-third-district-firms>
- PwC, 2026 Global AI Jobs Barometer: <https://www.pwc.com/gx/en/services/ai/ai-jobs-barometer.html>
- Anders Humlum / Emilie Vestergaard, «Large Language Models, Small Labor Market Effects», NBER Working Paper 33777 (2025, 丹麦全员数据的零效应估计): <https://www.nber.org/papers/w33777>
- Erik Brynjolfsson / Daniel Rock / Chad Syverson, «The Productivity J-Curve», American Economic Journal: Macroeconomics 13(1) (2021, 通用技术收益滞后于互补资本)

驾驭工程 (harness) 与超级个体

- Mitchell Hashimoto, Harness Engineering 命名文 (2026-02-05, mitchellh.com, 回源)
- OpenAI (Ryan Lopopolo), Codex harness 工程实践 (2026-02, openai.com, 回源)
- Li et al., Agent Harness 综述, TMLR (2026; arXiv 链接见下列研究报告脚注)
- LangChain, Terminal-Bench 自验证钩子实验 (2026-04, 回源)
- Addy Osmani, "验证是瓶颈, 而非生成": <https://addyosmani.com/blog/future-agentic-coding/>
- 《Harness Engineering 深度研究报告》(2026-06-29, 作者提供) ——本篇第九章数据的直接来源, 一手链接见该报告脚注
- Gary Becker, «Human Capital», 1964; Andrew Atkeson / Patrick Kehoe, «Modeling and Measuring Organization Capital», Journal of Political Economy 113(5), 2005 (资本四象限中两个已命名的格)

监管与责任

- EU AI Act Article 14, Human Oversight: <https://artificialintelligenceact.eu/article/14/>
- FDA, Clinical Decision Support Software Guidance: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software>
- NHTSA, Automated Vehicles Safety: <https://www.nhtsa.gov/vehicle-safety/automated-vehicles-safety>
- OECD AI Principles: <https://oecd.ai/en/ai-principles>

数据口径与使用说明：本篇为公开信息的整理与分析，不构成劳动市场预测。劳动市场部分主要使用“任务暴露”与“职业适用度”研究，暴露度不等同于失业率，任务可自动化不等同于职业消失；Stanford 早期职业者研究、WEF 与 Goldman Sachs 的岗位测算均有各自口径与假设，引用时请注意区间；贾湖骨笛的测年与音阶结论以 Nature (1999) 及后续考古研究为准。第二章各法域立场以 2026 年二季度公开资料为准，判例与立法动态更新较快，发表前请复核；第四章《长安的荔枝》情节据马伯庸原著，属文学文本，用作概念说明而非史料。第九章驾驭工程 (harness) 的行业数据转引自 2026 年 6 月《Harness Engineering 深度研究报告》，其中基准测试、市场规模与安全事件均为该报告汇总口径，发表前请沿其脚注逐项回源。正式引用请回溯原始来源。