行 业 深 度 研 究

大模型行业深度报告

模型层：从负毛利到 86% 毛利，价值正在向上游之外重新分配

数据截止日：2026 年 11 月 12 日

报告日期：2026 年 11 月 12 日

主源：Anthropic, PBC Form S-1/A（2026-11-10）及 SEC 原始披露文件

免责声明　本报告为行业研究，不构成任何证券的买入、卖出或持有建议，不设目标价，不预测股价。 报告中的估值分析仅用于说明市场价格隐含的假设。所引数据均标注来源与时点；凡属本报告的判断均在正文标明。 读者应自行核对原始披露文件。本报告不是审计，亦未对任何公司的财务报表发表审计意见。

关于本报告

这不是一份投资建议。 本报告不包含任何买入、卖出或持有评级，不设定目标价，不预测股价。第八章的估值分析用于回答"市场当前用什么价格给模型层定价、这个价格隐含了什么假设"，而不是回答"应该买还是卖"。

本报告的主源是原始文件。 核心财务与经营数据来自 Anthropic, PBC 于 2026 年 11 月 10 日向美国证券交易委员会提交的 S-1/A 修正案（注册号 333-299633，accession 0001628280-26-072484），全文约 1,480 万字符，本报告逐表提取并做自加校验（见附录 A）。行业侧数据来自监管机构原文、公司官方披露与权威机构统计，来源分级见附录 B。

每个数字都能追溯。 正文中出现的每一个 Anthropic 数字，都对应证据台账中的一条记录（编号 A-xx），台账列明数值、口径、时点、出处与来源级别。外部数字（监管原文、同业披露、第三方统计、行情数据）在各表图下方逐条注明来源与时点，并按来源分级登记在附录 B 的登记表中。凡属推导值，均注明推导链条；凡属本报告的判断，均明确标为判断而非事实（正文中以"本报告认为""本报告的判断"标出）。

已知的披露缺口。 Anthropic 的 S-1 未披露收入的地域拆分、未披露分产品收入、未披露 TAM 估算。本报告不填补这些缺口——不用外部估算冒充公司披露，也不用行业总量倒推公司份额。缺口本身是分析结论的一部分。

核心判断

一、模型层已经跨过毛利生死线，但跨过的方式不可外推。 Anthropic 的毛利率从 2024 年的 **−94.8%** 升至 2026 年前三季度的 **85.5%**（附注 8 倒算）。两年内 180 个百分点的摆动，在软件行业没有先例。但这条曲线不是一条平滑的学习曲线：它同时包含了收入结构从早期低价试用转向企业按量付费、推理效率的算法改进、以及云厂商在关联交易框架下给出的算力价格。三者中只有第二项可以外推。把 **85.5%** 当作模型层的稳态毛利率，是本报告认为当前市场最可能犯的错误。

二、收入爆发的速度超过了任何成本曲线的追赶能力。 2026 年第三季度单季收入 170 亿美元，同比增长 1,127%；同期算力与基础设施费用同比增长 382%（A-05、A-46）。收入增速是成本增速的三倍，这是毛利率跳升的直接原因。但同期公司签下的算力采购承诺是 **9,009** 亿美元（A-32），是 2026 年前三季度收入的 27 倍（约 13 年的当前年化收入；即使按 10 月底最宽的 run-rate 口径，仍约 11 年），是期末现金及短投的 6.8 倍。收入曲线与承诺曲线的斜率差，是模型层最大的结构性风险。

三、云厂商是模型层的供应商、渠道和股东，这不是三个关系，是一个关系。 44% 的收入经云厂商市场成交，且公司作为 principal 按总额法确认收入（A-19、A-20）；云厂商同时是可转债与优先股持有人（A-22）；同时是算力承诺的对手方。这意味着模型层披露的毛利率，是在关联交易定价下形成的毛利率。它既不是竞争性市场的价格，也不能直接与独立软件公司的毛利率比较。

四、模型层的护城河不在模型本身，在分发和算力承诺的期限结构。 2026 年 3 月至 9 月发布的模型版本贡献了 Q3 收入的 80% 以上（A-29）——能力领先只能维持约两个季度的收入优势。真正难以复制的是 9,009 亿美元的多年期算力锁定，以及 1.19 万家年付超 10 万美元的企业客户关系（A-26）。前者是资本壁垒，后者是切换成本，两者都不是技术壁垒。

五、按 **S-1** 定价区间，市场给模型层的是"高增长 + 终局高毛利"的双重假设。 74 美元中值对应市值 9,684 亿美元、企业价值 8,666 亿美元（A-40、A-41、A-39），对应滚动十二个月收入的 24.6 倍、Q3 年化收入的 12.7 倍、10 月底 run-rate 的 10.7 倍、2026 年前三季度年化调整后经营利润的 159 倍。这个价格不需要公司继续超预期增长就能成立，但需要 **85%** 的毛利率不回落。 第八章给出这个假设的敏感性边界。

六、媒体流传的 **1.4** 万亿美元估值不被 **S-1** 支持。 按 78 美元上限与 130.87 亿发行后股本计算，市值为 1.02 万亿美元；即使按含全部期权与转股的约 155 亿完全稀释股本计算，也仅为 1.21 万亿美元。要达到 1.4 万亿，股价需达到 107 美元（基本股本口径）或 90.3 美元（完全稀释口径），均高于披露的 70–78 美元区间。这是本报告在核对中发现的、媒体转述与原始文件之间最大的一处偏差。

八个问题与作答位置

本报告的骨架是八个问题。 下表把它们与作答位置一一对应，便于交叉核对，也便于读者看出哪些答案的证据基础较弱。

#	问题	作答位置	一句话答案	置信度
Q1	行业是什么、边界在哪	1.1	主体是模型层的第三段；以"是否自训前沿基础模型"划分与应用层的边界	高
Q2	需求从哪来、在长还是缩	2.2–2.4	企业按量付费是主体；Q3 收入的六到七成来自存量客户扩张，不是新客	中（依赖一项未披露参数）
Q3	供给从哪来、成本与瓶颈	第三章	四种算力取得方式并存，承诺把成本从可变推向固定；瓶颈在容量分配与第三方融资	高
Q4	产业链怎么分工、钱在哪一环	4.2–4.4	毛利留在模型层，但它是在云厂商三重身份下形成的，不是竞争性价格	中高
Q5	谁在竞争、壁垒是什么	5.2–5.5	四个梯队并存；四个护城河候选只有一个半通过	中（无审计级份额）
Q6	处在什么阶段	6.6	能力未见顶、成本近线性下降、渗透在前段、行业尚未整合	中高
Q7	政策与监管怎么管	第七章	三种范式：欧盟义务加罚则、中国事前备案、美国州法碎片化	高
Q8	什么在变	2.3、3.2、6.3、7.2	需求结构、成本结构、监管强度三条线同时在变，正反证据并列	中

怎么读这份报告

来源分级（完整定义与使用规则见附录 **B.1**）。 全文数字按五级标注：**A** 一手原文（法规、S-1、10-Q，可直接作结论依据）· **B** 可复现数据（附查询条件与落盘文件）· **C** 权威二手（尽量回溯原文）· **D** 一般二手（只用于发现线索，不单独支撑结论）· **X** 不可得（区分"做不到"与"不存在"，见 B.5）。推导值不套用来源分级，而是单独标注"推导"并写出链条与复跑脚本。

三条阅读路径。

- 十分钟：前言六条核心判断 → 8.3 三重约束 → 8.5 可比公司。
- 关心竞争与定价：第五章（竞争与护城河）→ 8.3、8.5、8.6（OpenAI 对照）。
- 复核数字：附录 A（提取与自加校验）→ 附录 B（来源登记与台账）→ `scripts/`（全部计算可复跑）。

本报告不做什么。 不做买卖评级、不设目标价、不预测股价；不用外部估算填补公司未披露的口径；不倒推公司市场份额；不把不同口径的市场规模相加。

目　录

第一章 行业定义与商业模式 6

1.1 边界：什么算"模型层" 6

1.2 收入模式：按 token 计费的"消费型"收入 6

1.3 成本结构：算力是唯一的真变量 7

1.4 三个反常，以及它们各自意味着什么 9

1.5 本章结论 11

第二章 需求：谁在付钱，为什么付 12

2.1 需求的三个层次与一个计量陷阱 12

2.2 企业需求：可量化的部分 12

2.3 需求的真实性检验：三个负对照 14

2.4 价格弹性与单位任务成本 14

2.5 需求的波动：季节性与项目制 15

2.6 本章结论 15

第三章 供给：算力、承诺与期限结构 16

3.1 算力的四种取得方式，与它们各自的代价 16

3.2 承诺的期限结构：本报告认为最重要的单一数字 16

3.3 收入增速与算力增速的剪刀差 18

3.4 芯片多元化：为什么它不是纯粹的技术选择 18

3.5 人才：5,300 人与 666 万美元人均收入 18

3.6 供给约束的三种表现形式 19

3.7 本章结论 19

第四章 价值链与价值分配 20

4.1 链条结构：谁向谁付钱 20

4.2 云厂商的三重身份：供应商、渠道、股东 20

4.3 价值分配：收入流向了哪里 21

4.4 一个反事实检验：如果算力按市价计价 22

4.5 与历史技术周期的结构对比 23

4.6 本章结论 23

第五章 竞争格局 24

5.1 竞争发生在三个维度，而不是一个 24

5.2 玩家图谱：四个梯队 24

5.3 开源：模型层的价格天花板 26

5.4 竞争的结果：三个可观察变量 26

5.5 护城河检验：四个候选，一个半通过 26

5.6 本章结论 27

第六章 行业阶段与技术趋势 29

6.1 判断阶段需要三个坐标系，而不是一个 29

6.2 能力曲线：从"更大"到"更会用" 29

6.3 成本曲线：通缩是结构性的 29

6.4 渗透曲线：从"工具"到"协作者" 31
6.5 技术趋势的四条主线 31
6.6 阶段判断及其对估值的含义 31
6.7 本章结论 32

第七章 政策与监管 33
7.1 三种监管范式：不是"松紧"之分 33
7.2 欧盟：GPAI 义务已生效，高风险条款被推迟两年 33
7.3 中国：备案制的实际形态 35
7.4 美国：没有联邦专法，约束来自三层 36
7.5 其他主要辖区：一句话状态 37
7.6 监管如何进入模型层的财务模型 37
7.7 中国模型层：结构、定价与资本化 38
7.8 本章结论 40
7.9 本章的局限 41

第八章 估值 42
8.1 为什么模型层难估值：三个结构性障碍 42
8.2 定价区间隐含的倍数 42
8.3 三重约束交叉检验 43
8.4 敏感性：什么会推翻结论 47
8.5 上市可比公司 48
8.6 OpenAI 的隐含区间 53
8.7 估值的边界：本报告不做什么 56

第九章 风险与局限 57
9.1 行业层面的五个风险 57
9.2 公司层面的四个风险 57
9.3 本报告的局限：未定、未测、已知缺陷 58
9.4 什么会让我们改变判断 59
9.5 本章结论 59

附录 60

附录 A 数据提取与校验方法 60

附录 B 来源分级与证据台账索引 62

附录 C 图表索引 65

第一章 行业定义与商业模式

1.1 边界：什么算"模型层"

一个行业报告最先要回答的问题不是"多大"，而是"算谁的"。大模型产业链可以切成四段：

层级	做什么	收入形态	代表主体
芯片与硬件	提供训练与推理算力	卖设备、卖系统	NVIDIA、AMD、Broadcom、台积电
云与基础设施	提供算力租用、机房、网络	按小时/按卡租用	AWS、Google Cloud、Azure、Oracle
模型层	训练基础模型，以 **API**、应用、订阅出售智能	按 **token** 计量、订阅、最低承诺	**Anthropic**、**OpenAI**、**Google DeepMind**、**xAI**、**DeepSeek**、阿里通义、字节豆包
应用层	在模型之上构建垂直产品	SaaS 订阅、按次收费	Cursor、Harvey、Perplexity、各类 Copilot

本报告的主体是第三层。这个边界有三处必须说清楚的模糊地带：

第一，云厂商同时是模型层玩家。 Google 既卖 TPU 和云，也做 Gemini；Amazon 既卖 Trainium 和云，也投了 Anthropic 并持有其可转债；Microsoft 既卖 Azure，也独家分发部分 OpenAI 模型。把"云"和"模型"当成两个独立行业分析，会系统性低估模型层的真实成本，也会系统性高估它的独立性。 本报告在第四章专门处理这种三重身份。

第二，模型层与应用层的边界正在消失。 Anthropic 的第一方应用（Claude Code 等）已经产生实质收入；OpenAI 的 ChatGPT 既是模型入口也是应用。反过来，Cursor、Harvey 这类应用公司也在自训小模型。本报告的处理方式是：以"是否自训前沿基础模型"为判据，自训的归模型层，不训的归应用层，无论其产品形态。

第三，开源模型不属于"公司"但属于"行业"。 Llama、Qwen、DeepSeek 的开源权重构成了模型层的价格天花板。它们没有收入，但它们的边际成本为零这一事实，直接约束了闭源模型层的定价权。第五章把开源作为竞争约束处理，而不是作为市场参与者处理。

地理范围：以美国与中国的模型层为主，其余辖区只在影响成本时出现。 这不是取舍，而是披露分布的结果——全球模型层唯一同时披露完整财务报表与算力承诺的一手文件，是一家美国公司的 S-1（Anthropic，2026-11-10）；中国的模型层有独立的价格体系与监管路径（第七章 7.3 与 7.7），因此不与美国公司并列做倍数比较。第八章的可比公司全部为美股上市公司，原因同此。

> 口径提示：本报告所说的"模型层收入"，指模型公司自身确认的收入。由于云厂商市场（CSP marketplace）按总额法确认（见 1.3），这部分收入里包含了支付给云厂商的通道费用。它不等于模型层从终端客户收取的净额，也不等于模型层留下的价值。

1.2 收入模式：按 token 计费的"消费型"收入

Anthropic 在 S-1 中对自身收入模式的表述是明确的：**pay-for-intelligence**（为智能付费）是主要模式，按 **token** 计量、按模型定价，占收入**"substantial majority"**（A-23）。此外有订阅收入（Claude Pro / Max / Ultra 及企业版），按订阅期直线确认，期限一般不超过 12 个月。

这个模式有三个可观察的后果：

其一，收入等于用量乘价格，而两者都在快速变化。 用量随客户渗透与单客户用量增长；价格随模型效率提升而下降。Anthropic 披露，Opus 5.5 相对 Opus 5 将客户的单位任务成本降低了约 **40%**（A-45），而公司称这一降价"在保持毛利的同时"完成。这意味着模型层的收入增长必须靠用量增长跑赢单价下降——这是一个典型的"卖得越多、单价越低"的行业，与半导体代工的价格曲线同构，而不是与 **SaaS** 的订阅曲线同构。

其二，收入确认的时点由客户行为决定，不由合同决定。 收入在客户实际消耗 token 时确认，"growing number of enterprise customers agree to annual or multi-year contracts with minimum spend commitments"（A-23）。最低承诺带来的是合约保障，不是收入确认。这解释了为什么公司的剩余履约义务（RPO）只有 138 亿美元（A-33），而 2026 年前三季度收入已达 332 亿美元——**RPO** 与收入的比例只有 **0.4** 倍，远低于典型企业软件公司。 这不是会计问题，是商业模式问题：模型层的收入没有多年期合同的"蓄水池"。

其三，渠道结构决定了收入的含金量。 见下节。

表 1-1 收入模式的三个层次

层次	计价方式	确认时点	可预测性	2026 年前三季度地位
按量付费（API / Platform）	按 token，按模型定价	客户消耗时	低，随用量波动	主要模式，占"substantial majority"
订阅（Pro / Max / Ultra / 企业版）	固定期间费用	订阅期内直线	中，有预付	次要，但提供现金流前置
最低承诺合同	承诺额 + 超出部分按量	实际消耗时	中高，有合约保障	增长中，不改变确认时点

数据来源：S-1/A 商业模式章节与附注 2（A-23）。

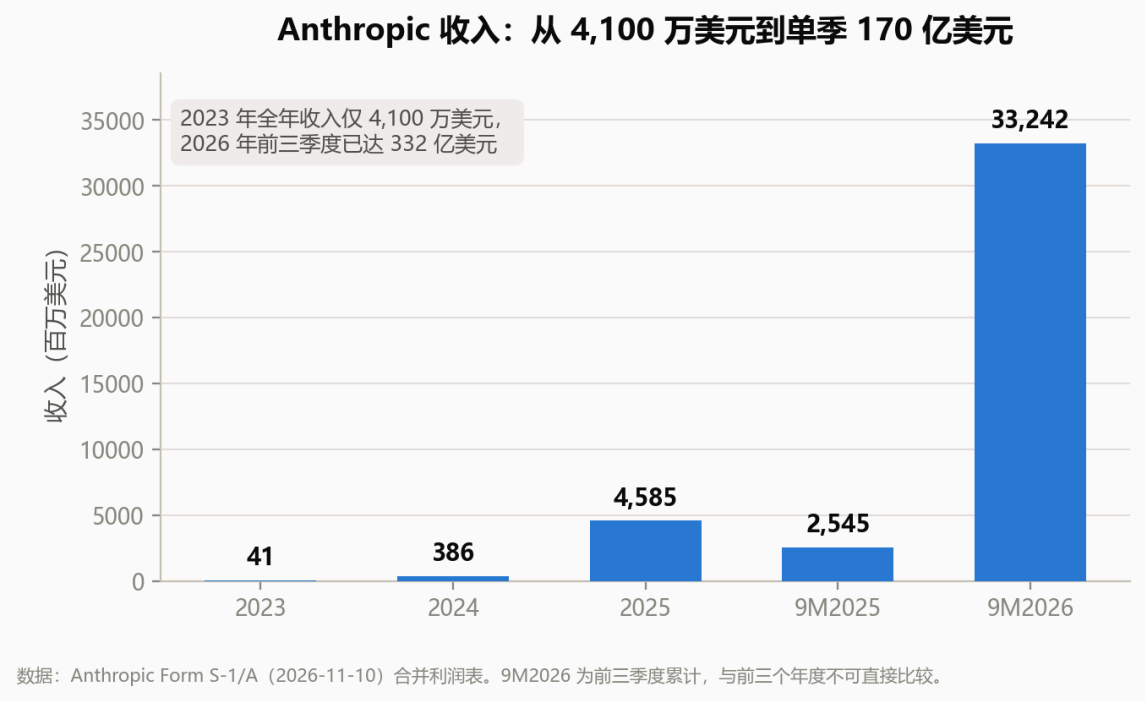


图 1-1

图 **1-1** 的读法：五个柱里只有前三个是完整年度，后两个是前三季度累计。 把 9M2026 的 332 亿与 2025 全年的 45.85 亿直接相除，得到 7.2 倍；按同口径（9M2026 vs 9M2025）相除，是 **13.1** 倍。两个数都不假，但含义不同——报告里凡出现"增长 **N** 倍"，都必须先看分母是不是同一个期间。

1.3 成本结构：算力是唯一的真变量

按性质列示，Anthropic 2026 年前三季度的营业费用为 345.02 亿美元，其中算力与基础设施 188.08 亿（54.5%）、人员 48.63 亿（14.1%）、销售营销与伙伴 55.29 亿（16.0%）、其他管理 53.02 亿（15.4%）（A-07）。

按功能列示则是另一幅图景：收入成本 48.12 亿、研发 192.40 亿、销售与营销 66.19 亿、管理费用 38.31 亿（A-15）。两种列示的合计相同（**345.02** 亿），但每一项的含义完全不同。 这个差别是本报告认为读懂模型层财务报表的关键：

- 功能口径的"研发"（**192.40** 亿）几乎就是性质口径的"算力与基础设施"（**188.08** 亿）。 差额 4.32 亿美元，大致是研发人员成本。也就是说，模型层 **97.8%** 的研发支出是算力（188.08 ÷ 192.40）。

- 功能口径的"收入成本"（**48.12** 亿）是服务付费客户所耗的算力与人员。 公司在附注中明确："substantially all of cost of revenue consists of compute and infrastructure expenses associated with customer-serving workloads"（A-16）。
- 免费用户的算力不在收入成本里，在销售费用里。 附注 8 脚注 (3) 明确：销售与营销费用包含"compute and infrastructure costs associated with providing free access to the Company's intelligence service"（A-17）。

> 这意味着：模型层披露的毛利率，只覆盖付费流量的服务成本，不覆盖免费流量和模型训练的成本。 一家把 90% 算力用于训练、10% 用于服务付费客户的公司，可以同时拥有 85% 的毛利率和巨额经营亏损。这两个数字不矛盾，它们描述的是不同的东西。任何用毛利率推断"模型层很赚钱"的论证，都漏掉了研发这一项——而研发是收入的 57.9%。

表 1-2 两种成本口径的映射（2026 年前三季度，百万美元）

按性质	金额	占收入	按功能	金额	占收入
算力与基础设施	18,808	56.6%	收入成本	4,812	14.5%
人员	4,863	14.6%	研发	19,240	57.9%
销售、营销与伙伴	5,529	16.6%	销售与营销	6,619	19.9%
其他管理	5,302	15.9%	管理费用	3,831	11.5%
Bartz 和解	—	—	Bartz 和解	—	—
合计	**34,502**	**103.8%**	合计	**34,502**	**103.8%**

数据来源：S-1/A 合并利润表与附注 8（A-07、A-15）。两种口径为同一总额的不同切分，不可相加。

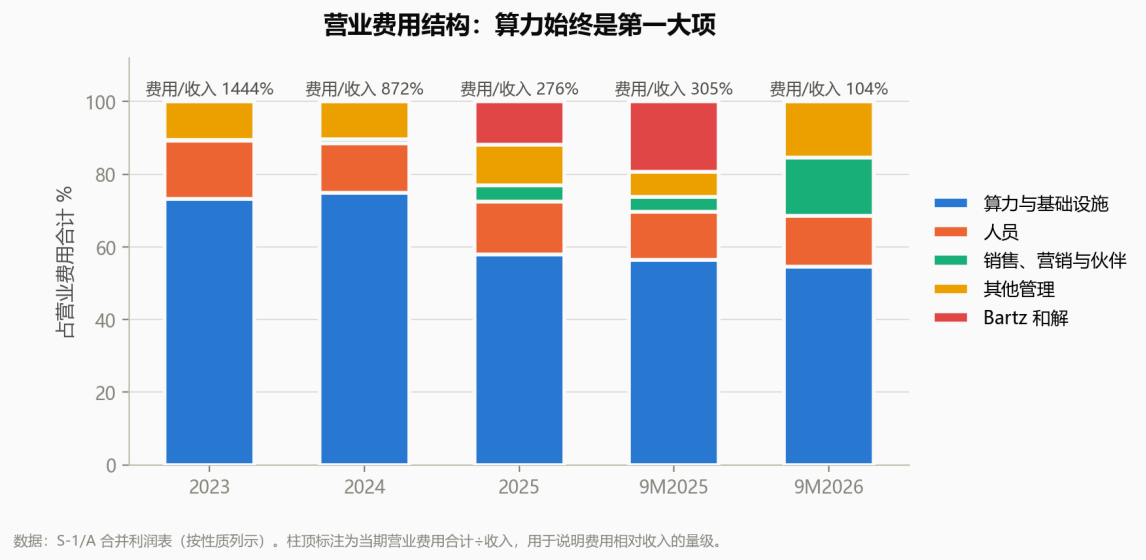


图 1-2

图 **1-2** 的读法：柱内是各项费用占营业费用合计的比重（归一化），柱顶才是费用合计÷收入。 两个问题必须分开看，否则会看错。

- 结构确实变了： 算力与基础设施占费用总额的比重，从 2023–2024 年的 **73%–75%** 降到 2025 年起的 **54%–58%**。但这不是算力投入减少——算力费用从 2024 年的 25.2 亿美元升到 9M2026 的 188.1 亿美元（A-07），是人员与销售、伙伴费用的增速更快，把它在费用盘子里挤薄了。
- 量级的变化远大于结构： 费用合计÷收入从 2023 年的 **1,444%**、2024 年的 **872%** 降到 9M2026 的 **103.8%**。

> 收入增长把费率摊薄了近 **14** 倍，而费用结构只动了约 **20** 个百分点。 如果只看结构图，会得出"算力重要性在下降"的结论；实际发生的是收入的爆炸式增长盖住了所有结构变化。 这也是本章反复用"占收入比"而不

> 是"占费用比"的原因。

算力成本的绝对量是理解这门生意的锚。 2026 年前三季度算力与基础设施费用 188.08 亿美元，年化约 250 亿美元。同期公司签下的算力采购承诺总额 9,009 亿美元，其中 2030 年底前预计发生 4,516 亿（A-32）。把承诺额按 **2030** 年前平均摊开，每年约 **1,000** 亿美元量级——是当前年化算力费用的四倍。 这个缺口只有两种填法：收入继续高速增长，或者承诺被重新谈判。

1.4 三个反常，以及它们各自意味着什么

反常一：小规模时毛利率为负，大规模时毛利率 86%

期间	收入	收入成本	毛利率
2023	41	26	36.6%
2024	386	752	**−94.8%**
2025	4,585	2,773	39.5%
2025 前三季度	2,545	1,796	29.4%
2026 前三季度	33,242	4,812	**85.5%**

数据来源：由 S-1/A 合并利润表与附注 8 倒算（A-18）。

2024 年，Anthropic 每收 1 美元收入，要花 1.95 美元的收入成本。这不是会计异常，是行业早期的真实状态：推理的固定成本（容量预留、冗余、冷启动）在低利用率下无法被摊薄，而早期为了争取客户，定价必然低于成本。

2026 年前三季度的 85.5% 则来自三个同时发生的变化：收入规模放大 13 倍导致固定成本被摊薄；推理效率改善（公司称单位任务成本下降 40%）；以及高毛利的企业按量付费占比上升。

关键问题是：这三个因素里，哪几个还能继续？ 摊薄效应会随基数变大而衰减；效率改进有物理下限；结构升级只能发生一次。因此本报告的基本判断是：**85.5%** 更接近这一轮周期的顶部区域，而不是新的起点。

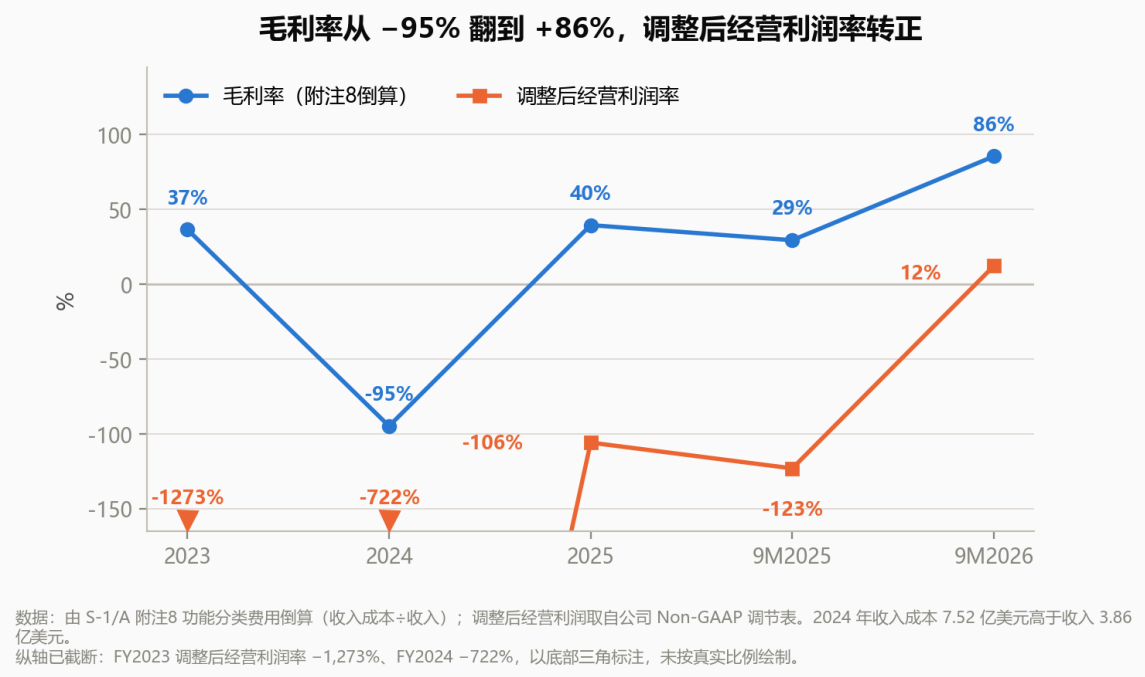


图 1−3

图 **1-3** 的读法：两条线的口径完全不同，不能相互印证。 毛利率是附注 **8** 倒算的会计口径（收入成本÷收入），只覆盖服务付费客户的成本；调整后经营利润率是公司 **Non-GAAP** 口径，在营业亏损基础上加回股权激励与非现金捐

赠。两者之间隔着研发、销售、管理与调整项四层。 图上毛利率在 2025 年就回到 39.5%，而调整后经营利润率仍是 −105.9%——这个 **145** 个百分点的缺口，就是"毛利率高不代表赚钱"的量化版本。

注意纵轴已截断： FY2023 调整后经营利润率为 −1,273%、FY2024 为 −722%，以底部三角标注，未按真实比例绘制。若不截断，整张图会被这两个点压扁到无法阅读——这本身也是规模效应的量级证据。

表 1-3 外部对照：同业披露口径下的毛利率

主体	期间	口径	毛利率	来源级别
OpenAI	FY2025	1 − 60.47 ÷ 130.7（付微软的 cost-of-revenue 占收入比重）→ 毛利率上限	**≤53.7%**	泄露文件，FT 核实，未审计
xAI （ SpaceX "SI" 分部）	2025 Q3	(813 − 分部 COGS) ÷ 813（百万美元）	**29.8%**	SEC 10-Q（一手）
xAI（同上）	2026 Q3	(7,560 − 2,675) ÷ 7,560（百万美元）	**64.6%**	SEC 10-Q（一手）
xAI（同上）	2026 前三季度	收入 10,939、分部口径	**61.3%**	SEC 10-Q（一手）
Anthropic	2026 前三季度	收入成本 ÷ 收入	**85.5%**	S-1/A（A-18）

数据来源：SpaceX 10-Q（CIK 0001181412，2026-09-30 期）；FT 核实的 OpenAI FY2025 文件；S-1/A。百分数由本报告计算。

这张表支持三件事，但都不能推得太远：

第一，规模摊薄这个机制是真的，而且不是 **Anthropic** 独有。 xAI 的分部毛利率在一年内从 29.8% 升到 64.6%，涨了 34.8 个百分点。这与本报告对 **Anthropic**（**29.4% → 85.5%**）的机制解释一致——早期固定成本无法摊薄，规模上来后迅速改善。因此 **85.5%** 本身不是造假，它是一个真实的规模效应读数。

第二，**85.5%** 仍显著高于可见的同业。 比 xAI 分部高 21 个百分点，比 OpenAI 的毛利率上限高 32 个百分点。

第三，但差距的精确值不可得。 xAI 的 "SI" 分部是混合口径：既含 Grok 的订阅与 API，也含对外出租算力（低毛利），还含 X 的广告收入（高毛利）。三个方向的影响互相抵消，无法分离出"纯模型业务"的毛利率。 OpenAI 那一行是上限而非实际值（60.47 亿只是成本中的一项）。

> 本报告的判断：外部数据支持"**Anthropic** 的 **85.5%** 高于同业"这个方向，但不支持量化差距。 残余差额是否来自关联交易定价（云厂商同时是股东、供应商与渠道），用公开信息无法判定——这需要算力采购的折扣率，而该数字未披露（见第九章"未定"清单）。
>
> 可检验的推论是：若关联交易定价是主因，那么当公司按 **S-1** 所述扩大非股东供应商的采购时，毛利率应向下收敛。 这是第九章列出的观察指标之一。

第八章的估值敏感性分析会把毛利率作为第一变量。

反常二：人均收入 666 万美元

按滚动十二个月收入 352.82 亿美元、期末员工 5,300 人计算，人均收入 665.7 万美元（A-31，本报告复算为 665.7 万，与公司披露的"over $6.6 million"吻合）。作为对照，同期 Alphabet、Microsoft 的人均收入在 200–300 万美元量级。

这个数字不是效率的证明，而是收入结构与员工结构错配的结果：收入由客户的 token 消耗驱动，而消耗由客户的工程师写代码产生，不需要模型公司配备同等人力；模型公司的 5,300 人里，大部分是研究员与工程人员，不是销售与实施人员。

它同时是一个脆弱性指标。 人均收入越高，意味着每新增一名员工对应的收入杠杆越大，也意味着一旦收入增速放缓，人员成本的刚性会迅速侵蚀利润。2026 年前三季度人员费用 48.63 亿，占收入 14.6%，这个比例在收入增速降到 50% 以下时会翻倍。

反常三：股东权益为负 1,498.73 亿美元

Anthropic 2026 年 9 月 30 日的股东权益为 **−1,498.73** 亿美元，累计亏损 1,561.93 亿美元（A-38）。这个数字的主体不是经营亏损，而是可转债公允价值变动：2026 年前三季度确认 −943.24 亿美元（A-10），累计起来构成了累计亏损的绝大部分。

可转债的账面金额是 1,305.02 亿美元，而其本金只有 62 亿美元（A-35）。差额来自 Amazon 可转债按公允价值计量——企业估值每上涨一轮，这笔负债的账面值就上涨一轮，亏损就扩大一轮。

> 这是模型层财务报表最容易被误读的地方。 一家公司可以在报告 1,040 亿美元净亏损的同时，经营性现金流为正 65.69 亿美元（A-49）。这两件事同时为真，因为前者主要是估值上升的会计后果，后者是真实的现金进出。用净亏损判断模型层的经营状况，会得到完全相反的结论。

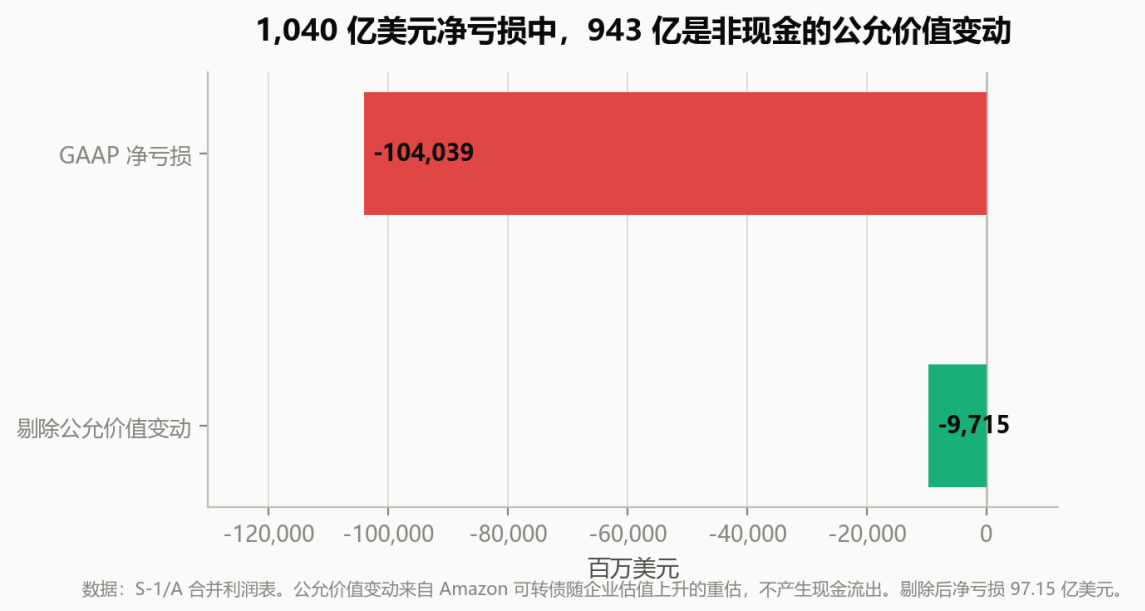


图 1−4

图 **1-4** 的读法：剔除公允价值变动后，**9M2026** 的净亏损只剩 **97.15** 亿美元。 这个数字仍不等于"经营亏损"——它还包含利息收入、税项等营业外项目，与营业亏损（−12.60 亿美元）不能互相替代（见表 1-2 的口径说明）。但它足以说明：把 **1,040** 亿当作"亏了这么多钱"是错的，正确的读法是"其中 **943** 亿来自估值上升本身"。

这个机制有一个容易被忽略的对称性：如果企业估值下降，同一科目会反向变成巨额利润。 也就是说，这家公司的报表利润在一个相当大的程度上与企业估值同向变动——估值越高，账面亏得越多。这不是财务技巧，是可转债按公允价值计量的必然结果；但它意味着任何用"净利润"给模型层定价的方法，在方向上是反的。

1.5 本章结论

1. 模型层的边界必须以"是否自训前沿基础模型"划定，否则会把云厂商和应用的收入混进来。
2. 收入是按 token 计量的消费型收入，没有多年合同蓄水池，RPO/收入仅 0.4 倍。
3. 成本结构上，97.8% 的研发支出是算力；算力同时出现在收入成本（服务付费客户）、研发（训练）、销售费用（免费流量）三处。
4. 毛利率只覆盖付费流量的服务成本，不覆盖训练与免费流量，因此高毛利与巨额亏损可以并存，且这不是矛盾。
5. 85.5% 的毛利率是三个一次性因素叠加的结果，本报告不认为它可以线性外推。

第二章 需求：谁在付钱，为什么付

2.1 需求的三个层次与一个计量陷阱

模型层的需求可以拆成三层，三层的驱动因素、价格敏感度和可预测性完全不同：

层次	谁在买	计价	驱动因素	可预测性
开发者 API	应用公司、企业内部工程团队	按 token	应用用量、模型能力、单价	低，随下游产品生死波动
第一方应用	企业知识工作者、专业人士	按席位或按量	工作流渗透率、替代人工的成本	中
消费者订阅	个人用户	月费	习惯、能力感知、免费替代品	中低，易流失

计量陷阱在于：三层需求的收入确认口径不同，但报表只给一个总数。 按量付费在消耗时确认，订阅在订阅期内直线确认。当一家公司从订阅转向按量、或从按量转向订阅时，收入增速会因口径切换而失真，而不是因需求变化。**Anthropic** 的 **S-1** 未披露三层各自的收入占比（A-24 同类缺口），只说明按量付费占"substantial majority"。本报告不推测具体比例。

2.2 企业需求：可量化的部分

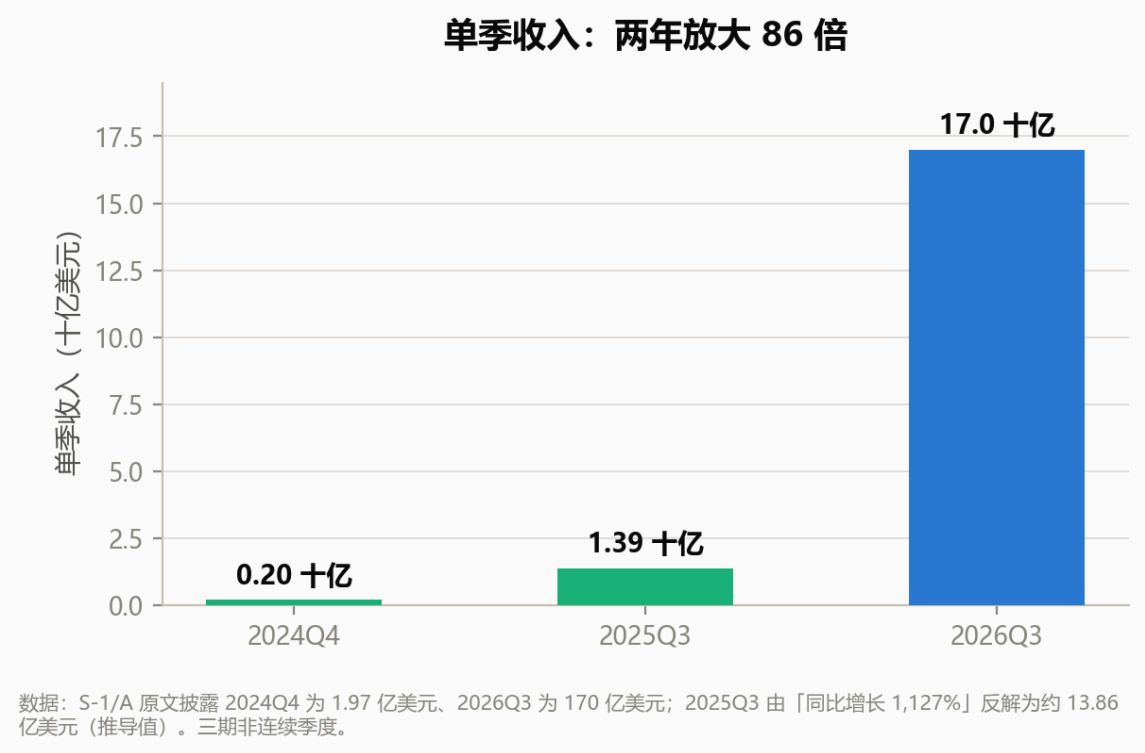


图 2-1

Anthropic 披露的客户指标提供了一个罕见的、可交叉验证的企业需求切面（A-26、A-25、A-27、A-28）：

表 2-1 企业需求的关键指标（截至 2026-09-30）

指标	数值	口径
年付 ≥10 万美元的客户数	11,900 家	滚动十二个月收入
年付 ≥100 万美元的客户数	1,900 家	同上
年付 ≥1,000 万美元的客户数	250 家	同上

指标	数值	口径
净收入留存率（NRR）	约 900%	剔除 Pro/Max/Ultra 订阅用户；基期为一年前同三个月
最大单一客户占收入	<10%	2026 年前三季度
前五大客户合计占收入	约 25%	2026 年前三季度
前五大以外客户占收入	79%（Q3 2026）、64%（Q1 2026）	单季

数据来源：S-1/A 客户指标章节与风险因素（A-25 至 A-28）。

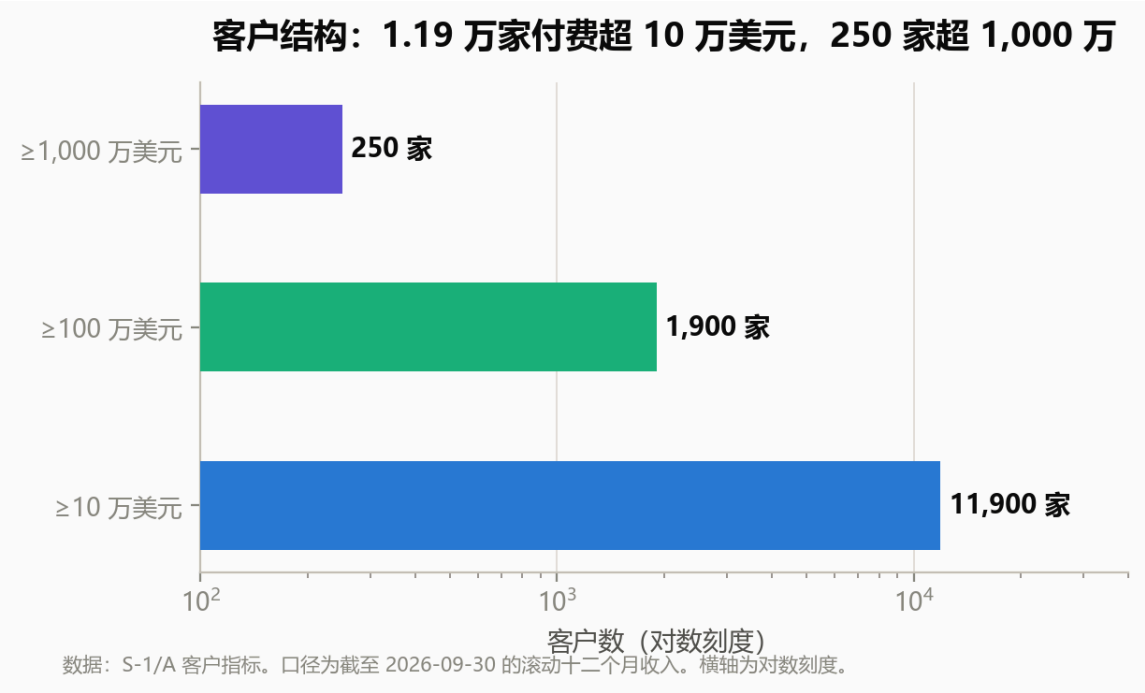


图 2-2

这组数字里最重要的是 **900%** 的 **NRR**，以及它同时伴随的集中度下降。 通常这两件事是矛盾的：高 NRR 往往来自少数大客户加购，会推高集中度。Anthropic 的情况相反——前五大客户占比仅 25%，且前五大以外的收入占比在两个季度内从 64% 升到 79%。这说明增长是"广谱"的，不是"大客户驱动"的。

> 推导（含假设，请谨慎使用）：设 Q3 2025 总收入为 R，Q3 2026 为 12.27R（由 +1,127% 反解，A-05）。NRR 900% 意味着基期已存在客户在 Q3 2026 贡献了 9 倍于其基期收入的金额。若忽略消费者订阅在基期的占比（该占比未披露，是本次推导唯一的自由参数），则新客户贡献 12.27R − 9R = 3.27R，约占 Q3 2026 收入的 **27%**；若消费者订阅占基期收入 20%，则新客户占 **41%**。 结论的方向是稳健的：**Q3 2026** 收入的六到七成来自一年前就已存在的客户的用量扩张，而不是新客拉新。 这个结论对"增长是否可持续"的含义是双面的——存量客户的扩张能力已被验证，但新客获取的贡献比表面增速看起来的要小。

客户结构的另一面是客单价，而且可以用两个门槛做一次下界估算。 1,900 家客户年付超过 100 万美元，250 家超过 1,000 万美元。以 2026 年前三季度 332.42 亿美元收入为分母：前五大客户占 25%，即 83.1 亿美元；其余 245 家若每家恰好付 1,000 万美元，按九个月均匀折算约合 750 万美元，合计 18.4 亿美元。两项相加至少 **101.5** 亿美元，占前三季度收入的 **30.5%**。 反过来说，其余约 1.17 万家客户至多贡献约七成收入。客户结构是"250 家撑起三成、长尾承担七成"——既不是"少数超大 + 大量微型"，也不是头部通吃，这个下界与"前五大 25%"同时成立。

交叉销售是可观察的第二个证据。 Q3 2026，年付 ≥10 万美元的客户中有超过 50% 同时使用了 Claude Platform 和至少一个第一方应用，而 2025 年 Q4 这个比例约为 30%（A-48）。半年内从 **30%** 到 **50%**，说明企业客户正在从"调用 **API** 做自己的产品"扩展到"直接采购现成应用"，需求在同一客户内部横向扩散。 这是比客户数增长更可靠的留存信号。

更新的读数：**Run-Rate Revenue 810** 亿美元（**2026-10-31**）。 表 2-1 的客户指标截至 9 月底，S-1/A 另披露了一个更近的数字：截至 10 月 31 日的 Run-Rate Revenue 约 **810** 亿美元，较 9 月 30 日的 710 亿美元增加 100 亿美元（A-55；710 亿为推导）。口径是"过去 28 天按量付费 × 13 + 当月订阅费 × 12"（A-54），公司同时明确警告：它"不能替代 GAAP 收入，也不是收入预测"（A-56）。这个指标在本报告中有三个用途，也有三条边界。

- 用途一：观测频率。 客户指标只到 9 月底，run-rate 把观察点推进一个月。**10** 月单月环比增加约 **14%**，说明第三季度的加速在季末之后没有停住。
- 用途二：**Q4** 收入的平价估计。 按定义，13 个 28 天期 ÷ 4 个季度 = 3.25 期/季；若 10 月的活动水平持平，run-rate ÷ 4 正好是一个季度的收入——约 **202.5** 亿美元，比**"Q4** 与 **Q3** 持平**"**的 **170** 亿高 **19%**。 这是第八章 DCF 基期敏感性的来源。
- 用途三：一个可比的对照口径。 它是唯一能与未上市公司直接对照的当期口径（8.6 用它与 OpenAI 的年末 run-rate 预期并列）。

三条边界： 其一，它不是 GAAP 收入，不进入任何 GAAP 倍数的主口径（8.2 的主口径仍是 Q3 年化）；其二，季节性会放大偏差——公司原文明确提示，在用量偏高的季节测量会系统性抬高年化值（A-56）；其三，它包含"尚未被限制或终止的违规使用客户"所产生的费用，公司自己承认这部分不会持续（A-56 同节风险因素）。因此本报告使用它的方式是"并列一个更宽的口径并标出边界"，而不是"用它替换 **GAAP** 收入"。

2.3 需求的真实性检验：三个负对照

需求数据最容易造假，也最容易自欺。本报告用三个可执行的检验来判断"增长是不是真的"：

检验一：增长是否伴随价格下降？ 如果收入增长完全来自涨价，那增长的可持续性取决于定价权；如果伴随单价下降，则增长来自用量，更可持续。Anthropic 披露 Opus 5.5 使单位任务成本下降约 40%（A-45），同时收入增长 12 倍。结论：增长来自用量，不是涨价。 这是正面信号。

检验二：增长是否集中在单一时点/单一事件？ 若增长来自某一次模型发布的脉冲，则不可持续。S-1 披露，2026 年 3 月至 9 月发布的模型版本贡献了 Q3 2026 收入的 80% 以上（A-29）。这条初看是风险，细看是行业特征：它说明收入半衰期约为两个季度——旧模型版本在半年内就失去大部分商业价值。这意味着"持续发布"不是竞争优势，而是入场费。 一旦发布节奏中断，收入会在两个季度内反映出来。

检验三：增长是否伴随应收账款的异常膨胀？ 收入可以通过放宽账期来"制造"。2026 年前三季度应收账款增加 70.76 亿美元（A-49），同期收入 332.42 亿美元，应收账款增量占收入 **21.3%**。这个比例偏高，但需要结合渠道结构看：CSP marketplace 贡献 44% 的收入，而云厂商的回款周期通常在 30–60 天（A-20）。应收账款占比高更像渠道结构的后果，而非信用放松的证据——但这需要后续季度验证，本报告将其列为待观察项。

2.4 价格弹性与单位任务成本

模型层的需求曲线有一个反直觉的性质：降价会扩大需求，但扩大得不够快时，收入会下降。

这可以用一个简单的恒等式表达：

> 收入 = 任务量 × 每任务 token 数 × 单价

模型效率提升会同时压低"每任务 token 数"和"单价"。若两者的乘积下降 40%（如 Opus 5.5 相对 Opus 5 的单位任务成本，A-45），则任务量必须增长 **67%** 才能保持收入不变。Anthropic 在 2026 年前三季度实现了 12 倍收入增长，说明任务量的增长远超这个门槛。

但这个门槛会随时间上升。 当单位成本再降 40% 时，同样的收入需要任务量再增 67%；而任务量的增长受限于真实工作流的渗透速度——一个人一天能审的合同份数、能写的代码行数是有限的。因此，模型层收入的长期增速，最终受限于"人类工作中有多少比例可以交给模型"，而不是受限于模型能力。

> 这是本报告对**"Scaling law** 决定收入**"**这一流行说法的修正：算力决定能力上限，能力决定可替代的任务范围，但收入取决于任务范围乘以渗透率。前两个环节有清晰的指数曲线，第三个环节是 **S** 形曲线。 用前两个环节的斜率外推收入，会系统性高估。

2.5 需求的波动：季节性与项目制

S-1 明确披露了需求的季节性："holiday and vacation periods in the summer and winter, when certain workflows slow down and overall customer activity declines"，并举例说许多公司在假期安排 code freeze，导致开发者提交与功能发布放缓（A-29 同节）。

这个披露的重要性被低估了。 它意味着模型层的收入具有与企业 **IT** 预算周期同步的性质，而不是与消费互联网同步。对估值的含义是：不应该用消费互联网的**"**高频、低波动**"**倍数，而应该用企业软件的**"**季度波动、预算周期**"**倍数。 第八章的可比公司选择会体现这一点。

另一重波动来自项目制。S-1 提到销售周期因客户而异，企业销售通常长于个人（A-20 同节）。项目制的含义是：收入确认可能滞后于合同签订，而 **RPO** 只有 **138** 亿美元（**A-33**），说明大部分合同不含长期最低承诺。这使模型层的收入更像**"usage-based SaaS"**而非**"subscription SaaS"**。

2.6 本章结论

1. 企业需求已被三项独立指标验证：1.19 万家年付超 10 万美元客户、900% 的 NRR、前五大集中度仅 25%。三者同时成立，排除了"少数大客户撑起增长"的解释。
2. 增长来自用量而非涨价，这是可持续性的正面证据；但收入半衰期只有约两个季度，持续发布模型是入场费而非优势。
3. 单位任务成本下降 40% 意味着任务量必须增长 67% 才能保持收入，价格弹性是模型层最需要跟踪的单一变量。
4. 需求呈企业 IT 预算周期，而非消费互联网周期；RPO/收入仅 0.4 倍，收入的可预测性低于典型企业软件。
5. 应收账款增量占收入 21.3%，偏高但可由渠道结构解释，列为待观察项。

第三章 供给：算力、承诺与期限结构

3.1 算力的四种取得方式，与它们各自的代价

模型层不拥有芯片，也不拥有数据中心。它通过四种方式取得算力，每种方式在报表上的落点不同：

方式	会计落点	现金节奏	风险特征
向云厂商按需租用	算力与基础设施费用	随用随付	价格随行就市，容量无保障
多年期承诺采购（take-or-pay）	承诺披露 + 使用时计入费用	按合同节点付	用不满也要付，形成刚性负债
自建/包租数据中心	固定资产 + 折旧	前置大额资本开支	建设期风险、利用率风险
设备租赁（厂商/金融方）	租赁负债	分期	供应链与融资方风险

Anthropic 的现状是四种方式并存，且重心正在从第一种转向第二、三、四种。S-1 明确描述了这一转变：

> "Historically, we have generally obtained compute from a limited number of CSPs under arrangements in which our compute and infrastructure expenses are paid as compute becomes available to us. In the future, we expect we will increasingly enter into separate arrangements with a greater number of different counterparties, each responsible for discrete services or inputs across the compute supply chain."

这段话的含义是：从"用了才付"转向"承诺了就要付"。 前者是可变成本，后者是固定成本。对财务报表的影响是：费用从"随收入波动"变成"先于收入发生"，经营杠杆被放大——收入好时利润弹性更大，收入差时亏损更深。

3.2 承诺的期限结构：本报告认为最重要的单一数字

截至 2026 年 9 月 30 日及之后，Anthropic 签订的采购承诺与租赁协议合计约 **9,009** 亿美元，其中 2030 年底前预计发生约 **4,516** 亿美元，未来 12 个月内到期约 **707** 亿美元，其余主要在 2031–2045 年（A-32）。

把这个数字放进五个分母里看：

分母	数值	承诺/分母	含义
2026 年前三季度收入	332.42 亿美元	**27.1** 倍	约 27 个前三季度的收入，或约 13 年的当前年化收入
2026-10-31 Run-Rate Revenue（口径见 2.2）	810 亿美元	**11.1** 倍	约 11 年的当前 run-rate 收入（较年化收入更宽的口径）
2026 年前三季度算力费用	188.08 亿美元	**47.9** 倍	约 48 个前三季度的算力支出，或约 36 年的当前年化算力支出
期末现金及短期投资	1,318.45 亿美元	**6.8** 倍	现有现金无法覆盖承诺
IPO 后备考现金及短投	2,126.84 亿美元	**4.2** 倍	募资后仍无法覆盖

数据来源：S-1/A（A-32、A-07、A-37、A-39、A-55）。比值由本报告计算（`scripts/runrate.py`）。

前三行都是当期的经营流量（第一、三行是九个月的实际发生额，第二行是 10 月底的年化率），后两行是存量；承诺是横跨二十年的累计额。 比值的含义是「多少个当期水平才能抵一次承诺」，不是「需要几年还清」。按当前年化收入计，承诺约相当于 **13** 年；即使按最宽的 run-rate 口径，仍相当于 约 **11** 年；按当前年化算力支出计约 **36** 年。

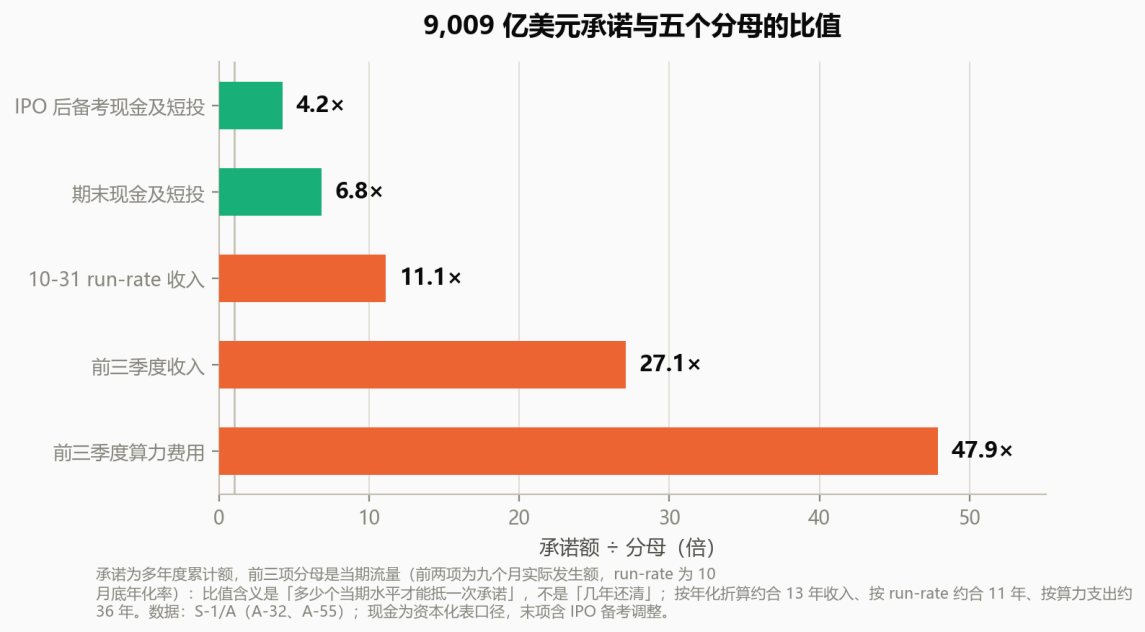


图 3–1

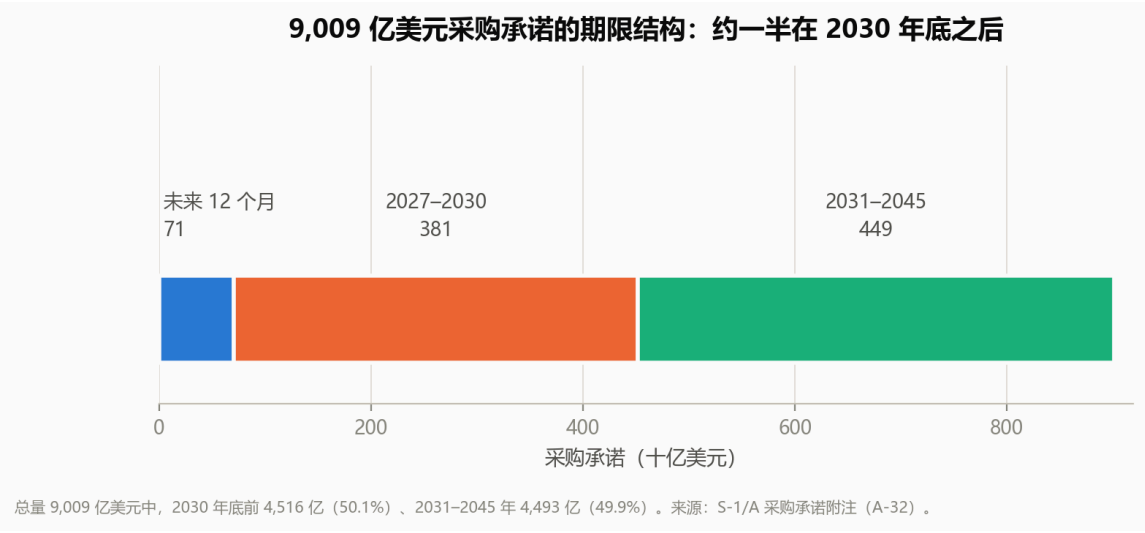


图 3–2

这张表是理解模型层风险的核心。一家公司承诺了相当于约 36 年当前年化算力支出的采购额，而其现金（即使完成 IPO）只相当于承诺额的 24%。这意味着：

- 承诺必须靠未来的经营现金流兑现，而不是靠现有资金。2026 年前三季度经营性现金流为正 65.69 亿美元（A-49），年化约 88 亿；即使按这个速度，覆盖 4,516 亿美元的 2030 年前承诺需要 51 年。
- 承诺的兑现依赖收入继续高速增长。若 2027 年收入达到 1,500 亿美元（本报告不做此预测，仅作敏感性说明），算力费用按当前 56.6% 的收入占比约为 850 亿，仍低于承诺隐含的年均 1,000 亿量级。
- 承诺本身是可重新谈判的。S-1 提到多方可的结构与融资方依赖（见 3.4），这意味着部分承诺可能在执行中调整。但调整的代价是容量损失，而容量损失直接限制收入上限。

> 本报告的判断：**9,009** 亿美元承诺不是"资产负债表外的负债"，而是一份"以未来收入为抵押的期权"。它在需求持续超预期时是护城河（锁定了别人拿不到的容量），在需求放缓时是绞索（用不满也要付）。同一份合同在不同的需求情景下符号相反，这正是它难以被传统财务比率捕捉的原因。

3.3 收入增速与算力增速的剪刀差

Anthropic 披露了 2026 年第三季度的一组对比：收入同比增长 **1,127%**，算力与基础设施费用同比增长 **382%**（A-46）。

这组数字是本报告认为最值得关注的效率证据，原因在于它同时否定了两种流行叙事：

它否定了"算力投入必然线性转化为收入"的叙事。 如果算力与收入是固定比例，两者增速应该一致。实际是 3:1，说明单位算力产生的收入在快速上升。

它也否定了"模型层永远烧钱"的叙事。 费用增速远低于收入增速，意味着经营杠杆正在起作用。

但需要给这个结论加上三条限制：

1. 这是单季同比，基数效应极强。 2025 年 Q3 收入基数只有 13.86 亿美元，任何增长都会显得巨大。
2. 算力费用中不含资本开支。 2026 年前三季度购置固定资产仅 9.25 亿美元，租赁资产与设备预付款 19.55 亿美元（A-50），合计 28.80 亿——相对于 **9,009** 亿美元的承诺，表内资本开支几乎可以忽略。 这说明大量算力投入尚未进入费用，未来会以折旧或租金的形式到来。
3. 效率改进有物理上限。 单位任务成本下降 40%（A-45）来自模型架构、量化、蒸馏、缓存等工程手段，这些手段的边际收益会递减。

3.4 芯片多元化：为什么它不是纯粹的技术选择

Anthropic 的算力横跨四种芯片：AWS Trainium、Google TPU、NVIDIA GPU、AMD GPU（A-47）。S-1 把这称为"multi-chip compute architecture"，并强调可以"orchestrate workloads to optimize for price and performance"。

从供给安全角度看，这是理性的： 任何单一芯片供应商的产能、交付节奏、定价都可能成为瓶颈，多元化降低单点依赖。

从成本角度看，这是被迫的： 模型层没有议价权。NVIDIA 的 GPU 面对的是全行业的超额需求，云厂商的自研芯片（Trainium、TPU）只对自己的客户开放且往往与承诺绑定。所谓"优化价格性能比"，实际是在有限的几个选项里做配置，而不是在完全竞争市场里采购。

从财务角度看，这制造了一个隐性风险： 不同芯片的工具链、算子库、并行策略不同，跨芯片训练与推理需要额外的工程投入。这部分投入不体现在"算力与基础设施"这一行里，而体现在研发人员成本里，因此从报表上不可见。

S-1 还披露了一项具体安排：2026 年 3 月与 Broadcom 签订设备租赁协议，采购专用加速器及相关组件，Broadcom 随后将部分安排转让给指定金融方，并可能继续转让（A-35 同节）。这是"厂商 + 金融方"的三方结构：芯片来自 **Broadcom**，资产由金融方持有，**Anthropic** 租用。 这种结构把资本开支移出表外，但把租金变成长期固定费用——与承诺采购的经济实质相同。

3.5 人才：5,300 人与 666 万美元人均收入

Anthropic 的员工数从 2024 年底约 940 人增至 2026 年 9 月底约 5,300 人（A-30），不到两年增长 4.6 倍。同期收入从 2024 年全年 3.86 亿美元增至 2026 年前三季度 332.42 亿美元。

5,300 人是一个需要放在行业里看的数字。 对比：OpenAI 在 2025 年公开的员工规模在数千人量级；Google DeepMind 约 5,000–6,000 人；Meta 的 AI 部门在 2025 年重组后约数千人。模型层的第一梯队，人员规模都在 **5,000** 人上下，而不是几万人。 这是一个由"研究密度"决定的行业结构——能力来自少数顶尖研究员，而不是来自人力规模。

S-1 对人才风险的表述值得引用：公司的文化"low-ego, truth-seeking"，并把人才密度列为竞争优势（A-44 同节）。这类表述在招股书里通常是套话，但在模型层有实质含义： 顶尖研究员的边际产出差异极大，一名核心研究员的离开可能使某个技术路线的进度延后数个季度。这使得模型层的人才风险无法通过"多招人"对冲。

3.6 供给约束的三种表现形式

形式一：容量分配。 S-1 明确写道，即使有合同承诺，"our ability to obtain such capacity remains dependent on those providers' operational execution … and broader business considerations, such as capacity constraints, internal demand for compute to support their own AI and computing operations"。翻译过来：云厂商会优先满足自己的模型业务。 这不是合同违约，是合同留出的裁量空间。

形式二：融资方依赖。 "certain of our compute providers and infrastructure partners depend on third-party financing or credit support to acquire chips and develop capacity, and if such financing or support becomes unavailable or more costly, contracted capacity may be delayed, repriced, reduced, or not delivered."（A-35 同节）这意味着承诺的兑现还取决于第三方融资市场的状况——一个模型公司无法控制、也无法对冲的变量。

形式三：交付时序。 多方可结构下，"we may be obligated to pay a counterparty once it performs its discrete obligation, even if delays or failures by others prevent us from assembling those discrete services or inputs into usable compute"。换句话说：钱可能付了，算力却拼不起来。 这是把供应链风险直接转化为现金流出风险。

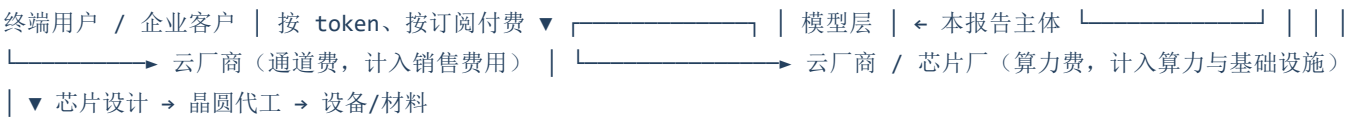
3.7 本章结论

1. 模型层的算力取得方式正从"按需租用"转向"多年承诺"，成本结构从可变转向固定，经营杠杆被放大。
2. 9,009 亿美元承诺相当于约 36 年当前年化算力支出、约 11 年 10 月底的 run-rate 收入、6.8 倍当前现金；它是一份"以未来收入为抵押的期权"，符号随需求情景翻转。
3. 表内资本开支（28.80 亿美元）相对承诺额几乎可忽略，说明大量算力成本尚未进入费用表，未来将以折旧或租金形式到来。
4. 收入增速是算力费用增速的 3 倍，效率改进真实存在，但基数效应极强且有物理上限。
5. 供给约束不只来自芯片产能，还来自云厂商的分配裁量、第三方融资可得性与多方可结构的交付时序——三者都不在模型层的控制范围内。

第四章 价值链与价值分配

4.1 链条结构：谁向谁付钱

大模型产业链的现金流方向可以用一句话概括：下游按 **token** 付费，上游按算力收费，中间层承担全部的资本开支风险。



这条链上有两个不对称：

不对称一：模型层对下游是"按量收费"，对上游是"按承诺付费"。收入随客户用量波动，成本却由多年期 take-or-pay 合同锁定。这是全链条唯一承担"需求下行风险"的环节。芯片厂按订单生产，云厂商按合同收款，只有模型层在中间承担用不满的风险。

不对称二：模型层对下游没有定价权，对上游也没有议价权。下游面对开源模型的零边际成本替代（第五章），上游面对全行业的超额需求。模型层被夹在两个都有替代品的市场之间，这是它毛利率波动巨大的结构性原因。

4.2 云厂商的三重身份：供应商、渠道、股东

理解模型层最关键的一件事，是认识到云厂商同时扮演三个角色，而这三个角色在报表上分散在不同科目，在分析时却必须合并看：

身份	具体表现	报表落点	对模型层的影响
供应商	提供算力、机房、网络	算力与基础设施费用（2026 前三季 188.08 亿，占收入 56.6%）	决定成本曲线的斜率
渠道	CSP marketplace 代收代付	收入按总额法确认；通道费计入销售与营销	抬高收入，同时抬高销售费用
股东	持有优先股与可转债	可转债按公允价值计量，变动计入损益	估值上升→账面巨亏；同时摊薄股权

数据来源：S-1/A 附注 2、附注 8、风险因素（A-16、A-19、A-20、A-22、A-35）。

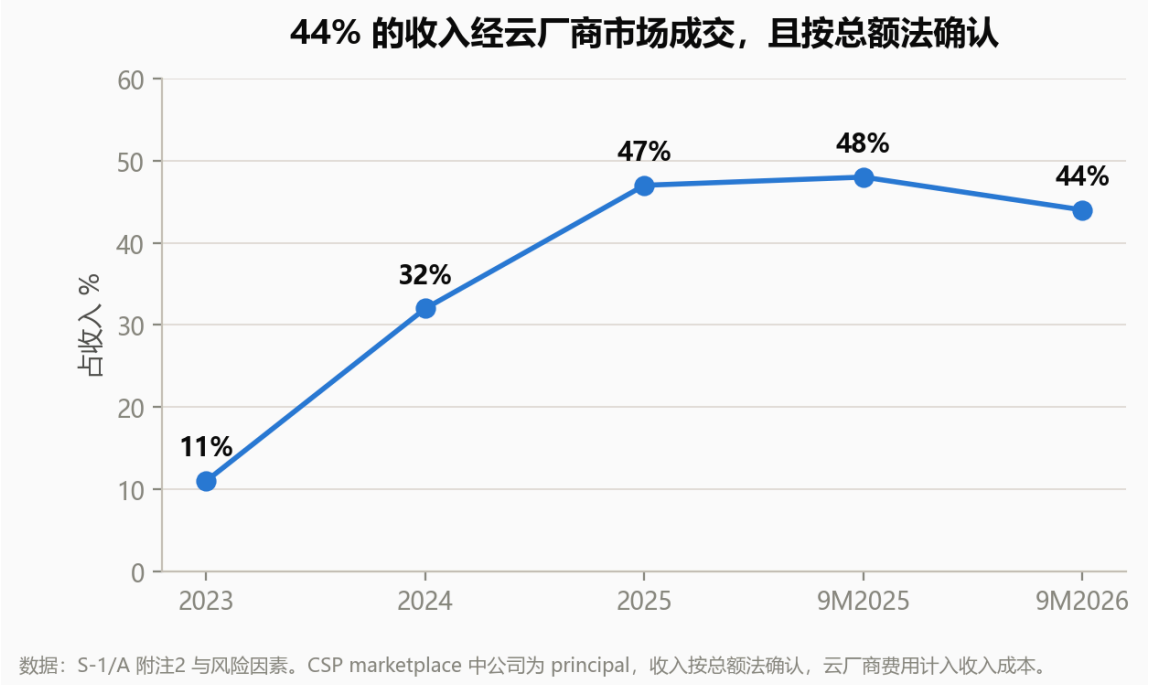


图 4–1

这三重身份构成一个闭环：模型层向云厂商买算力（付钱）→ 通过云厂商市场卖模型（云厂商抽成）→ 云厂商把收到的钱投回模型层（买股权/可转债）→ 模型层用这笔钱继续买算力。

从会计上看，这个闭环的每一环都是独立、公允、可辩护的。从经济上看，它是一个自我循环。判断这个循环是否健康，唯一的标准是：终端客户是否在用真金白银为 **token** 付费，且付费额是否超过模型层支付给上游的算力成本。

表 4-1 闭环的四个环节与可验证的数字

环节	方向	2026 年前三季度金额	可验证性
终端客户付费	客户 → 模型层	收入 332.42 亿美元	高（经审计报表）
模型层付算力	模型层 → 云/芯片	算力与基础设施 188.08 亿美元	高
模型层付通道费	模型层 → 云	未单独披露（含于销售与营销 66.19 亿）	低
云厂商投资	云 → 模型层	累计融资 1,411 亿美元（含全部投资方）	中（总额，未拆分）

第四行是本报告认为最需要警惕的披露缺口。公司披露了累计融资 1,411 亿美元（A-36），但未披露其中有多少来自算力供应商。而可转债的账面价值（1,305.02 亿）相对于本金（62 亿）的 21 倍溢价（A-35、A-38），说明供应商身份带来的股权收益，量级上已经超过了它作为供应商收取的算力费。

> 这是本章最重要的判断：在模型层，云厂商赚的不是算力费，是股权增值。算力费是按成本加成定价的、可谈判的、有上限的；股权增值是按模型层估值定价的、无上限的。因此，"云厂商在补贴模型层"和"模型层在给云厂商打工"这两种说法都不准确——真实的关系是：云厂商用算力换股权，模型层用股权换算力。这是一种以股权为媒介的资源置换，而不是买卖。

4.3 价值分配：收入流向了哪里

把 2026 年前三季度的 332.42 亿美元收入按去向拆开：

表 4-2 收入的去向（2026 年前三季度）

去向	金额（亿美元）	占收入	性质
算力与基础设施	188.08	56.6%	支付给云厂商/芯片供应链
人员	48.63	14.6%	支付给员工（含部分股权激励）
销售、营销与伙伴	55.29	16.6%	含渠道费、市场费用、免费用户算力
其他管理	53.02	15.9%	含捐赠匹配、法务、设施
营业亏损	**−12.60**	**−3.8%**	收入不足以覆盖全部费用

数据来源：S-1/A 合并利润表（A-07）。占收入比由本报告计算，合计 103.8%。

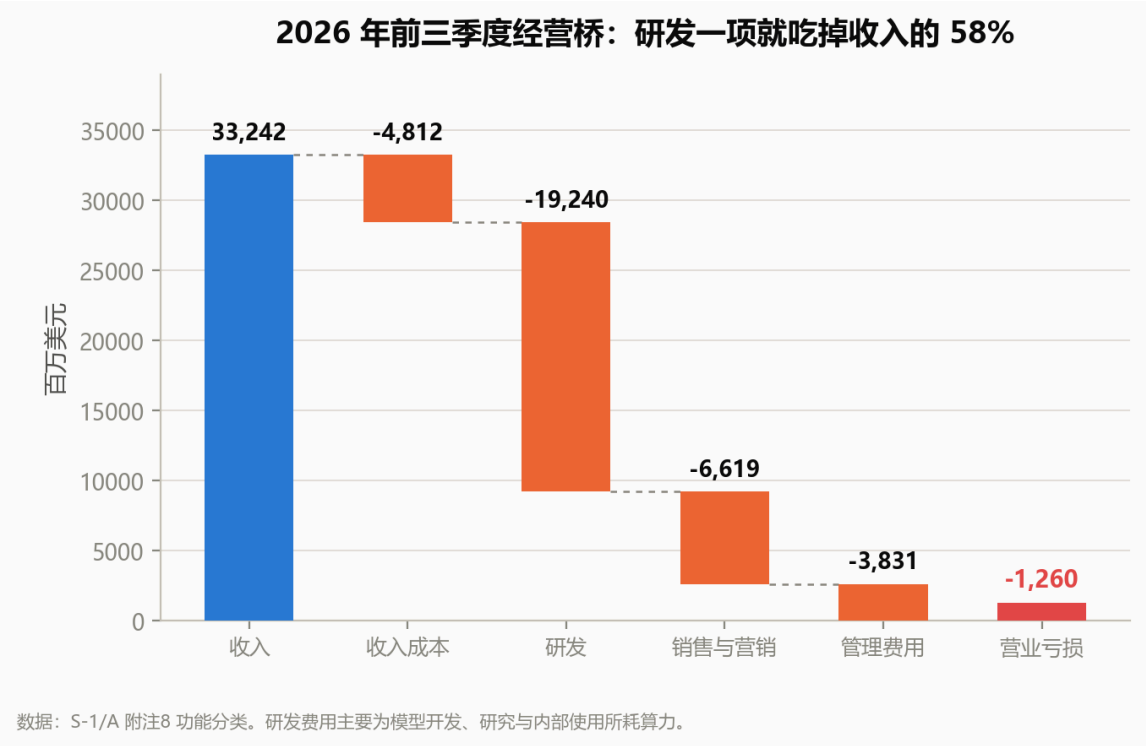


图 4–2

这张表说明：模型层目前没有留下经营利润。 全部收入，加上额外 3.8%，都付给了上游、员工和渠道。

公司口径的"调整后经营利润"为正 40.81 亿美元、利润率 12%（A-12、A-13）。但这个口径加回了 **32.55** 亿美元的股权激励和 **20.86** 亿美元的非现金慈善捐赠（A-11、A-12 调节表）。换句话说，**12%** 的调整后利润率，全部来自把两项非现金支出加回。

> 本报告的判断：把 **SBC** 加回是标准的 **Non-GAAP** 做法，但在模型层这一做法具有误导性。 对一家每年以数亿美元股权补偿员工的公司，SBC 是真实的股东成本——它体现为股本的持续稀释。2026 年前三季度 SBC 32.55 亿，占收入 9.8%；同期发行后股本 130.87 亿股，而完全稀释股本约 155 亿股（A-42），摊薄悬顶约 **18%**。加回 **SBC** 后的 **12%** 利润率，在扣除真实稀释成本后接近零。

4.4 一个反事实检验：如果算力按市价计价

模型层披露的毛利率是在关联交易框架下形成的。要判断这个毛利率是否可持续，可以做一次思想实验：

假设 **1**： 当前算力采购全部按现货市价结算，无任何股东折扣。 假设 **2**： 云厂商的通道费按行业常见的 3%–5% 收取。

在假设 2 下，44% 的收入经云厂商市场成交（A-20），通道费约为 $332.42 \times 44\% \times 4\% \approx$ **5.85** 亿美元，占收入 1.8%。这部分已在销售费用中，影响有限。

真正的未知数是假设 **1**。 若当前算力有 10% 的股东折扣，则成本被低估约 18.81 亿美元，营业亏损将从 −12.60 亿扩大到 −31.41 亿；若折扣为 20%，亏损扩大到 −50.22 亿。由于 **S-1** 未披露算力采购的定价机制，这个折扣率无法从公开信息中验证，本报告将其列为"已知未知"（**known unknown**），并在第八章的估值敏感性中作为独立情景处理。

> 这就是本章要说明的核心问题：模型层的毛利率不是市场竞争的结果，而是关联交易定价的结果。 在云厂商既是股东又是供应商的结构下，毛利率可以被人为维持在任何水平。因此，用毛利率给模型层估值，等于用一个由交易对手方决定的参数做分母。

4.5 与历史技术周期的结构对比

把模型层放进更长的技术周期里看，它的结构与三个历史形态有相似之处，但有一处关键不同：

对比对象	相似之处	关键不同
半导体代工（Foundry）	资本密集、产能承诺、规模效应极强	代工有明确的制程节点与客户长约，模型层没有
电信运营商（3G/4G 时代）	巨额资本开支、按用量收费、被上游设备商挤压	运营商有牌照壁垒与频谱稀缺性，模型层没有
云计算（IaaS）	规模效应、按用量计费、前期亏损后期盈利	云的边际成本随规模下降且客户切换成本高，模型层两者都弱

最关键的不同在于切换成本。 云计算的客户一旦把系统迁移上去，迁移成本极高，因此云厂商可以承受多年亏损并在后期提价。模型层的客户切换成本接近于零——改一个 API endpoint 就能换模型，S-1 自己承认客户会"route individual tasks to the models that are most optimized"（A-44 同节）。没有切换成本，就没有定价权；没有定价权，前期亏损就无法在后期收回。

> 这是本报告对"模型层会像云计算一样在规模后盈利"这一类比的反驳。 云计算的盈利来自锁定，模型层的收入来自不锁定。两者的规模效应相似，但变现路径相反。

4.6 本章结论

1. 模型层是全链条唯一承担需求下行风险的环节：对下游按量收费，对上游按承诺付费。
2. 云厂商的三重身份（供应商、渠道、股东）构成一个自我循环；判断循环健康与否的唯一标准是终端付费是否超过算力成本。
3. 2026 年前三季度，收入的 103.8% 被算力（56.6%）、人员（14.6%）、销售（16.6%）、管理（15.9%）吃掉，经营亏损 12.60 亿美元。
4. 调整后 12% 的利润率全部来自加回 SBC 与非现金捐赠；扣除 18% 的摊薄悬顶后接近零。
5. 毛利率是在关联交易定价下形成的，不是市场竞争的结果；算力折扣率是已知未知，无法从公开信息验证。
6. 模型层缺乏云计算的客户锁定，因此不能套用"先亏损后盈利"的云叙事。

第五章 竞争格局

5.1 竞争发生在三个维度，而不是一个

行业讨论习惯把模型层竞争简化为"谁的模型更强"。这是把三个独立的竞争维度压缩成了一个。实际的竞争同时发生在：

维度	竞争内容	可观察指标	领先者的优势能维持多久
能力	模型在基准与真实任务上的表现	榜单、客户迁移、单位任务成本	约 2 个季度（见 5.5）
成本	每 token、每任务的推理成本	单位任务成本降幅、毛利率	6–12 个月，随工程扩散
分发	客户从哪接触到模型	渠道占比、企业客户数、交叉销售	数年，取决于合同与集成深度

这三个维度的领先者不一定是同一家公司，而且优势的持久性差一个数量级。能力领先半年度就消失，成本领先一年被追上，分发领先可以维持数年。因此，判断模型层竞争格局，关键不是"谁现在最强"，而是"谁的优势落在最难复制的维度上"。

5.2 玩家图谱：四个梯队

按"是否自训前沿基础模型"与"分发能力"两个轴，模型层可以分成四个梯队。本节的分类依据是各公司公开披露的模型发布与产品形态，不涉及能力排序——能力排序需要独立评测，本报告不做。

第一梯队：前沿自训 + 自有分发

主体	模型族	分发渠道	关键结构性事实
Anthropic	Mythos（含 Fable）、Opus、Sonnet、Haiku	自有 API + 三大云市场	44% 收入经云市场；9,009 亿美元算力承诺
OpenAI	GPT 系列	自有 + Azure	与 Microsoft 的独家云关系
Google DeepMind	Gemini	自有 + Google Cloud + 安卓/搜索入口	自有 TPU，是唯一"芯片—云—模型"全栈自持者
Meta	Llama	开源权重 + 自有产品	以开源换生态，不以模型收入为主要目标
xAI（2026-02 起为 SpaceX 子公司）	Grok	自有 + X 平台 + 对外算力	2026-02 SpaceX 全股票收购；财务并入 SpaceX "SI" 分部，已无独立报表

数据来源：Anthropic 部分来自 S-1/A（A-44）；xAI 状态来自 SpaceX 10-Q（SEC 一手，CIK 0001181412）；其余为公开信息，属本报告对公开产品形态的整理。

> 关于 **xAI** 的一个必要更正： 2026 年 2 月 SpaceX 以全股票方式收购 xAI，SpaceX 于 2026 年 6 月在纳斯达克上市（SPCX）。此后**"xAI** 的估值**"**与**"xAI** 的收入**"**都不再是可独立观察的量。 其财务并入 SpaceX 的 "SI" 分部，而该分部同时包含 X 的广告收入与 2026-08 并入的 Cursor。任何按分部倒推出来的**"xAI** 收入**"**，都不是纯净的模型业务收入。 本报告因此不对 xAI 单独做任何倍数计算。

这一梯队里有一个结构性差异值得单独指出：只有 **Google** 同时拥有芯片（**TPU**）、云、模型三层。 这意味着 Google 的模型业务不受"算力采购承诺"约束——它的算力是内部转移定价，不是外部采购。当其他玩家在为 **9,000** 亿美元级别的承诺承担需求风险时，**Google** 承担的是资本开支风险，两者的风险性质完全不同。 这是本报告认为模型层最被低估的结构性不对称。

第二梯队：前沿自训 + 第三方分发

Mistral AI（欧洲）、Cohere（企业向）等。这一梯队的共同特征是不做全栈，专注模型本身，通过云市场或直销分发。它们的战略空间受制于第一梯队的定价与开源模型的下压，中间地带正在变窄。

第三梯队：中国模型层

阿里通义千问、字节豆包、DeepSeek、智谱、月之暗面、MiniMax、腾讯混元等。这一梯队的结构特征与海外不同：

- 分发高度依赖自有流量入口（阿里云、抖音、微信、浏览器），而不是企业直销；
- 价格战更激烈，因为开源权重与国产算力降低了进入门槛；
- 受监管前置审批约束（生成式 AI 服务备案制，见第七章）。

中国模型层的收入规模、盈利状况与海外不可直接比较，本报告在第七章单列，不并入全球模型层的估值框架。

第四梯队：开源权重提供者

Llama、Qwen、DeepSeek 的开源版本。它们不是市场参与者，但是价格约束者。 见 5.3。

5.2.1 一个外部锚：企业 API 支出的实际份额

上面的梯队划分依据的是"是否自训 + 是否自有分发"，是结构判断。结构判断需要与钱的实际流向对照。

表 5-1 企业 LLM API 支出的份额（2026 年，截至 9 月）

主体	2026	2025	2024
Anthropic	**55%**	40%	24%
OpenAI	31%	27%	50%
Google	9%	21%	—
Anthropic + OpenAI	**86%**	67%	—
含 Google 三家合计	**95%**	—	—
xAI + 开源合计	略低于 **5%**	—	—

数据来源：Menlo Ventures《2026: The State of Enterprise AI》，2026-11-03。方法：调查近 400 名美国企业决策者（Morning Consult 执行，2026-09-08 至 09-26），按 AI 支出加权；份额定义为"基于生产环境 **API** 使用比例估算的支出美元"。

这张表支持三件事：

1. 结构判断与钱流一致。 本报告把 Anthropic、OpenAI、Google 列为第一梯队，表 5-2 显示三家合计占企业 LLM API 支出的 95%。梯队划分不是凭印象的。
2. 份额在向 **Anthropic** 集中。 从 2024 年的 24% 到 2026 年的 55%，两年翻倍以上；同期 OpenAI 从 50% 降到 31%。
3. 开源在企业 **API** 支出里几乎不占份额。 xAI 与开源合计不足 5%。

但第 **3** 点恰好印证了第五章的核心论点，而不是推翻它。 开源不占支出份额，却在压制价格——它通过让客户在续约时有一个免费替代选项来压缩单价，而不是通过自己产生收入。用一个支出份额去衡量一个价格约束，本来就会得到"接近零"的读数。

三条必须写明的口径限制：

- 分母是"企业 **LLM API** 支出"，不是模型公司的全部收入。 Anthropic 的消费者订阅、第一方应用收入不在其中。这个份额不能用来推算任何公司的总收入结构。

- 同一份报告里还有另一个分母。 Menlo 另称前沿实验室合计只占"企业 AI 部署支出"（1,227 亿美元）的 69%。两个分母不可互相换算，也不可与其他机构的**"LLM 市场份额"**直接比较。
- 调查对象是美国企业决策者，不是全球。 结论的外推范围限于美国企业市场。

> 来源级别提示：这是第三方调查估算，不是审计数据。 它比单一公司的披露弱，但比行业叙事强——它是本报告能找到的、唯一把"竞争格局"折算成钱的公开口径。

5.3 开源：模型层的价格天花板

闭源模型层的定价权，受制于一个简单的事实：开源权重的边际成本为零，且能力在追赶。

这个约束的传导路径是：

1. 企业客户在"闭源前沿模型"与"自部署开源模型"之间做选择；
2. 当开源模型的能力足以覆盖 80% 的日常任务时，闭源模型只能对剩余 20% 的高难度任务收取溢价；
3. 因此闭源模型层的收入结构会向"高难度、高价值"任务集中，总量增长但均价下降。

Anthropic 的数据与这一判断一致：单位任务成本下降约 40%（A-45）而收入增长 12 倍（A-05），说明降价确实换来了用量。但这条路径有一个终点：当开源模型覆盖 **95%** 的任务时，闭源模型的市场只剩下 **5%** 的极端任务，其收入规模可能无法支撑 **9,000** 亿美元的算力承诺。

> 本报告的判断：开源不是模型层的竞争对手，是模型层的定价上限。 它不会抢走客户，但会让客户在每一次续约时要求降价。这解释了为什么模型层必须在开源追上之前，把收入转化为分发关系与算力锁定——因为价格优势会消失，关系不会。

5.4 竞争的结果：三个可观察变量

竞争格局不能用"谁强"来描述，只能用结果变量来描述。本报告跟踪三个：

变量一：单位任务成本。 这是竞争最直接的读数。Anthropic 披露 Opus 5.5 相对 Opus 5 下降约 40%（A-45）。若这一降幅在行业层面普遍成立，则意味着每 **12–18** 个月，同样任务的价格腰斩。 模型层必须用翻倍的用量来维持收入。

变量二：毛利率的分化。 若各家的毛利率在同一时间同向变动，说明是行业性因素（算力价格、推理效率）；若分化，说明是结构性因素（渠道、客户结构、关联交易定价）。**Anthropic 2024** 年毛利率 **−94.8%**、**2026** 年前三季度 **85.5%**（**A-18**），这个幅度不可能是行业普遍水平，更可能是规模效应与关联交易定价的叠加。这提示：用单一公司的毛利率代表行业，会严重误导。

变量三：客户集中度与留存。 前五大客户占比 25%、NRR 900%（A-25、A-27），说明 Anthropic 的客户结构在分散而非集中。这是竞争格局中最正面的信号——如果增长依赖少数大客户，那么任何一个客户的流失都会造成断崖；25% 的集中度意味着单一客户流失的影响可控。

5.5 护城河检验：四个候选，一个半通过

模型层最常见的四个"护城河"叙事，本报告逐一检验：

候选一：技术领先

判决：不成立，是入场费而非护城河。

证据：2026 年 3 月至 9 月发布的模型版本贡献了 Q3 收入的 80% 以上（A-29）。这意味着半年前的模型版本只贡献不到 **20%** 的收入。 技术领先的商业价值半衰期约为两个季度。一家公司必须持续发布才能维持收入，一旦停摆，收入在两个季度内反映。

这与护城河的定义相反：护城河是"不做事也能守住的东西"。技术领先是"不做就失去的东西"。

候选二：客户切换成本

判决：不成立。

证据：S-1 自己描述客户会"route individual tasks to the models that are most optimized to meet their needs along the capability-cost-speed curve"（A-44 同节）。客户在多模型之间路由，是产品设计的核心功能，不是例外情况。

切换成本存在的唯一场景是深度集成（客户把模型嵌进自己的产品并针对特定模型调优）。交叉销售数据（**50%** 的 **≥10** 万美元客户同时使用平台与应用，**A-48**）说明集成在加深，但集成加深的是"用得多"，不是"换不了"。

候选三：算力锁定

判决：成立，但符号双向。

证据：9,009 亿美元采购承诺（A-32）。在一个算力供不应求的市场里，锁定容量就是锁定收入上限——竞争对手即使有更好的模型，没有算力也服务不了客户。

但这份承诺同时是负债：用不满也要付。它是护城河还是绞索，取决于需求能否持续超过承诺隐含的水平。 本报告在第三章已经论证：承诺隐含年均约 1,000 亿美元的算力支出，是当前年化算力费用的四倍。

候选四：企业与分发关系

判决：部分成立。

证据：11,900 家年付超 10 万美元的客户、1,900 家超 100 万美元、250 家超 1,000 万美元（A-26）；前五大仅占 25%（A-27）；NRR 约 900%（A-25）。

这是四个候选里最接近传统护城河的一个。 企业客户的采购流程、安全审查、集成工作、内部培训都构成摩擦，摩擦就是切换成本。但它的强度取决于模型是否被嵌入客户的关键工作流——如果只是"员工用 Chat 界面问答"，切换成本接近零；如果嵌入了代码库、客服系统、合同流程，切换成本就很高。

Anthropic 的数据显示集成在加深：Q3 2026 有超过 50% 的大客户同时使用平台与第一方应用，而 2025 年 Q4 约 30%（A-48）。这是本报告认为该公司最值得跟踪的单一指标——它比收入增速更能预测三年后的竞争地位。

表 5-2 护城河检验汇总

候选	判决	关键证据	可持续年限（本报告估计）
技术领先	✗ 入场费	半年内版本贡献 >80% 收入	0.5 年
客户切换成本	✗ 不成立	客户在多模型间路由是产品功能	<1 年
算力锁定	√ 但双向	9,009 亿美元承诺	5–10 年（合同期）
企业与分发关系	◐ 部分成立	1.19 万家企业客户、NRR 900%、交叉销售 30%→50%	3–5 年

"可持续年限"为本报告基于合同期限与行业惯例的估计，非实测数据，使用时请注意其主观性。

5.6 本章结论

1. 竞争发生在能力、成本、分发三个维度，三者的优势持久性差一个数量级：能力约两个季度，成本约一年，分发可达数年。

2. 只有 Google 同时自持芯片、云与模型，其算力是内部转移定价而非外部采购——这是模型层最被低估的结构性不对称。

3. 开源不是竞争对手，是定价上限；它通过迫使闭源模型持续降价来压缩收入，而不是通过抢客户。

4. 四个护城河候选中，技术领先与切换成本不成立；算力锁定成立但符号双向；企业与分发关系部分成立。

5. 最值得跟踪的单一指标不是收入增速，而是"同时使用平台与第一方应用的大客户占比"——它从 30% 升到 50%，是集成加深的直接证据。

第六章 行业阶段与技术趋势

6.1 判断阶段需要三个坐标系，而不是一个

"行业处在什么阶段"这个问题，通常被回答成"处在早期"或"处在泡沫期"。这两种回答都没有信息量，因为它们没有说明用什么坐标测量。本报告用三个相互独立的坐标系：

坐标系	测量什么	当前读数	含义
能力	模型能可靠完成的任务范围	快速扩张，但边界仍清晰	技术曲线未见顶
成本	单位任务的价格	每年下降约 40%（公司披露口径）	通缩型行业
渗透	有多少工作流真正被改变	企业采用快，个人渗透慢	S 形曲线的前段

三个坐标系的读数不一致，这正是当前阶段的核心特征：能力与成本在指数变化，渗透在 S 形变化。用前两个外推第三个，是当前市场最主要的分析错误。

6.2 能力曲线：从"更大"到"更会用"

S-1 描述了一个值得注意的机制：**"closing the loop"**——Claude 本身参与下一代 Claude 的研发，包括运行实验、生成与处理训练数据、自动化评测与安全审计、编写与审查代码（A-44 同节）。

这个机制的经济含义是：研发的边际成本在下降。传统上，模型能力的提升需要线性增加研究员与算力；如果模型能承担一部分研发工作，同样的投入可以产出更多进展。公司把这称为"compute multipliers"——用同样的算力获得更多智能。

> 本报告的判断（属推断，非公司披露）：如果这一机制成立，它会同时推高三件事——研发速度、算力需求的增速、以及单位算力的产出。三者同向，因此从外部很难观察到它是否真的在起作用。唯一可观察的间接证据是：收入增速（+1,127%）显著高于算力费用增速（+382%）（A-46）。但这个证据是混杂的——它同时可以被"利用率提升"、"定价结构变化"、"关联交易定价"解释。本报告不把**"compute multiplier"**当作已验证的事实，只当作公司提供的、尚未被独立验证的机制解释。

S-1 对"Powerful AI"的定义是"认知能力在大多数专业领域匹配或超过顶尖人类专家"（A-44 同节）。这是公司的愿景表述，不是可验证的预测。报告引用它，只用于说明公司自身的技术路线假设，不作为行业判断的依据。

6.3 成本曲线：通缩是结构性的

Anthropic 披露的单位任务成本下降幅度是：**Opus 5.5** 相对 **Opus 5** 下降约 **40%**（A-45）。

这个数字需要放在正确的框架里理解。它不是"同一模型变便宜了"，而是"新一代模型完成同样任务更便宜"。两者的经济含义完全不同：

- 如果是同一模型降价，那是价格战，会压缩毛利；
- 如果是新一代模型更高效，那是技术进步，可以在保持甚至提高毛利的同时降价。

公司明确说这是后者："Our compute efficiency allows us to deliver these reductions while preserving our margins."（A-45）

这与毛利率从 **29.4%**（**9M2025**）升到 **85.5%**（**9M2026**）是一致的（A-18）：降价 40% 的同时毛利率提升 56 个百分点，说明成本下降的速度快于价格下降的速度。这是一个健康的通缩——它扩大市场而不是压缩利润。

6.3.1 外部检验：独立机构测出的成本曲线

公司自述的"40%"需要有外部对照，否则只是自我评价。目前可用的两组独立数据如下：

表 6-1 独立测得的 AI 成本下降速度

指标	数值	口径	来源
同等能力成本下降	**−47%** / 季度（约 **13×** / 年）	"达到指定准确率所需的每题美元成本"，5 个基准（AIME、Chess、FrontierMath、GPQA Diamond、Mystery Game），2023–2026	Epoch AI，2026-09-22
其中：能力刚成为 SOTA 时	−66% / 季度（75× / 年）	同上，logistic 回归定 SOTA 时点	同上
其中：同一能力 8 个季度后	−32% / 季度（4.7× / 年）	同上	同上
例证	o3 在 GPQA Diamond 达 75% 约 $0.30/ 题（2025-01）→ 同等能力约 $0.0004/题（2026 年中），**725×**	同上	同上
平台实付均价	**$0.77 → $0.14** / 百万 **token**（约 **−82%**）	OpenRouter 平台，2025-01 → 2026-09	OpenRouter，2026-10-22
同期平台 token 量	0.5T → 约 100–160T / 周	同上（厂商自述 164T；本报告查公开端点 2026-11-09 为 109.4T）	同上

来源：epoch.ai《The Plunging Price of Thought》；OpenRouter 数据经公开帖文与公开端点。两者的时间单位不同（季度 **vs 18** 个月），不可直接换算成同一指标。

这三条外部数据能确认什么、不能确认什么：

能确认的： 成本下降是真实且极快的，量级与公司自述一致。尤其值得注意的是 **Epoch** 的分解——刚成为 **SOTA** 的能力降价最快（**−66%**/季度），**8** 个季度后降到 **−32%**/季度。 这直接支持本报告在 6.3 节的判断：成本下降的速度会衰减，因为快降的那部分来自"最新能力被迅速商品化"，而这一过程有终点。

不能确认的： 公司自述的"40%"无法被证实或证伪。它测的是"我们这一代比上一代便宜多少"（代际口径），**Epoch** 测的是"市场上达到同等能力最便宜要多少钱"（市场口径）。 后者包含"换一家更便宜的供应商"这个选项，前者不包含。两个口径不可换算，任何把它们对齐的说法都需要一个未公开的折算假设。

> 本报告的判断：外部数据支持"成本曲线陡峭"这一事实，但不支持任何具体的年化降幅数字。 本报告因此不把"每年降 40%"当作预测输入，只当作"降幅很大且会衰减"的方向判断。

6.3.2 一个被低估的需求侧证据：agent 消耗的 token 已超过人类

OpenRouter 披露的结构数据提供了一个本报告认为最重要的需求侧观察：**2026** 年 **9** 月下旬，**agent** 产生的 **token** 为每周约 **131** 万亿，人类交互约 **18** 万亿；**agent** 在 **2026** 年 **2** 月 **6** 日首次超过人类。

来源：OpenRouter（2026-10-22）。这是平台口径，不是全市场口径；**OpenRouter** 只是众多通道之一，其客户结构可能偏向开发者与 **agent** 场景。

这条数据的经济含义，比"**token** 总量涨了多少倍"更重要： 如果 token 消耗的主体已经不是人在对话框里打字，而是 agent 在后台反复调用，那么需求的价格弹性会显著下降——agent 不看价格，它看任务完成率。这解释了为什么单价降 **82%** 而支出仍在高速增长：降价打开的是机器用量，不是人类用量。

但这个机制有终点。 成本下降来自模型架构、量化、蒸馏、缓存、专用芯片等工程手段，每一项的边际收益都递减。当这些手段用尽时，成本曲线会从指数转为线性，届时降价将直接压缩毛利。 本报告无法判断这个转折点在何时，但可以指出：毛利率的走向是判断转折点是否到来的最灵敏指标。

6.4 渗透曲线：从"工具"到"协作者"

S-1 把产品演进描述为从"工具"到"Virtual Collaborators"——能够持续推理、跨多步骤执行、使用内部工具的 AI（A-44 同节）。Claude Code 被举为"最清晰的体现"。

从收入结构看，这个转变已经在发生：

- 超过 50% 的大客户同时使用平台与第一方应用（A-48），说明客户不再只买"API 能力"，而是买"完成工作的产品"。
- 半年内发布的模型贡献 80% 以上收入（A-29），说明客户在持续迁移到新能力，而不是固守旧用法。
- NRR 约 900%（A-25），说明存量客户在快速加深使用。

但从渗透率看，这个转变仍在前段。公司引用"知识工作者人口在 2028 年将超过 10 亿"作为目标市场（A-44 同节）——这是一个公司自己的估算，不是独立统计，本报告不采用其数值。可观察的替代指标是：11,900 家年付超 10 万美元的客户。放在全球企业总数（数亿家）里，这是一个很小的渗透率。

> 本报告的阶段判断：模型层处在"能力已经够用、渗透刚刚开始"的位置。这个位置的特征是——收入增长快（因为从零开始），但增长的质量取决于渗透深度而非能力提升。因此，未来两年的关键变量不是"模型还能强多少"，而是"多少工作流真正被重构"。

6.5 技术趋势的四条主线

主线一：从单次问答到长时程任务。S-1 反复强调"多步骤"、"持续推理"、"agent loop"、"multi-agent orchestration"（A-44 同节）。技术含义是从"生成"转向"执行"。商业含义是：每个任务的 **token** 消耗量级提升，因为长时程任务需要反复推理与工具调用。这是抵消单价下降的主要力量。

主线二：从单一模型到模型路由。客户在不同模型间路由任务（A-44 同节），意味着模型层的产品形态从"一个最强模型"变成"一组按成本性能排列的模型 + 一个路由器"。Anthropic 的四类模型（Mythos/Opus/Sonnet/Haiku）就是这个结构的体现。商业含义是：收入不再由最强模型决定，而由路由效率决定。

主线三：从通用到垂直。第一方应用（Claude Code 等）是针对特定工作流优化的产品。这条主线的含义是模型层正在向应用层延伸，与 Cursor、Harvey 等应用公司正面竞争。这既是增长来源，也是与客户（应用公司）的利益冲突。

主线四：从外部采购算力到自建。S-1 明确说公司计划"expand our own data center footprint to secure additional long-term compute"（A-44 同节）。这意味着模型层正在向上游垂直整合，未来会同时承担芯片采购、数据中心建设、能源供应的风险。这条主线会让模型层的资产负债表越来越像重资产公司，而不是软件公司。

6.6 阶段判断及其对估值的含义

把三个坐标系合起来，本报告对当前阶段的判断是：

判断	依据	置信度
能力曲线未见顶	半年内版本贡献 80% 收入，说明新能力仍在创造商业价值	中高
成本曲线仍在下降但接近线性	单代下降 40%，但工程手段有物理下限	中
渗透曲线在前段	1.19 万家企业客户，相对全球企业总数渗透率低	中
行业尚未进入整合期	多个前沿玩家并存，价格仍在下降而非寡头定价	中高

对估值的含义是直接的：

- 如果能力与成本仍在快速改善，那么高增长假设成立，估值可以承受高倍数；

- 如果渗透在前段，那么增长持续时间长，但增长速度会从指数转为线性；
- 如果行业未整合，那么毛利率无法长期维持在当前水平，因为竞争会把它压回。

第八章的估值分析会把这三点转化为具体的敏感性变量。

6.7 本章结论

1. 行业处在"能力够用、渗透刚起步"的位置，三个坐标系的读数不一致是当前阶段的核心特征。
2. 单位任务成本下降 40% 是技术进步而非价格战——降价同时毛利率上升，这是健康的通缩。
3. 成本下降的工程手段有物理下限，毛利率是判断转折点最灵敏的指标。
4. 四条技术主线中，最重要的是"从生成到执行"：它提升单任务 token 消耗，是抵消单价下降的主要力量。
5. 模型层正在向应用层（第一方应用）与上游（自建数据中心）双向延伸，资产负债表会越来越像重资产公司。

第七章 政策与监管

7.1 三种监管范式：不是"松紧"之分

讨论 AI 监管最常见的错误，是把不同辖区放在一根"严格—宽松"的刻度上比较。这是错的，因为它们管的不是同一件事。

辖区	监管范式	管的是什么	对模型层的经济传导
中国	事前备案制	能不能向公众提供服务	影响发布节奏与上市时间
欧盟	事后义务制	上线之后必须履行什么义务	影响合规成本与罚则上限
美国	无联邦专法 + 州法碎片 + 算力管制	算力能不能买到、州法怎么适用	影响供给与法律不确定性

这三种传导路径进入财务模型的位置完全不同。 备案制不罚款，但它决定一个模型版本能不能上线、什么时候上线；欧盟的义务不阻止上线，但它规定上线后必须做什么、做不到罚多少；美国的联邦层面没有专门法，但出口管制直接作用于算力这一项最大的成本。

因此本章不判断"哪里监管更严"，只分析"每一套规则通过哪一个变量影响模型层的生意"。 这是本章与一般政策综述的区别。

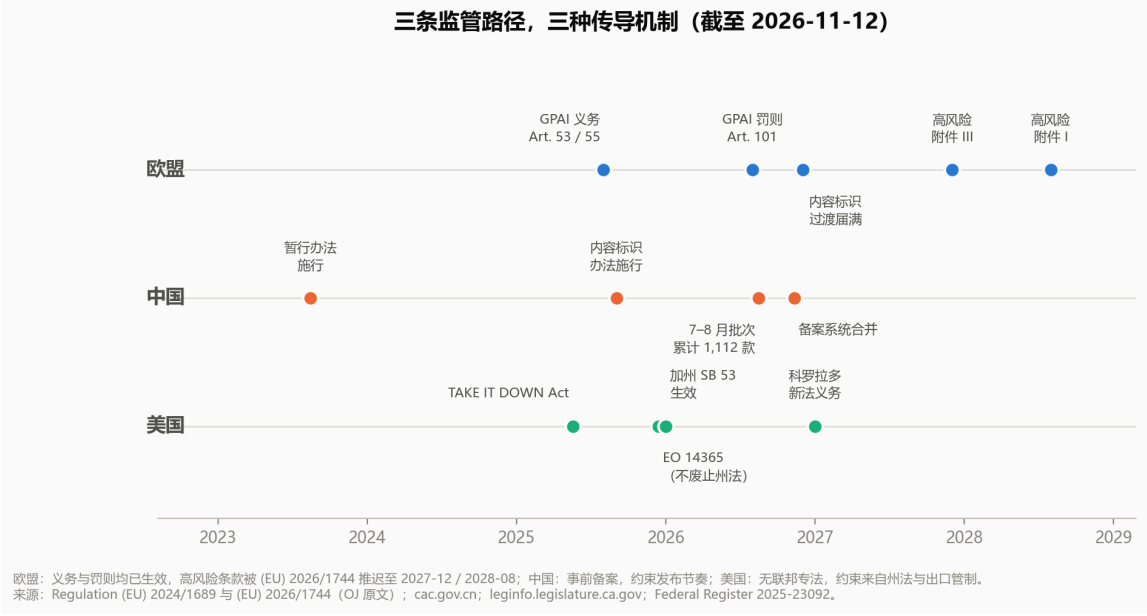


图 7–1

图 **7-1** 的读法： 三条泳道的时间尺度不同。欧盟是按日期逐条生效（可精确到某一天），中国是按批次发牌（只能读到累计数），美国是按执法与管制动作（没有统一生效日）。同一条横轴上放三种节奏，本身就是结论：欧盟的时间表最可预测，美国的最不可预测，而不可预测性本身是成本。

7.2 欧盟：GPAI 义务已生效，高风险条款被推迟两年

7.2.1 一个必须先纠正的时间表错误

2024 年原文《人工智能法案》（Regulation (EU) 2024/1689）规定高风险义务自 2026 年 8 月 2 日起适用。这一条已经被改掉了。

Regulation (EU) 2026/1744（**"Digital Omnibus on AI"**），2026 年 7 月 8 日通过，刊登于 OJ L 系列 2026/1744（2026 年 7 月 24 日）。本报告核对了该条例对第 113 条第 3 款的修订原文：

> "(40) in Article 113, the third paragraph is amended as follows: (a) point (a) is replaced by the following…; (b) point (c) is replaced by the following: '(c) Chapter III, Sections 1, 2, and 3, with the exception of Article 6(5), shall apply from: (i) **2 December 2027** as regards AI systems classified as high-risk pursuant to Article 6(2) and Annex III; and (ii) **2 August 2028** as regards AI systems classified as high-risk pursuant to Article 6(1) and Annex I;'; (c) the following point is added: '(d) Articles 102 to 110 shall apply from 27 July 2026.'"

来源：Publications Office of the European Union，CELEX 32026R1744（一手）。

注意这次修订改了什么、没改什么： 它只替换了 (a)、(c) 两项，并新增 (d) 项。第 **3** 款的 **(b)** 项——即 **Chapter V**（通用目的 **AI** 模型，**GPAI**）自 **2025** 年 **8** 月 **2** 日起适用、但 **Article 101** 除外——原封未动。

表 7-1 欧盟 AI 法案关键义务的时间表（截至 2026-11-12）

义务	条款	适用日期	状态
禁止性实践	Art. 5	2025-02-02	已适用
新增两项禁止（非自愿私密影像、CSAM 生成）	Art. 5(1)(ba)(bb) 、 Art. 5(1a)(1b)	**2026-12-02**	未适用
GPAI 模型义务（技术文档、下游提供者信息、版权政策、训练数据摘要）	**Art. 53**	**2025-08-02**	已适用
系统性风险 **GPAI** 附加义务（模型评估与对抗测试、风险缓解、严重事件报告、网络安全）	**Art. 55**	**2025-08-02**	已适用
GPAI 罚则	**Art. 101**	**2026-08-02**	已适用
其他义务罚则（含 Art. 50 透明度）	Art. 99(4)	2026-08-02	已适用
机器可读标识（2026-08-02 前已上市系统过渡期）	Art. 50(2)、新增第 4 款	**2026-12-02**	未适用
高风险：附件 III 独立系统	Chapter III §1–3	**2027-12-02**（原 2026-08-02）	未适用
高风险：附件 I 嵌入产品	Chapter III §1–3	**2028-08-02**（原 2027-08-02）	未适用
市场监督等治理条款	Art. 102–110	2026-07-27	已适用
2025-08-02 前已投放市场的 GPAI 过渡	Art. 111(3)	2027-08-02	未适用

来源：Regulation (EU) 2024/1689 第 113 条；Regulation (EU) 2026/1744 修订条款（均为一手，本报告核对原文）。

7.2.2 系统性风险的认定门槛：一个常被引错的数字

条例第 51(2) 条规定，累计训练算力超过 **10^{25} FLOPs** 的 **GPAI** 模型被推定为具有高影响力，提供者须在两周内通知委员会（Art. 52(1)），并可举证反驳（Art. 52(2)）。

流传更广的**"10^{23} FLOP** 是 **GPAI** 认定门槛**"**不在条例条文里——它出自欧盟委员会的 GPAI 范围界定指南。两者的法律效力不同：**10^{25}** 是条例的推定门槛，**10^{23}** 是指南的操作口径。 报告引用时必须写明是哪一层来源。

7.2.3 对模型层的经济含义：罚则上限是可计算的

Art. 101 的罚则上限为全球年营业额的 **3%** 或 **1,500** 万欧元，取较高者。以 Anthropic 滚动十二个月收入 352.82 亿美元计（A-12），3% 约为 **10.6** 亿美元。

但这是上限，不是预期成本。它只在最严重的违规情形下适用，且罚款需经委员会决定。对一家年收入 **350** 亿美元的公司，这个上限不构成经营约束。真正构成约束的是 Art. 53 的两项要求：

1. 训练数据摘要（training data summary）——公开到足以让"具有合法利益的各方"理解模型是如何训练的。这一条的合规成本不在罚款，在于它可能迫使公司披露训练数据的构成。
2. 版权政策——须制定并遵守欧盟版权法下的退出机制（opt-out）政策。这是对训练数据可得性的直接约束。

> 本报告的判断：欧盟监管对模型层的真实成本不在罚则，在训练数据的透明度义务。罚则上限（约 10.6 亿美元）相对于收入可以忽略；而训练数据摘要可能改变的是"能不能用这批数据"这一根本前提。市场讨论欧盟监管时普遍聚焦罚款金额，这是抓错了变量。

7.3 中国：备案制的实际形态

7.3.1 两套统计，不可相加

中国的生成式 AI 服务实行备案制（面向境内公众提供服务）与登记制（仅调用已备案模型能力的应用/功能）。这是两套独立统计，任何把它们相加的说法都是口径错误。

表 7-2 生成式 AI 服务备案与登记累计数（截至 2026-08-31）

项目	2026 年 7–8 月新增	累计
备案（模型/服务）	124 款（含 7 款手机端侧）	**1,112** 款
登记（调用已备案模型的 App/功能）	133 款	**731** 款

来源：国家互联网信息办公室《关于发布生成式人工智能服务已备案信息的公告（2026年7月至8月）》，2026-09-14 发布（一手）。

两条口径提醒：

1. 官方不按"第 **N** 批"编号，而是按双月区间命名。官网汇总页现有 13 期附件，覆盖 2024-04 至 2026-08。引用时须写"**2026** 年 **7–8** 月批次"，不能写"第 **13** 批"。
2. 第三方聚合站流传的"1,991 / 2,074 条备案记录"与官方 1,112 口径不同，不可混用。
3. **9–10** 月批次截至 **2026-11-12** 尚未公告，上述累计数随时会跳变。

7.3.2 备案制真正约束的是什么

备案制没有罚则驱动的合规成本，它的约束是时间：一个模型版本从训练完成到可以面向公众提供服务之间，需要一个审批周期。

这条约束的重要性，取决于收入函数对版本更新的敏感度——而这一点恰好有实测数据：

> Anthropic 披露，**2026** 年 **3** 月至 **9** 月发布的模型版本贡献了 **Q3** 收入的 **80%** 以上（A-29）。半年前的版本只贡献不到 **20%**。

把这两件事放在一起，结论是直接的：在一个收入高度依赖最新版本的行业里，上线审批的周期长度直接等于收入的折损。这不是中国的特殊问题——欧盟的合规审查、美国的州级审查都有同样的机制，只是工具不同。

7.3.3 与备案配套的三项制度

制度	文号 / 时点	核心内容	对模型层的影响
《生成式人工智能服务管理暂行办法》	七部门令第 15 号，2023-07-10 公布，**2023-08-15** 施行	第 2 条适用于向境内公众提供生成文本/图片/音频/视频的服务（内部研发不对外提供的不适用）；第 7 条训练数据合法来源与知识产权；第 17 条具舆论属性或社会动员能力的须安全评估 + 算法备案	准入与训练数据合规
《人工智能生成合成内容标识办法》	国信办通字〔2025〕2 号，**2025-09-01** 施行	第 3 条显式标识（用户可感知）+ 隐式标识（元数据）；第 5 条元数据须含生成合成属性、提供者名称/编码、内容编号；第 6 条传播平台核验元数据；禁止恶意删除/篡改标识	推理侧成本（每次生成须写入元数据）
算法/深度合成/区块链备案系统合并	2026-11-09 起	四类备案合并进 `beian.cac.gov.cn`，原系统同日停用	行政流程，非经营性

来源：gov.cn 令第 15 号、国信办通字〔2025〕2 号、cac.gov.cn 2026-11-09 通告（均为一手）。

内容标识办法的经济含义值得单独指出：它把一项推理侧的固定成本强加给了每一次内容生成。这是所有监管工具中唯一直接作用于边际成本的一项——其他工具作用于准入、数据或罚则，只有标识办法作用于"每生成一次就要多做一件事"。

7.3.4 在途的出口管制：状态容易写混

商务部、科技部 **2026** 年第 **48** 号公告（2026-10-16）拟将 AI 领域 **23** 项技术纳入禁止/限制出口目录，意见征询至 **2026-11-15**。

这里有两个状态必须分开：- 目录本身尚未生效（仍在征求意见）；- 对目录所列技术的"临时登记"已即时生效（ecomp.mofcom.gov.cn）。

把这两者写成"已实施出口管制"是错的。截至 2026-11-12，正确表述是"拟纳入、临时登记已生效、正式目录未定"。

一项未从原文核实的边界（须标注）：据商务部答记者问的转述，境外本地化部署受控模型属出口管制，而 **API** 调用、**App**/网页访问、**SaaS**、轻量端侧部署不属。这条对"中国模型出海"的判断影响很大，但本报告只见到二手转述，未见答记者问原文。引用时须写明"据商务部答记者问"。

7.4 美国：没有联邦专法，约束来自三层

表 7-3 美国 AI 监管的实际结构（截至 2026-11-12）

层级	状态	依据
联邦综合立法	无。第 119 届国会未通过综合性 AI 治理框架	立法检索（二手汇总，未见 enacted 综合法案）
联邦专门法	《TAKE IT DOWN Act》Public Law 119-12（2025-05-19 签署）：NCII/深度伪造刑事化 + 平台 48 小时下架	govinfo PLAW-119publ12（一手）
行政令	**EO 14365**《Ensuring a National Policy Framework for AI》，2025-12-11 签署，90 FR 58499	Federal Register 2025-23092（一手）
EO 的实际效力	设司法部 **AI** 诉讼工作组（30 天内）挑战州法；商务部 90 天内评估州法；以联邦拨款为条件。**EO** 本身不废止州法	同上
加州 **SB 53**（前沿 AI 透明度）	Chapter 138, Statutes of 2025，2025-09-29 签署，**2026-01-01** 生效。适用门槛：训练算力 **>10^{26}** 整数或浮点运算，或"大型前沿开发者"上年营收 **>5** 亿美元	leginfo（一手，条文已核）
加州 **SB 1047**	2024-09-29 被州长否决，未生效	公开记录
科罗拉多	SB 24-205 已被 **SB 26-189** 废止并重立（2026-05-14 签署）；新法义务自 **2027-01-01** 起，AG 规则亦须 2027-01-01 前制定	leg.colorado.gov（一手）

一个必须标注的不确定： 科罗拉多 SB 24-205 原定 2026-06-30 生效，2026-04 有联邦法院裁定暂停执行（xAI 诉讼、DOJ 加入）。这一诉讼细节本报告只见二手报道，未见法院文书。 正确表述是："SB 24-205 已于 2026-05-14 被 SB 26-189 取代；其间是否有联邦法院禁执令，未核实。"

7.4.1 碎片化本身就是成本

美国的特殊之处不是"没有监管"，而是"监管来自 **50** 个州且互相冲突"。

EO 14365 设立诉讼工作组挑战州法，说明联邦政府自己认为州法构成了负担。但 EO 不废止州法，白宫 2026-03-20 的立法建议也未成法。结果是：州法仍可执行，且处于诉讼中。

对模型层的含义是具体的：一家公司要面向全美提供服务，实际适用的标准是"最严的那个州"。 加州的 10^{26} FLOP 与 5 亿美元营收双门槛、科罗拉多的自动化决策文档义务，构成事实上的全国基线——不是因为它们有全国效力，而是因为合规成本不允许按州分版本。

7.5 其他主要辖区：一句话状态

表 7-4 其他辖区（截至 2026-11）

辖区	状态	依据级别
英国	无 **AI** 专法，且短期不会有：2026-05-13 国王演讲未提 AI 法案，原定针对前沿模型的立法被搁置；改推《Regulating for Growth Bill》（监管沙盒 + 增长义务），该法案尚未提交议会	二手（多家律所一致）
日本	《**AI** 推进法》 2025-06-04 公布、**2025-09-01** 全面施行；促进型基本法，无禁止条款、无罚则	二手（未查 e-gov 原文）
韩国	《**AI** 基本法》**2026-01-22** 施行，亚洲首部综合性 AI 框架法；有透明度/标注、高影响 AI 义务、境外提供者国内代表人；罚则上限 3,000 万韩元且有一年缓期（约 2027-01 起执行）	law.go.kr（一手法条页）
新加坡	无综合性 **AI** 法、无 **AI** 专门监管机构；自愿性框架路线，2026-01 发布《Model AI Governance Framework for Agentic AI》；约束层落在部门监管（MAS/IMDA/MOH）	二手（MDDI 答复、律所综述）

这张表的读法不是"谁先进"，而是"合规投入该往哪里放"。 有约束力的只有欧盟（义务已生效、罚则已生效）与中国（备案已运行）；英国、日本、新加坡目前不产生新增合规义务；韩国刚生效且罚则低、有缓期。

7.6 监管如何进入模型层的财务模型

把 7.2–7.5 的内容折算成财务变量，只有五项传导路径：

表 7-5 监管工具 → 财务变量 → 可观察指标

监管工具	作用的财务变量	传导强度	可观察指标
备案/上线审批（中国）、合规审查（欧盟）	收入（通过发布节奏）	强	新模型从训练完成到公开可用的间隔
训练数据摘要 / 版权 **opt-out**（欧盟 Art. 53）	训练成本与数据可得性	强	训练数据来源披露的完整度；数据授权支出
算力出口管制（美国）	供给与单位算力成本	中（对非美主体）	算力采购单价；交付延迟
内容标识元数据（中国）	推理侧边际成本	弱	单次生成的计算开销
罚则（欧盟 Art. 101、韩国）	尾部风险	弱	无（未发生即为零）

这张表的核心结论：监管对模型层影响最大的是"发布节奏"，不是"罚款"。

理由已经在 7.3.2 给出：半年内发布的版本贡献 **80%** 以上收入（**A-29**）。 这意味着收入函数对"新版本何时可用"的弹性极高，而对罚则的弹性接近于零（罚则上限约为收入的 3%，且需经完整行政程序）。

这解释了一个看起来矛盾的现象：模型公司对欧盟罚则的反应平淡，对训练数据披露和上线审批的反应激烈。前者是尾部成本，后者是收入函数的分母。

7.7 中国模型层：结构、定价与资本化

> 按本报告的覆盖边界（用户选择 **A**）：中国模型层单列一章，不并入全球模型层的估值框架，也不对中国个股做估值。

7.7.1 市场规模：两个都成立但互相矛盾的口径

IDC《中国 AI 软件市场半年度追踪，26H1》（2026 年 10 月下旬发布）给出的核心数字：

表 7-6 中国 MaaS 市场（2026 年上半年）

指标	数值
MaaS 市场规模	106.3 亿元人民币，同比 **+889.4%**
占公有云 PaaS 比重	5.2% → **21.0%**
Token 消耗	2,857.9 万亿，同比 +432.5%
模态结构（收入）	大语言模型 47.0%、视频生成 32.7%、其他 20.2%

份额有两个口径，结论完全不同：

口径	第一名	份额
LLM MaaS 细分收入	阿里云	**46.6%**（智谱 10.6%、火山引擎 10.2%、百度 6.7%、科大讯飞 3.5%）
整体 **MaaS** 收入（全模态）	火山引擎	**45.1%**（阿里云 25.2%、智谱 5.0%）

同一天的多篇报道出现过"阿里第一"与"火山第一"两个标题。这不是打架，是两个口径。 前者只算大语言模型，后者含视频生成（火山靠 Seedance）——视频生成占 **MaaS** 收入的 **32.7%**，足以颠覆 **LLM** 口径下的排名。

> 来源级别提示：**IDC** 原报告在付费墙后，本报告只取到媒体转述（**5** 家以上媒体数字一致），未见 **idc.com** 新闻稿原文。按权威二手处理。 引用时必须写明口径。

7.7.2 定价：七家厂商的旗舰模型（2026-11-12）

表 7-7 中国主要模型厂商 API 定价（人民币 / 百万 token）

厂商	2026 旗舰模型	输入	输出	缓存命中	备注
阿里通义千问	qwen3.8-max（1M 上下文）	12	36	—	夜间 22:00–08:00 四折；Batch 五折
阿里（预览）	qwen4-max-preview	无按量价	无按量价	—	仅订阅制（39/79/139/312 元/月），预览期 11-11 至 11-25
字节豆包	doubao-seed-2.1-pro	6.00	30.00	1.20	输入长度档 [0,1024]K
字节豆包	doubao-seed-2.1-turbo	3.00	15.00	0.60	[0,256]K
DeepSeek	deepseek-v4-pro（1M）	4.5 / 9.0	13.5 / 27.0	0.15 / 0.30	空闲/高峰价；高峰为工作日 9–12、14–18 时
DeepSeek	deepseek-flash	1 / 2	4 / 8	0.02 / 0.04	空闲价为高峰价一半
智谱	GLM-5.3（1M）	8	28	2	缓存存储限时免费
智谱	GLM-5.3-Flash	0.8	2.8	0.23	多模态
月之暗面	kimi-k3.1-pro（1,048,576）	13.00	68.00	1.30	Batch 为标准价 60%

厂商	2026 旗舰模型	输入	输出	缓存命中	备注
月之暗面	kimi-k3	20.00	100.00	2.00	—
MiniMax	MiniMax-M3（≤512K）	4.20（永久五折 **2.10**）	16.80（**8.40**）	0.84（0.42）	五折为长期政策
MiniMax	MiniMax-M3.1-Flash	1.00	4.00	0.10	—
腾讯混元	Hy4	6	18	0.3	**2026-12-31** 前按目录价 **50%**（实付 3/9/0.15）
腾讯混元	Hy3	1	4	0.25	—

来源：各家官方定价页（阿里云百炼、火山方舟、DeepSeek API 文档、智谱开放平台、Kimi 开放平台、MiniMax 开放平台、腾讯云 TokenHub），查询日 2026-11-12。

三条必须写进正文的口径提醒：

1. **qwen4-max** 没有 **token** 单价。 它不是"很贵"，是不存在按量计价的报价——预览期只走 Token Plan 订阅（含 20k–200k Credits），且预览期于 2026-11-25 结束后下线。报告若写**"qwen4-max** 定价 **X** 元/百万 **token"**，即为编造。
2. 腾讯混元已迁至 **TokenHub**。 旧文档页（最高只到 Hunyuan-a13b 等小模型）仍在可检索范围内，引用旧页会漏掉 **Hy4** 这一代旗舰。
3. 促销价与目录价必须分开。 MiniMax 的"永久五折"是厂商措辞而非合同承诺；腾讯 Hy4 的五折明确止于 2026-12-31；阿里夜间四折是时段定价。用促销价做横向比较会系统性低估中国市场的实际价格水平。

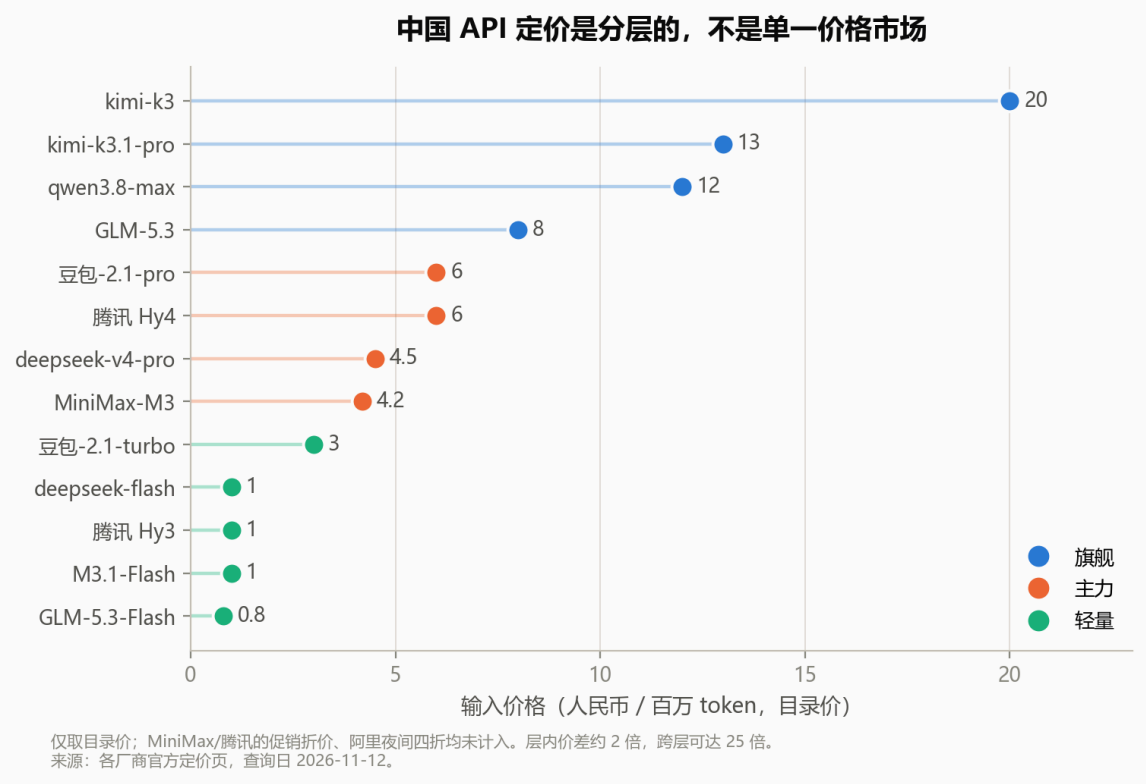


图 7–2

图 **7-2** 只画目录价，促销与时段价不进图——否则同一家厂商会因为在不同时间被查到而出现两个点。图上每一层的宽度就是这一层内部的最大价差，层与层之间的空档就是"能力分层"的价格表现。

7.7.3 一个可观察的结构性事实：价格带

把上表按"输入价"排序，可以看到一个清晰的价格带结构：

- 旗舰层（输入 8–20 元）：kimi-k3（20）、kimi-k3.1-pro（13）、qwen3.8-max（12）、GLM-5.3（8）；
- 主力层（输入 4–6 元）：豆包-2.1-pro（6）、Hy4（6）、deepseek-v4-pro（4.5 低谷）、MiniMax-M3（4.2）；
- 轻量层（输入 ≤3 元）：豆包-turbo（3）、deepseek-flash（1）、Hy3（1）、M3.1-Flash（1）、GLM-5.3-Flash（0.8）。

相邻层的价差约为 **2** 倍，跨层价差可达 **25** 倍。 这个结构与海外一致：模型层不是单一价格市场，是按能力分层的多个市场。因此"中国 **API** 价格是海外的几分之一"这类笼统说法没有意义——必须指定层。 本报告不做跨币种的价格比较，因为缺少同层同能力的配对。

7.7.4 资本化：从叙事驱动转向 ARR 验证

表 7-8 2026 年中国模型层主要资本事件

时间	主体	事件	数字	来源级别
2026-01-08	智谱	港股上市（"大模型第一股"）	发行价 11.6 港元，首日市值 579 亿港元	权威二手
2026-01-09	MiniMax	港股上市	首日 +109%，市值超 1,000 亿港元	权威二手
2026-05	月之暗面	D 轮	约 20 亿美元，投后 >200 亿美元	二手
2026-06	DeepSeek	首轮外部融资	>70 亿美元	二手
2026-09	智谱	配售 + 可转债	20 亿 + 30 亿美元	二手
2026-10-06	月之暗面	Pre-IPO 轮	估值约 **500** 亿美元，拟 2027Q1 港股 IPO	彭博社，公司未确认
2026-10-06	DeepSeek	新一轮	至少 **800** 亿元，估值约 **5,000** 亿元	彭博社，公司未确认
2026-11-10	智谱	二级市场	收盘约 482 港元、市值约 2,350 亿港元，较 6 月高点回撤约 78%	二手（行情可复核）

> 不要把彭博社的 **500** 亿美元 / **5,000** 亿元估值当作已确认事实。 两家公司均未公告。本报告引用它们时，只用于说明"一级市场存在这个量级的报价"，不用于任何倍数计算。

可以写进报告的稳健表述： 2026 年中国模型层完成了"上市 + 一级市场向头部集中"两件事。智谱从 **6** 月高点到 **11** 月 **10** 日回撤约 **78%**，是"估值从叙事驱动转向 **ARR** 验证"的直接证据。

7.7.5 与海外的四条结构性差异

1. 分发依赖自有流量入口（阿里云、抖音、微信、浏览器），而非企业直销。后果：**C** 端渗透快，**B** 端付费意愿弱。
2. 开源权重 + 国产算力降低了进入门槛，价格战比海外更激烈。后果：单位价格更低，但难以判断是效率优势还是补贴。
3. 备案制构成事前准入，发布节奏受行政审批约束（7.3.2）。
4. 收入规模与盈利状况不可直接与海外比较：中国厂商多不以模型 API 为独立核算主体，MaaS 收入混在云业务中，缺少可比的独立损益。

> 这四条合起来的意思是：中国模型层的竞争强度可能高于海外，但可观测性远低于海外。 本报告因此不做中外模型层的估值对比——缺少可比口径的对比，比不做对比更危险。

7.8 本章结论

1. 三种监管范式管的不是同一件事：中国管准入（发布节奏）、欧盟管义务（合规成本与罚则）、美国管供给与州法碎片。用"松紧"刻度比较是错的。

2. 欧盟高风险条款已被 **Regulation (EU) 2026/1744** 推迟至 **2027-12-02 / 2028-08-02**，但 **GPAI** 义务未被推迟。 按 2024 年原文写时间表会错。10^{25} FLOP 是条例门槛，10^{23} 是指南口径，不能混。
3. 欧盟监管的真实成本不在罚则（上限约收入的 3%），在训练数据摘要与版权 **opt-out** 义务——它作用于"能不能用这批数据"这一根本前提。
4. 中国备案制不罚款，它约束时间；而收入函数对版本更新的弹性极高（半年内版本贡献 80% 收入），因此审批周期直接等于收入折损。
5. 监管影响模型层的主通道是收入（通过发布节奏），不是成本（通过罚则）。 市场讨论监管时聚焦罚款金额，是抓错了变量。
6. 中国 **MaaS** 的两个份额口径（阿里 **46.6% /** 火山 **45.1%**）不可混用；差异来自视频生成占 MaaS 收入 32.7%。
7. 中国模型层竞争强度可能高于海外，但可观测性远低于海外——MaaS 收入混在云业务中，缺少独立损益，本报告因此不做中外估值对比。

7.9 本章的局限

项目	状态	影响
欧盟 2026/1744 之后是否还有新修订	未定（EUR-Lex 当前版本日期 2026-07-27；本报告核对的是 OJ 原文）	时间表可能再变，须标注"截至 2026-11-12"
IDC 报告原文	未取得（付费墙，仅有媒体转述）	106.3 亿元 / 46.6% / 45.1% 三个数按权威二手处理
商务部 2026 年第 48 号公告附件	未取得（页面只写"AI 领域 23 项技术"，未列门槛）	流传的"10^{26} FLOPs / 2 万亿参数"来自二手解读，未从原文核实
科罗拉多 SB 24-205 的联邦禁执令	只见二手报道，未见法院文书	该州状态按"已被 SB 26-189 取代"表述，禁执令列为未核实
日本《AI 推进法》条文	未打开 **e-gov** 原文，施行日来自律所综述	表 7-4 中日本一行标为二手
商务部答记者问（API 调用不属出口管制）	未见原文	7.3.4 的边界判断须标注转述来源
模型层的独立损益（中国）	不存在（MaaS 收入混于云业务）	无法做中外估值对比
中国 9–10 月备案批次	截至 **11-12** 未公告	表 7-2 累计数会跳变

第八章 估值

> 本章不提供评级、不设定目标价、不预测股价。 估值分析的目的是回答：当前定价隐含了什么假设，这些假设在什么条件下不成立。 所有计算过程见 `scripts/dcf.py`，输出见 `evidence/dcf.json`。

8.1 为什么模型层难估值：三个结构性障碍

障碍一：没有稳态可比。 传统估值依赖"成熟期的利润率"作为终值锚。模型层的终值利润率取决于三个未定的变量：算力成本的长期曲线、模型层的竞争结构、以及它与云厂商的利润分配比例。三者都没有历史数据可供外推。

障碍二：没有历史。 Anthropic 的商业化始于 2023 年一季度，到 2026 年三季度只有三年半的数据。任何基于历史增长率的推断，样本都不足一个完整的技术周期。

障碍三：关联交易定价。 44% 的收入经云厂商市场成交，云厂商同时是股东与算力供应商（A-19、A-20、A-22）。收入、成本、毛利率三个关键参数，都部分由交易对手方决定。 这使得任何基于报表比率的估值，都隐含了"当前定价关系会延续"的假设。

> 本报告的应对方式：不追求"算出内在价值"，而是做"约束检验"——用三组互相独立的信息（市场价格、公司承诺、行业上限）互相校验，找出它们不兼容的地方。不兼容处就是风险所在。

8.2 定价区间隐含的倍数

按 S-1 披露的 70–78 美元区间、发行后 130.87 亿股、备考调整后现金 2,126.84 亿美元、可转债 1,108.73 亿美元计算（A-39、A-40、A-41）：

表 8-1 IPO 定价区间的倍数（2026-11-12）

指标	$70	$74（中值）	$78
发行后市值（十亿美元）	916.1	968.4	1,020.8
企业价值 EV（十亿美元）	814.3	866.6	918.9
EV / 滚动十二个月收入（352.82 亿）	23.1×	**24.6×**	26.1×
EV / Q3 年化收入（680 亿）	12.0×	**12.7×**	13.5×
EV / 前三季度年化收入（443.23 亿）	18.4×	**19.6×**	20.7×
EV / 10-31 run-rate（810 亿，口径见 2.2）	10.1×	**10.7×**	11.3×
市值 / 滚动十二个月收入	26.0×	27.5×	28.9×
市值 / 前三季度年化调整后净利（74.55 亿）	122.9×	129.9×	136.9×
EV / 前三季度年化调整后经营利润（54.41 亿）	149.6×	159.3×	168.9×

数据来源：S-1/A（A-12、A-14、A-39、A-40、A-41）。倍数由本报告计算。

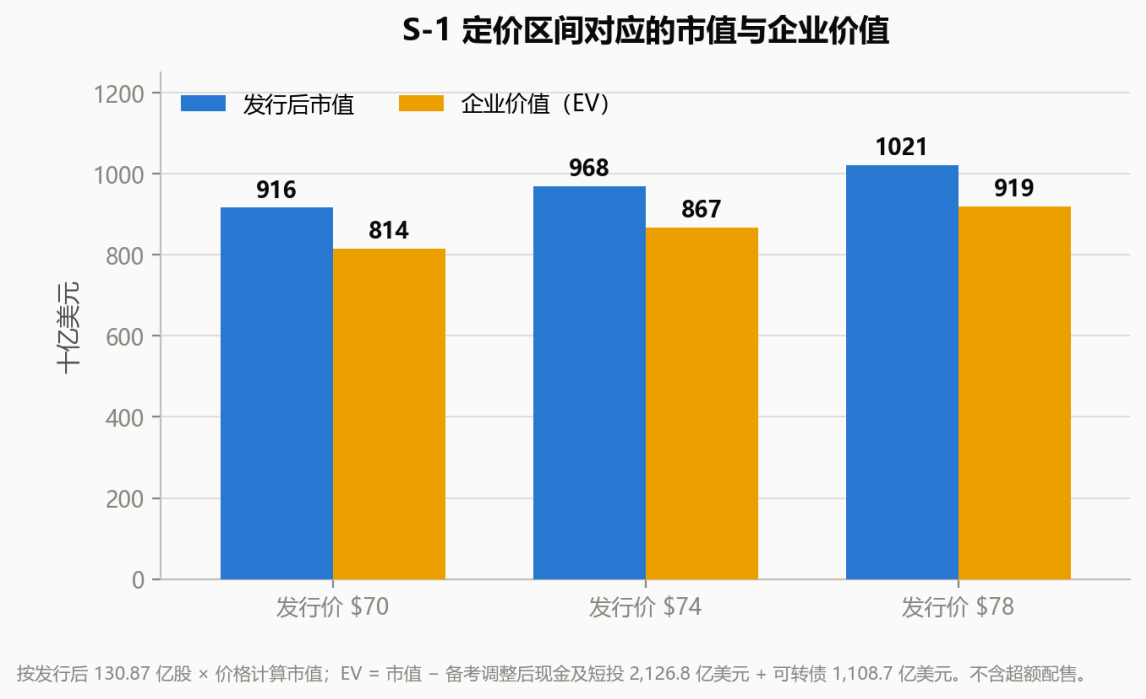


图 8-1

图 **8-1** 的读法：市值与企业价值相差 **1,018** 亿美元，全部是备考净现金（2,126.84 − 1,108.73，A-39），占中值市值的 **10.5%**。本报告对 **Anthropic** 自身的估值全程用 **EV** 而非市值（唯一例外是与 OpenAI 的对照，那里只能用市值，理由见 8.6.3），理由是这个差额来自 IPO 募资本身——用市值会把这笔刚募到、还没花的钱算进对业务的定价里。 但 EV 口径也有代价：它隐含假设这笔现金最终会以面值转化为产能或股东价值。对一家正在把现金投进算力与研发的公司，这个假设大体成立；但它确实是假设，不是事实。

这张表里最需要注意的是"用哪个收入做分母"。 四个分母给出四个完全不同的倍数：10.7 倍（10 月底 run-rate）、12.7 倍（Q3 年化）、19.6 倍（前三季度年化）、24.6 倍（滚动十二个月）。差异全部来自增长速度——当收入一年增长 **12** 倍时，任何历史口径都会严重低估当前规模。 run-rate 与 Q3 年化之间相差 19%（810 对 680），仅这一个月的用量增长就使倍数从 12.7 倍降到 10.7 倍。

> 本报告的取法：以 **Q3** 年化（**12.7** 倍）为主口径。 理由有二：一是它最接近"如果当前速度维持一年"的收入，与前瞻倍数的含义最接近；二是它是 GAAP 收入，可与 8.5 的可比公司倍数直接对照。公司自报的 **run-rate**（**10** 月底 **810** 亿）更近，但按 **S-1** 自己的定义是 **"28** 天按量费 × **13** + 当月订阅费 × **12"**，不是 **GAAP** 收入，因此本报告只把它作为并列口径——在 8.6 的对照里必须用它（因为对方也只有 run-rate）。必须同时提示：Q3 单季收入 170 亿美元中，有超过 80% 来自半年内发布的模型（A-29），这个收入基础本身是新的、未经完整季度周期检验的。

用调整后利润算出的 **150–169** 倍（**EV**/调整后经营利润），或按市值算的 **123–137** 倍（市值/调整后净利），本报告认为参考价值有限。 原因在第四章已论述：调整后利润加回了 32.55 亿美元 SBC 与 20.86 亿美元非现金捐赠（A-12）。这是"未扣除真实股东成本"的利润。 本报告列出它，是为了说明"即使按公司最有利的口径，估值仍是三位数倍数"，而不是为了用它做估值。

8.3 三重约束交叉检验

估值的问题不是"倍数高不高"，而是"什么条件下这个倍数成立"。本报告用三组互相独立的信息交叉检验：

约束一：市场价格隐含的增速（反向 DCF）

给定企业价值 866.6 亿美元，反解"前五年收入增速需多快才能支撑"。模型设定：2026 年收入 502.4 亿美元（前三季度实测 332.42 亿 + 第四季度按 Q3 的 170 亿持平），2027–2031 年恒定增速，2032–2036 年线性降至 8%，此后永续 3%；**FCF** 利润率以终值为锚按比例爬坡——2027 年为终值的 8%（终值 25% 时即 2%），此后依次 24%、40%、56%、72%、88%，第 7 年达到终值并维持。这组路径全部是选择，不是实测；表 8-2 展示结论对这些选择的敏感度。

表 8-2 隐含前五年收入增速（行：终值 FCF 利润率；列：WACC）

终值 FCF 利润率	WACC 8%	9%	**10%**	11%	12%
10%	41.1%	45.7%	50.0%	53.8%	57.5%
15%	33.1%	37.5%	**41.5%**	45.2%	48.7%
20%	27.7%	32.0%	35.8%	39.3%	42.6%
25%	23.7%	27.8%	**31.5%**	34.9%	38.1%
30%	20.4%	24.4%	28.0%	31.4%	34.5%

数据来源：本报告计算（`scripts/dcf.py`）。WACC 与终值 FCF 利润率均为选择，非实测。

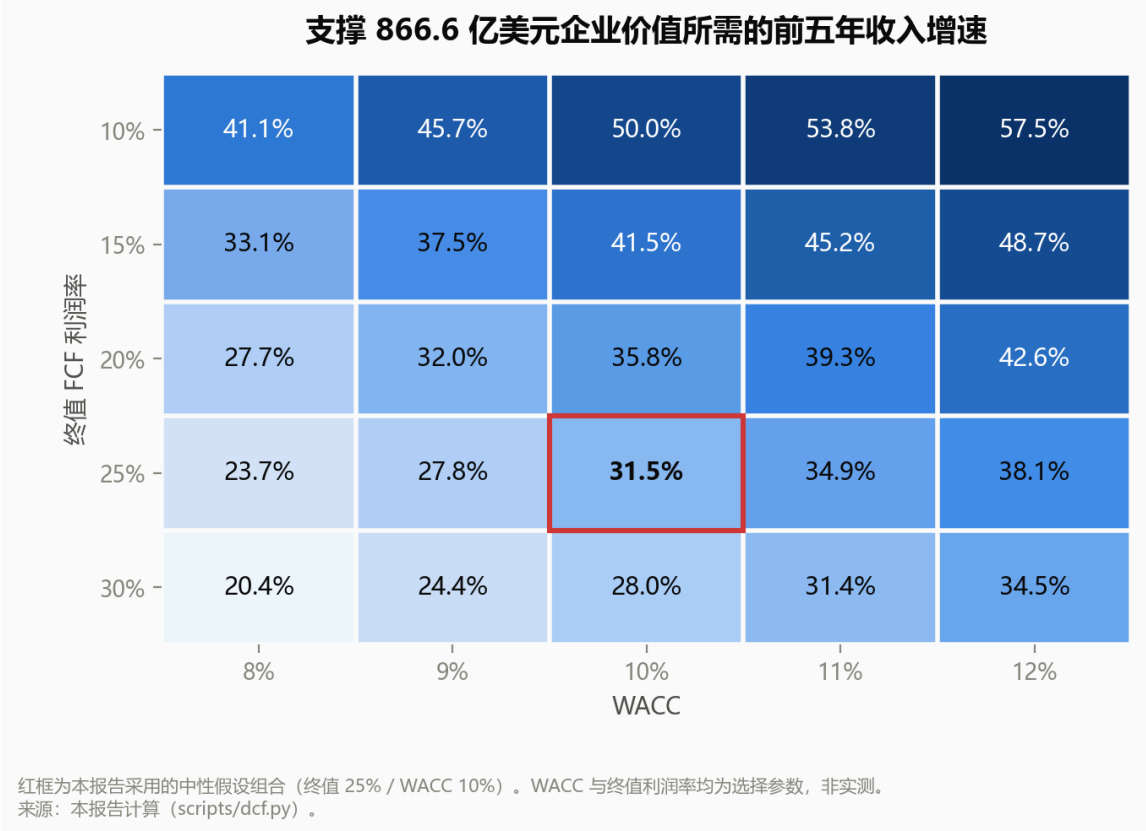


图 8-2

读法： 若假设模型层的稳态 FCF 利润率能达到软件行业的 25%、WACC 为 10%，则 866.6 亿美元的企业价值隐含五年 **31.5%** 的收入复合增速，2030 年收入约 1,500 亿美元，2036 年约 4,370 亿美元。

31.5% 是一个偏低的门槛。 公司当前增速是三位数，即使大幅减速，五年 31.5% 也不算苛刻。按这个口径，定价是保守的。

基期的敏感性。 502.4 亿的基期假设第四季度与第三季度持平，这是偏保守的一档：10 月底 run-rate 810 亿按定义折算的季度收入约 **202.5** 亿（2.2），比 Q3 高 19%。以此重解，2026 年收入约 **534.9** 亿，同一企业价值隐含的五年增速为 **30.3%**（基准假设下 31.5%），2030 年收入约 **1,540** 亿。门槛下移 **1.2** 个百分点，结论方向不变（`scripts/runrate.py`）。

但这个结论有一个致命的对手方——见约束二。

约束二：采购承诺要求的收入规模

9,009 亿美元采购承诺中，4,516 亿美元将在 **2030** 年底前发生（A-32，原文为 "before December 31, 2030"）。从 2026 年 9 月底到 2030 年底约 **4.25** 年，年均算力支出约 **1,063** 亿美元（推导；按均匀分摊，实际节奏前低后高）。

若算力占收入的比例维持在当前的 56.6%，则 2030 年收入需达到约 **1,880** 亿美元；若算力占比降至 45%，需 **2,360** 亿美元；降至 35%，需 **3,040** 亿美元。

注意方向：算力效率越高（占收入比重越低），同样的采购义务反而要求更多收入来覆盖。义务是固定的美元金额，分母越小，所需收入越大——这不是笔误，而是这份合同最反直觉的地方。

表 8-3 承诺要求的收入规模与所需增速

2030 年算力占收入	所需 2030 年收入（十亿美元）	CAGR（4.25 年），自 Q3 年化 680 亿起	CAGR（4.25 年），自 10-31 run-rate 810 亿起	收入需为 run-rate 的倍数
65%	163	22.9%	18.0%	2.0×
55%	193	27.9%	22.7%	2.4×
45%	236	34.0%	28.6%	2.9×
35%	304	42.2%	36.5%	3.7×

数据来源：承诺额来自 S-1/A（A-32）；收入要求、两组 CAGR 与倍数由本报告计算（`scripts/dcf.py`、`scripts/runrate.py`）。算力占收入的比例为选择，当前实测值 56.6%（A-07）。

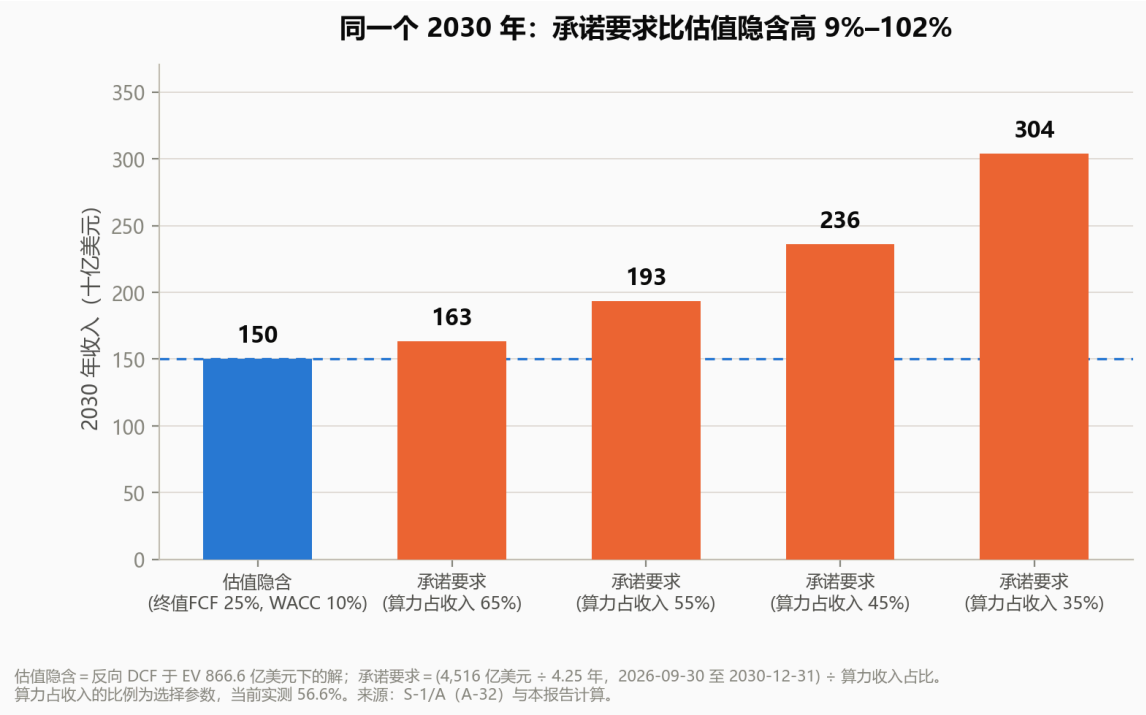


图 8-3

把约束一与约束二并排放，矛盾就出现了：

> 估值隐含：**2030** 年收入约 **1,500** 亿美元（**31.5%** 增速）。承诺要求：**2030** 年收入约 **1,880** 亿美元（**56.6%** 算力强度；自 **Q3** 年化起需 **27.0%** 增速，自 **run-rate** 起需 **21.9%**）；强度降至 **35%** 则需 **3,040** 亿美元（**42.2% / 36.5%**）。在最有利于估值的假设下（算力强度不再下降），承诺要求仍比定价隐含高约 **25%**；若强度按行业规律继续下降，差距扩大到约 **1** 倍。

> 换基期不改变这个矛盾。基数更高只会让所需增速更低（表 8-3 的第三列对第二列），但收入水平的差距（1,880 对 1,500、3,040 对 1,500）一点没变。

这不是会计问题，是两组假设不能同时成立：定价假定公司在 2030 年只做到 1,500 亿收入，承诺要求它在同一年做到 1,880–3,040 亿。三种可能的解释，每一种对应一种完全不同的投资含义：

解释	含义	可观察的验证信号
A. 承诺会被重新谈判或延期	大部分承诺（尤其是 2031–2045 年的 4,493 亿）在执行中调整；实际算力支出低于合同额	未来财报中"采购承诺"科目下调；算力费用增速持续低于承诺隐含节奏
B. 收入会超过估值隐含的水平	市场低估了增长，定价偏保守	2027 年收入超过 1,000 亿美元（即增速 >100% 后再维持一年）
C. 承诺会被履行，但现金流为负	收入增长但无法覆盖算力支出，需要持续融资，股权被稀释	经营性现金流转负；再次大规模股权或债务融资

本报告的判断：**A** 的可能性最高，**C** 的破坏性最大。理由：承诺的期限结构（4,493 亿美元在 2031–2045 年，跨越 15 年）本身说明它不是硬性的短期义务；S-1 也明确描述了多方可结构与融资方依赖，这些结构在执行中天然留有调整空间。但"最可能"不等于"无代价"——承诺的重新谈判意味着容量减少，而容量减少直接限制收入上限。换句话说，A 与 B 是互斥的：你不能既指望承诺被削减，又指望收入超过估值隐含的水平。

约束三：行业规模的上限

约束一与约束二的矛盾，最终要由"模型层这个行业能有多大"来裁决。这是本报告认为最需要外部锚的地方，而外部锚在这里恰好失效了——失效的方式本身就是一个结论。

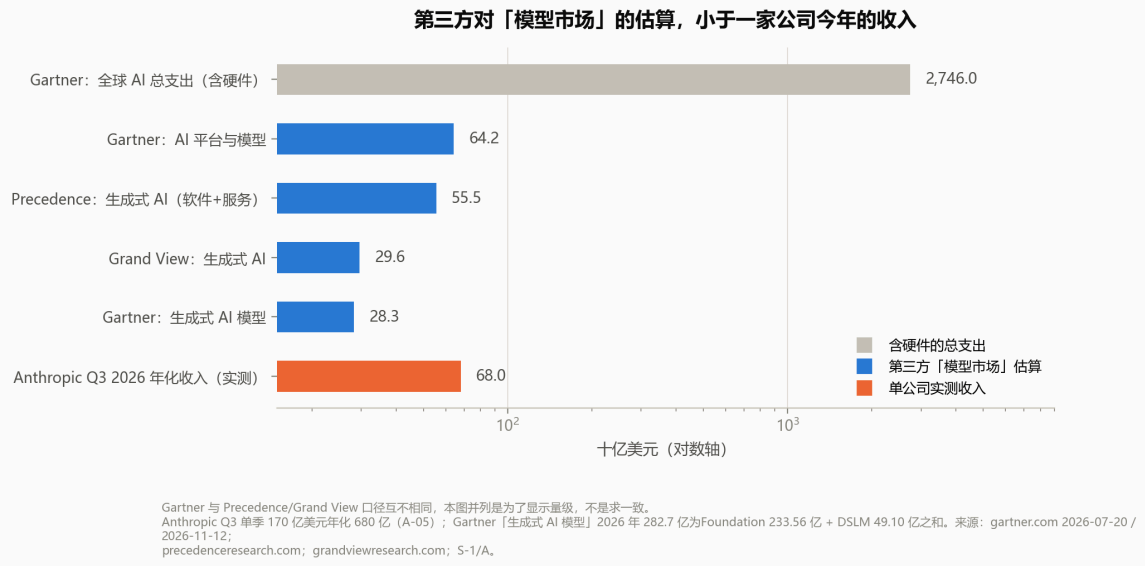


图 8–4

表 8–4 第三方对"AI 市场规模"的估算（2026 年，口径逐条不同）

机构	2026 年数值（十亿美元）	口径	发布日
Gartner	**2,746**	全球 AI 总支出，含服务器、芯片、网络、设备	2026-11-12
Gartner	**64.25**	"AI 平台与模型"（四个子段合计）	2026-07-20
Precedence Research	**55.51**	生成式 AI，软件 + 服务，不含硬件	2026-09-24
Grand View Research	**29.6**	生成式 AI，软件 + 服务	2026-06
Gartner	**28.27**	仅生成式 **AI** 模型（基础模型 233.56 亿 + 专用模型 49.10 亿）	2026-07-20
Anthropic（实测）	**68.0**	**Q3** 单季 **170** 亿美元年化	S-1/A

来源：gartner.com 三份新闻稿；precedenceresearch.com；grandviewresearch.com；S-1/A（A-05）。单位原为百万美元者已折算，口径照原文。

这张表最重要的不是"口径不同"，而是最后一行与倒数第二行的关系：

> **Anthropic** 第三季度单季收入年化（**680** 亿美元），是 **Gartner** 估算的整个"生成式 **AI** 模型"市场（**282.7** 亿美元）的 **2.4** 倍，也超过了 **Gartner**"**AI** 平台与模型"全部四个子段之和（**642.5** 亿美元）；即使按 **2026** 年全年收入估计 **502.4** 亿美元计，仍是前者的 **1.8** 倍、后者的 **78%**。

这在逻辑上只能有三种解释，每一种都影响估值方法：

解释	含义	对估值的影响
A. 口径不重叠	Gartner 的"模型"段不含经云市场转售的消耗、不含第一方应用订阅（Claude Code 等），而 Anthropic 的收入包含这些	第三方 **TAM** 无法作为模型层收入的上限，因为它测的不是同一件事
B. 估算已过时	2026 年的实际增速远超机构在年中做出的预测	所有基于 TAM 的长期预测都需要上修
C. 收入确认口径偏松	总额法（principal）确认使收入包含渠道环节的价值	收入与"终端用户支出"之间需要一个折算系数

本报告无法在 **A/B/C** 之间裁决——这需要 **Gartner** 的报告原文（付费墙）与 **Anthropic** 的分产品收入拆分（不存在于披露中）。

> 这个无法裁决本身，就是本章对约束三的结论：第三方市场规模数字不能用来检验估值，因为它与被检验对象的口径没有对齐。
>
> 因此本报告不采用任何 **TAM** 数字作为估值上限，也不采用任何"渗透率"推算。 8.5 节改用已上市公司在市场上的实际倍数作为外部锚——那是唯一与 Anthropic 处在同一计价体系里的信息。

一条必须写明的口径风险： 若解释 C 成立，则 Anthropic 的 680 亿美元年化收入中，有一部分在可比口径下应被折算掉。这会使本章所有倍数上升，第八章各情景的企业价值相应下调。 本报告在第四章已论证总额法的依据（S-1 明确写"the Company is the principal and recognizes revenue on a gross basis"），但不掌握折算系数。

8.4 敏感性：什么会推翻结论

表 8-5 企业价值对增长与利润率的敏感性（十亿美元，WACC 固定 10%）

前五年增速	终值 FCF 15%	22%	28%
20%	279	409	521
40%	803	1,178	1,500
60%	2,059	3,019	3,843

数据来源：本报告计算（`scripts/dcf.py`）。WACC 固定 10%，仅变动增速与终值利润率，以便单独观察两个变量的影响。

这张表说明：估值对增速的敏感度高于对利润率的敏感度，但没有压倒性差距。 增速从 20% 提到 40%，企业价值从 279 亿升至 1,178 亿（4.2 倍）；终值利润率从 15% 提到 28%，企业价值从 803 亿升至 1,500 亿（1.9 倍）。增长是第一变量，利润率是第二变量，但两者量级相同——这与市场只争论毛利率的倾向不符。

> 这与市场的直觉有偏差。 市场讨论模型层时最常争论"毛利率能不能维持"，但从 DCF 结构看，增速的杠杆约为利润率的 **2.2** 倍（4.2 倍 vs 1.9 倍）。原因是终值占比高达 66%–78%（见表 8-6），大部分价值来自 **2036** 年之后的永续假设，而永续假设由终值利润率决定；但近期现金流折现后，近期增速的权重更大。

表 8-6 三个情景的核心假设与结果

情景	前五年增速	终值 FCF 利润率	WACC	企业价值	2036 年收入	终值占比	相对中值定价
熊市	20%	15%	11%	235	228	66%	−73%
基准	40%	22%	10%	1,178	685	73%	+36%
牛市	60%	28%	9%	4,713	1,807	78%	+444%

数据来源：本报告计算。全部假设为选择。

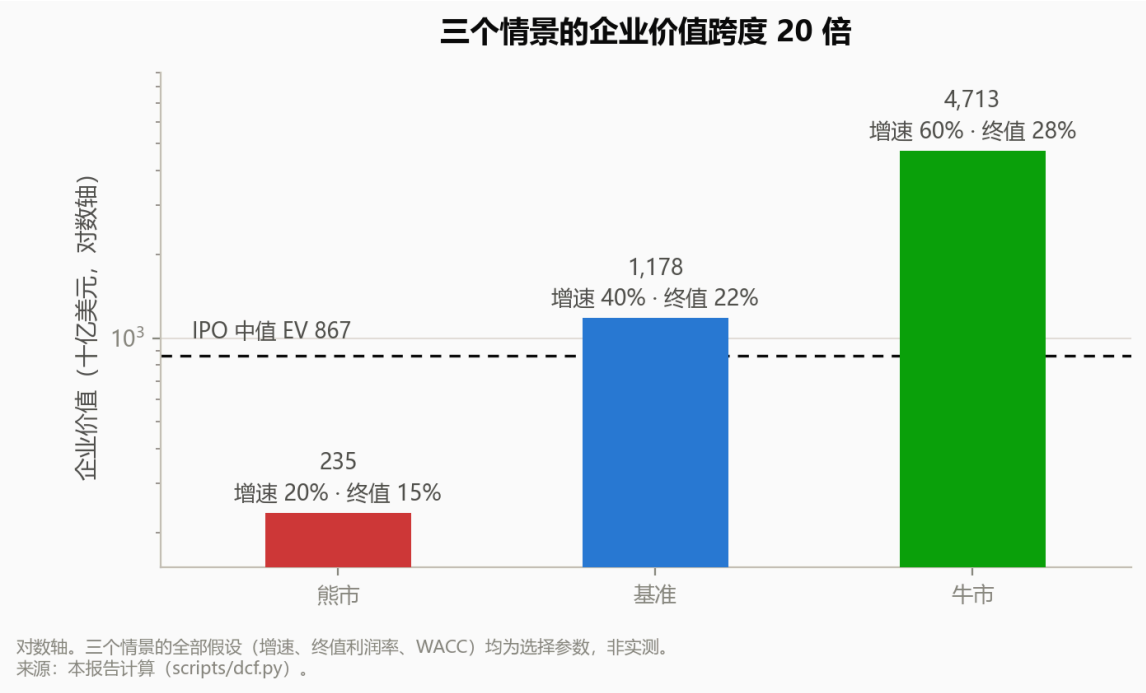


图 8-5

三个情景的跨度是 **20** 倍，这本身就是一个结论：在缺乏稳态可比、缺乏历史、且关键参数由关联交易决定的情况下，模型层的估值不是"算出来的"，是"假设出来的"。任何声称能给出精确内在价值的分析，都隐藏了未被检验的假设。

本报告愿意承担的判断是：基准情景（**40%** 五年增速、**22%** 终值 **FCF** 利润率）比熊市情景更可能。40% 的增速在 2030 年给出约 **1,930** 亿美元收入，恰好覆盖承诺在当前 56.6% 算力强度下的要求（1,880 亿）；若算力强度降到 35%，承诺需要 3,040 亿，基准情景就不够——按同样的四年口径，需要约 57% 的年均增速（4.25 年口径约 53%）。基准与熊市的分歧不在模型能力，而在算力强度与承诺可调整性这两个财务变量上。但牛市情景（60% 增速持续五年）需要模型层在 2036 年达到 1.8 万亿美元收入——这个量级超过当前全球软件行业的总规模，本报告认为概率很低。

> 这是一个可以用外部数据检验的判断。若第三方对 2030 年 AI 模型层市场规模的中性估算显著低于 1,500 亿美元，则基准情景也不成立；若接近或高于 1,880 亿美元（承诺在当前算力强度下的要求），则基准偏保守。第八章 8.5 节给出可比公司与行业规模的对照。

8.5 上市可比公司

8.5.1 一个必须先承认的事实：没有一家上市公司是"模型层"

8.3 节论证了第三方市场规模数字不能用来检验估值，因为它们与被检验对象的口径没有对齐。上市公司是唯一与 **Anthropic** 处在同一计价体系里的外部锚——它们和 Anthropic 一样，被同一批投资者用同一套倍数、在同一天定

价。

但可比公司的选取有一个无法回避的缺陷：市场上不存在纯粹的上市模型层公司。

主体	为什么不能直接当可比
OpenAI	未上市，无公开财务报表（见 8.6）
Anthropic	本章的被估值对象本身
xAI	2026-02 并入 SpaceX，财务进入混合口径的 "SI" 分部（第五章 5.2）
Mistral、Cohere 等	未上市

因此本节选的七家不是"同行业公司"，而是"与模型层有实质收入或成本关联的上市公司"。这个替代关系必须在读每一个倍数时都记住：它们衡量的是市场对 **AI** 相关现金流愿意付多少钱，不是市场对模型层愿意付多少钱。

8.5.2 口径与数据

样本七家：Microsoft、Alphabet、Amazon、Meta、NVIDIA、Oracle、Palantir。选取理由是它们分别对应模型层的四种关联：分发渠道（MSFT、GOOGL、AMZN）、自训模型（GOOGL、META）、算力供给（NVDA）、企业 AI 应用（PLTR、ORCL）。

口径定义：

- 股价：2026-11-12 收盘价（Yahoo Finance 前复权收盘，经浏览器 UA 与本地代理取得）。
- 股本：最近季度加权平均基本股数（SEC XBRL `WeightedAverageNumberOfSharesOutstandingBasic`）。
- 收入与经营利润：滚动十二个月（**TTM**），由 XBRL 单季数据还原后四季相加，而非直接取年报。这一步不可省——七家的财年结束日分布在 1 月（NVDA）、5 月（ORCL）、6 月（MSFT）与 12 月（其余四家），直接用年报会把横截面混入约半年的时点差；改用 TTM 后，七家的期末只相差两个月（NVDA 2026-07-26、ORCL 2026-08-31，其余 2026-09-30）。
- 企业价值 **EV** = 市值 + 有息负债 − 现金及短期投资。有息负债取最近资产负债表日的长期债务与短期债务之和；ORCL 的 `LongTermDebtNoncurrent` 在其 XBRL 中已停用，本报告改用 `LongTermNotesAndLoans` + `NotesPayableCurrent`（2026-08-31：1,177.1 + 76.2 = 1,253.3 亿美元）。
- 经营利润：GAAP `OperatingIncomeLoss`，不调整股权激励——这一点在 8.5.5 与 Anthropic 对比时会造成不对称，届时单独处理。

表 8-7 七家上市公司的倍数（2026-11-12）

公司	财务期末	市值	净现金	EV	TTM 收入	TTM 同比	经营利润率	EV/收入	EV/经营利润
MSFT	2026-09-30	3,870	+29.0	3,841	346.4	+17.9%	47.1%	11.1×	23.5×
GOOGL	2026-09-30	3,597	+139.0	3,458	471.1	+22.2%	33.3%	7.3×	22.0×
AMZN	2026-09-30	2,874	−65.7	2,940	797.5	+15.4%	13.1%	3.7×	28.2×
META	2026-09-30	1,605	−1.2	1,606	241.6	+27.5%	32.9%	6.6×	20.2×
NVDA	2026-07-26	4,203	+23.2	4,179	303.0	+83.4%	65.2%	13.8×	21.2×
ORCL	2026-08-31	381	−88.3	469	71.8	+21.6%	32.1%	6.5×	20.3×
PLTR	2026-09-30	415	+10.6	404	7.3	+87.1%	45.6%	55.5×	121.5×

单位：十亿美元。净现金 = 现金 + 短期投资 − 有息负债，负值为净负债。数据来源：SEC XBRL companyfacts（一手）+ Yahoo Finance 收盘价；倍数由本报告计算（`scripts/build_comps.py`、`scripts/comps_table.py`）。

8.5.3 第一个结果：市场在按经营利润率定价，不是按增速

把上表按 **EV/经营利润**排序，会看到一个反直觉的收敛：

> 七家里的六家落在 **20.2×–28.2×** 之间（标准差 **2.7**），跨越 **15%–83%** 的收入增速和 **13%–65%** 的经营利润率。
>
> 唯一的例外是 **Palantir，121.5×**。

这个收敛可以用一个恒等式解释——它不是发现，是算术：

EV/收入 = (EV/经营利润) × 经营利润率

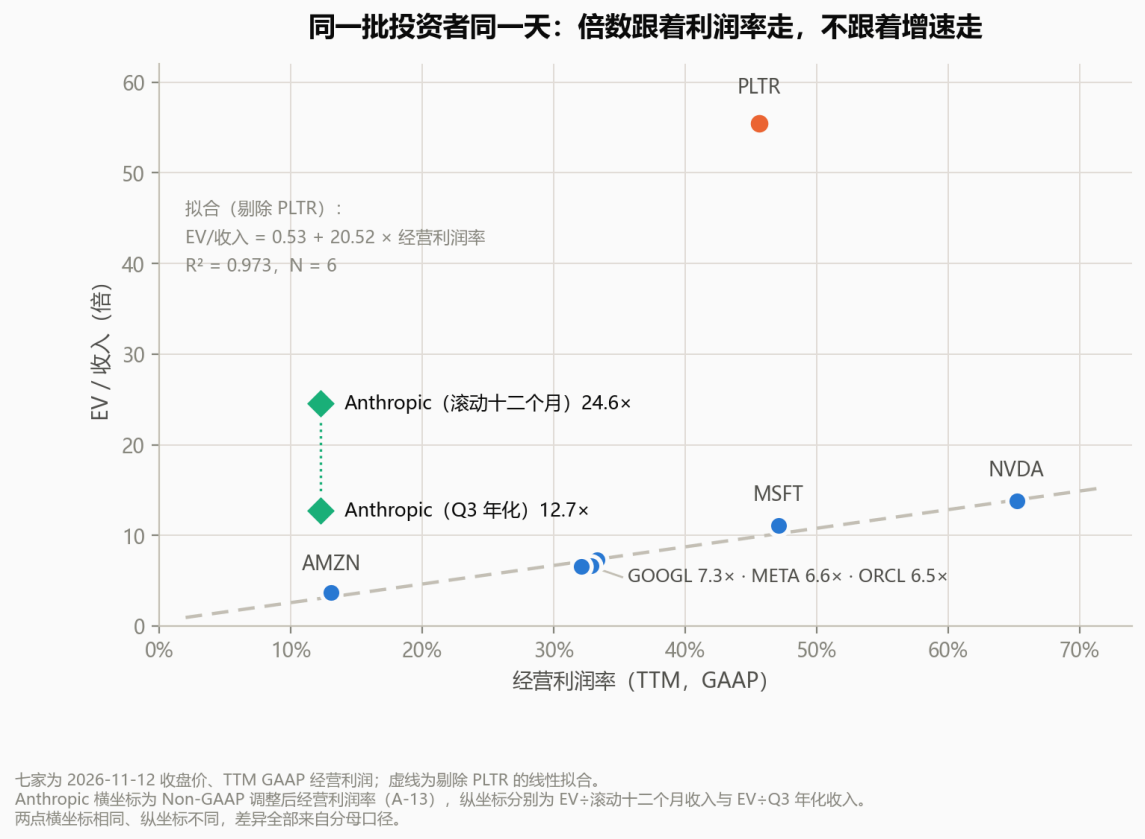


图 8-6

用六家（剔除 PLTR）做线性拟合：**EV/收入 = 0.53 + 20.52 × 经营利润率**，**R^2 = 0.973**；同样六家用收入增速解释 EV/收入，**R^2** 只有 **0.571**。拟合斜率 20.5 与表 8-7 的 EV/经营利润均值 22.6 接近，两者互为印证。

这两个 **R^2** 的差，是本节的第一个结论：在这组样本里，横截面倍数的差异九成以上由当前利润率解释，增速只解释一半左右。换句话说，市场给一家公司的估值倍数，主要看它现在多赚钱，而不是看它增长多快。

但必须立刻加上限定：这条关系是恒等式的重述，不是因果发现。EV/经营利润之所以收敛在 20 倍附近，也可能只是因为七家都是成熟公司——它们的增长都已进入可预测区间，因此市场退回到用盈利定价。对一家增速 **100%** 以上的公司，这条关系是否成立，本报告无法从这组数据回答。

8.5.4 第二个结果：同样 85% 的增速，市场给 13.8× 和 55.5×

本节原定的问题是"同等增速的公司市场给多少倍"。样本给出的答案是没有稳定答案。

指标	NVIDIA	Palantir
TTM 收入增速	+83.4%	+87.1%

指标	NVIDIA	Palantir
经营利润率	65.2%	45.6%
EV/收入	**13.8**×	**55.5**×
EV/经营利润	21.2×	121.5×

同样的增速，相差 **4** 倍的收入倍数、**5.7** 倍的经营利润倍数，而利润率低的反而倍数高。

三个可能的解释，本报告无法用公开数据区分：

1. 收入的可预测性不同。 NVIDIA 卖硬件，订单可见度以季度计；Palantir 卖订阅，剩余履约义务可跨年。市场对可预测的收入付溢价。
2. 稀缺性溢价。 Palantir 是七家里唯一"纯 AI 叙事 + 高增长 + 无硬件周期"的标的，资金没有第二个同样干净的入口。
3. 规模与流通盘。 Palantir 的 EV 只有 4,042 亿美元，是七家里最小的两家之一；同样的资金流入对小盘股的估值推动更大。

> 这个不确定性对 **Anthropic** 的估值是直接相关的：上面三条解释里，**Anthropic** 在第 **1** 条上强于 **NVIDIA**、弱于 **Palantir**（模型层的收入没有多年合同，**RPO**/收入仅 **0.4** 倍）；在第 **2** 条上强（前沿模型公司只有两三家）；在第 **3** 条上不适用（发行后市值近 **1** 万亿美元）。三条方向不一致，因此不能通过类比得出一个倍数。

8.5.5 把 Anthropic 放进这张表

Anthropic 中值定价对应 EV 866.6 亿美元（8.2）。用四个分母：

分母	数值（十亿美元）	EV/分母	同口径的可比基准	相对倍数
滚动十二个月收入	35.28	**24.6**×	七家 EV/收入中位 7.3×	3.4 倍
Q3 单季年化	68.0	**12.7**×	同上	1.7 倍
10-31 run-rate	81.0	**10.7**×	同上	1.5 倍
前三季度年化调整后经营利润	5.44	**159.3**×	六家 EV/经营利润均值 22.6×	7.0 倍

前三行的差别全部来自增长。 一家一年增长 12 倍的公司，滚动十二个月收入必然严重低估当前规模；而 run-rate 又比 Q3 年化再高 19%，把倍数再压低到 10.7 倍——这正是 **8.2** 节选择 **Q3** 年化为主要口径（**GAAP**、口径可比）、同时并列 **run-rate** 的理由。

最后一行才是问题所在。 用可比的 EV/经营利润锚反解：

> 要用六家的平均 **22.6× EV/**经营利润撑住 **24.6×** 的 **EV/**收入，需要 **109%** 的经营利润率。
>
> 这个数字在算术上就不成立——利润率的上限是 **100%**。

所以 **Anthropic** 的当前倍数，在"按当前盈利能力定价"的框架里无法被解释。 它只能由未来盈利解释——这正是 8.3 节反向 DCF 得出的结论（需要五年 31.5% 增速 + 25% 终值 FCF 利润率），两条互相独立的路径指向同一件事。

而 **159.3×** 还是所有口径里最低、对公司最有利的一个倍数。 它用的是公司 Non-GAAP 调整后经营利润，加回了 32.55 亿美元股权激励与 20.86 亿美元非现金捐赠（A-12）：

表 8-8 同一家企业价值下，Anthropic 的 EV/经营利润取决于怎么算利润

利润口径	9M2026（百万美元）	年化	EV/经营利润
GAAP 营业亏损	−1,260	−1,680	不适用（负）
加回非现金捐赠，仍计股权激励	+826	+1,101	**787×**
公司 Non-GAAP（两项都加回）	+4,081	+5,441	**159.3×**

数据来源：A-08、A-11、A-12。年化为 9M × 4/3，本报告计算。

同一家公司、同一个企业价值，倍数可以是 **159** 倍、**787** 倍，或"不适用"。这不是会计细节，它说明：对一家高速增长、大额股权激励、大额非现金项目的公司，**"EV/经营利润"**这个指标本身失去了分辨力。用它跟成熟公司比较，得到的差异主要来自口径，不来自经营。

七家里唯一可以类比的是 **Palantir** 的 **121.5×**。它证明市场确实会给一家高增长公司超过 100 倍的经营利润倍数——因此 **159.3×** 不是史无前例的。但它同时也说明，这类倍数的成立依赖于"增长会持续到利润追上倍数"，而不是依赖于当前利润。

表 8-9 按各可比公司的 EV/收入给 Anthropic 的隐含企业价值

参照	EV/收入	隐含 Anthropic EV（十亿美元）
AMZN	3.7×	130
ORCL	6.5×	231
META	6.6×	235
GOOGL	7.3×	259
MSFT	11.1×	391
NVDA	13.8×	487
七家中位	**7.3×**	**259**
Anthropic IPO 中值（实际）	**24.6×**	**867**
PLTR	55.5×	1,957

按 TTM 收入 352.82 亿美元计算。若改用 **Q3** 年化收入 **680** 亿，所有隐含值同比例放大 **1.93** 倍，但相对位置不变。

这张表的用途不是"算出 **Anthropic** 该值多少"，而是标出市场自己的报价区间有多宽：同一批投资者，同一天，对七家 AI 关联公司的收入出价从 3.7 倍到 55.5 倍。**IPO** 中值的 **24.6** 倍落在这个区间的上半段，高于中位 **3.4** 倍，低于 **Palantir**。

8.5.6 本节能支持与不能支持的结论

能支持的：

1. 模型层的定价无法用"当前盈利能力"解释。用可比的 EV/经营利润锚反解，需要 109% 的利润率——算术上不成立。
2. 定价必须由未来盈利解释，且这个解释与 **8.3** 节的反向 **DCF** 一致（两条独立路径：一条从市场价格反解增速，一条从可比倍数反解利润率）。
3. 市场对高增长 **AI** 资产的报价区间极宽（3.7×–55.5× 收入），且在这个区间内，当前利润率比增速更能解释倍数位置（R^2 0.973 vs 0.571）。
4. **159×** 的调整后经营利润倍数不是史无前例——Palantir 的 121.5× 是同一量级的先例。

不能支持的：

1. 不能判断 **IPO** 定价偏高或偏低。 可比倍数给出的是市场愿意支付的价格区间，不是价值。
2. 不能把七家中的任何一家当作"同等增速的可比"。 增速最接近的两家（NVDA 83%、PLTR 87%）本身相差 4 倍，说明增速不是定价的主导变量。
3. 不能外推到模型层的稳态倍数。 七家都是已经盈利的成熟公司，模型层的终局利润率尚未可知（第八章 8.1 的三个障碍）。

8.5.7 本节的口径局限

局限	影响方向	量级
不扣除长期投资组合	高估 GOOGL、AMZN 的 EV 与倍数	GOOGL 非流动投资 1,352 亿若扣除，EV/经营利润由 22.0× 降至 21.2×
股本用加权平均基本股数，非最新实际流通股	方向不定	各家的差异在 1% 量级
ORCL 的债务与现金同日（2026-08-31），但收入 TTM 止于同日	无	已对齐
NVDA 财务期末 2026-07-26，比其余早约两个月	若三季度继续高增长，其 TTM 收入偏低、倍数偏高	方向已知，幅度未计算
七家均非纯模型层业务	倍数包含非 AI 业务的估值	无法分离（PLTR 最接近纯 AI，也是唯一极端值）
经营利润为 GAAP，未剔除股权激励	与 Anthropic 的 Non-GAAP 口径不对称	已在表 8-8 单独处理

最后一条尤其重要：本节所有对比都在"拿 **GAAP** 利润的成熟公司"与"**Non-GAAP** 利润的成长公司"之间进行。本报告选择不用同一口径强行对齐，因为对齐需要假设一个 Anthropic 的稳态股权激励比例，而这个比例没有依据——用假设去消除口径差异，只会把一个已知的偏差换成一个未知的偏差。

8.6 OpenAI 的隐含区间

8.6.1 先看数据来源的性质

OpenAI 不是上市公司，没有 **S-1**、没有经审计的公开报表。 本节所有收入数字的来源分为三类，结论的可信度受最低一类限制：

数据	数值	来源级别
周活用户 / 付费订阅 / 企业收入占比 / API 吞吐	9 亿+ / 5,000 万+ / >40% / 150 亿 token 每分钟	**A**：公司官方博客（**2026-03-31**）
融资额与投后估值	1,220 亿美元 / **8,520** 亿美元	**A**：公司官方博客（**2026-03-31**）
2026-03 月收入	约 20 亿美元/月	A（同上）
FY2025 收入与净亏损	130.7 亿 / −385 亿美元	**B**：泄露文件，**FT** 独立核实，未审计
2026 年化 run-rate	7 月约 300 亿 → 8 月中 >400 亿 → **9** 月末接近 **500** 亿	**B**：公司向投资人披露，经媒体报道
2026 Q3 单季收入 / 经营亏损	100 亿 / −109 亿美元	**B**：投资人材料，经 **WSJ** 报道
年末 run-rate 预期	**≥700** 亿美元	**B**：公司向投资人（**2026-10-09**）
新一轮融资	目标最高 200 亿，投后 **1.25** 万亿美元，未关闭；首笔约 110 亿（2026-11-07）	B：媒体报道
至 2030 年算力支出计划	约 7,500 亿美元（计划，非签约）	B：媒体报道

> 口径警告一：广为流传的**"OpenAI ARR 逼近 700 亿美元"**不是公司口径。 OpenAI 自己给投资人的数字是 9 月底约 500 亿美元。700 亿是投资人为与 **Anthropic** 做可比而做的总额法调整——Anthropic 按云伙伴全额确认收入（渠道分成记为费用），OpenAI 只计自己那一份。据 CNN 引述的消息源，"700 亿不是 OpenAI 给的"。

8.6.2 三个分母给出三个倍数

表 8-10 OpenAI 的市值/收入倍数（按不同分母与不同估值）

分母	数值（十亿美元）	× 8,520 亿（2026-03 已成交投后）	× 12,500 亿（洽谈中，未关闭）
Q3 单季 × 4（机械年化）	400	21.3×	31.3×
公司口径 run-rate（9 月末）	500	**17.0×**	25.0×
投资人总额法调整后（年末预期）	700	12.2×	**17.9×**

数据来源：估值来自公司官方博客与媒体报道；收入口径见 8.6.1 的来源表。倍数由本报告计算。

8.6.3 一个必须纠正的日期错配

把 **3** 月的估值除以 **9** 月的收入，会把 **OpenAI** 的倍数低估约一半。

- OpenAI 最后一个已成交估值是 **2026** 年 **3** 月的 8,520 亿美元，当时月收入约 20 亿美元（年化约 240 亿）。
- 分母用的是 **9** 月末的近 500 亿 run-rate。

用 **3** 月估值除以 **3** 月收入，倍数是 **35.5** 倍；除以 **9** 月收入，是 **17.0** 倍。 两倍的差距不来自任何经营变化，只来自分子与分母的时点不同。

Anthropic 这边本来也有类似的错配：分子是 11 月 12 日的定价区间，分母若用 Q3 收入则止于 9 月 30 日，相隔约六周。**10** 月 **31** 日的 **run-rate** 披露正好补上了这个缺口，把分母推进到与 OpenAI 的年末预期相近的时点。

> 因此本节唯一把两边都放在各自最近期的比较是：**Anthropic** 按 **11** 月的定价区间除以其 **10** 月 **31** 日的 **run-rate**，**OpenAI** 按 **10** 月底洽谈的 **1.25** 万亿美元除以其年末 **run-rate** 预期。 要说明的是，这仍然不是同一时点——一边是 10 月 31 日已发生的用量、一边是年末的预期。它只是"各自最近期"，差距里仍有相当一部分来自这个时点差。

表 8-11 各自最近期对照（市值/年化收入）

	Anthropic	OpenAI
估值	9,684 亿美元（IPO 中值，11-12）	12,500 亿美元（洽谈中，10-31，未关闭）
分母	810 亿（10-31 run-rate，公司披露）	700 亿（年末 run-rate 预期，非公司披露）
倍数	**12.0×**	**17.9×**
分母性质	公司披露的年化估计（当期已发生用量外推，不做增长假设；非 GAAP；未经第三方独立验证，A-57）	预测（年末，非公司披露）
口径	总额法（S-1 明确）；run-rate 为公司自报口径	投资人总额法调整（非公司口径）

这张表最诚实的读法是：它不足以支持**"谁更便宜"**的结论。

理由有三条，每一条都足以单独推翻比较：

1. 分母性质不同，而且两者不是同一时点。 Anthropic 的 810 亿是 10 月 31 日公司披露的经营指标——它把 10 月一个月的用量年化，不是 GAAP 收入，由内部系统计算、未经第三方独立验证（A-57）；OpenAI 的 700 亿是

年末预期，且这个预期本身是公司给投资人的口径，未经审计。一个是已发生用量的外推，一个是还没发生的事。

2. 折算系数未公开。 OpenAI 的 700 亿是第三方做的总额法调整，幅度 +40%（500→700）；若退回公司口径的 500 亿，OpenAI 的倍数是 25.0 倍。同一次比较在 **17.9** 倍与 **25.0** 倍之间摆动，摆幅（**40%**）与两家倍数的差距（**12.0 vs 17.9**，约 **49%**）是同一量级——结论仍由口径选择驱动。需要更正一个容易说反的直觉：这次的差距（49%）其实小于用 Q3 年化做的比较（119%），与用 3 月 8,520 亿估值做的比较（49.6%）基本持平。也就是说，run-rate 让分母更可比，但没有让两家的差距变小到"可以下判断"的程度。
3. 只能用市值，不能用企业价值。 Anthropic 有 S-1 备考资产负债表（现金 2,126.84 亿、可转债 1,108.73 亿，A-39、A-40），OpenAI 没有公开资产负债表。两家现金与债务结构不同，市值/收入不可比，而 **EV**/收入算不出来。

8.6.4 一个差距更大的变量：经营亏损

表 8-12 相近收入量级下的经营损益对照

	Anthropic（2026 前三季度）	OpenAI（2026 Q3 单季）
收入	332.42 亿美元	100 亿美元
经营损益	**−12.60** 亿美元	**−109** 亿美元
经营利润率	**−3.8%**	**−109%**

数据来源：Anthropic 为 S-1/A 性质分类口径（A-11）；OpenAI 为投资人材料经 WSJ 报道（未审计）。百分数由本报告计算。

两家在相近的收入量级上，经营亏损差一个数量级，方向也相反——一家接近盈亏平衡，一家单季亏损超过收入。

这个对比同样必须带口径警告： 两者的成本归类不同。Anthropic 把训练算力记在研发、把免费层算力记在销售费用，两者都在经营费用内；OpenAI 的成本归类未公开。因此**"−109%"**不能直接理解为**"OpenAI** 的经营效率比 **Anthropic** 差 **28** 倍"。

但它能支持一个更弱、也更稳的判断： 若 OpenAI 的亏损结构真实，则它比 Anthropic 更依赖持续融资来覆盖经营现金流出；而融资能力取决于估值能否维持。这使 **OpenAI** 对 **8.3** 节约束二（承诺与现金流的错配）的暴露程度高于 **Anthropic**，而它的算力计划（**7,500** 亿美元至 **2030** 年，且为"计划"而非"签约"）在量级上与 **Anthropic** 的 **9,009** 亿美元承诺相当。

8.6.5 本节结论

1. **OpenAI** 没有公开报表，所有收入数字中只有 **2026-03** 的官方博客属一手；其余为未审计的投资人材料或媒体转述。
2. **"700** 亿美元 **ARR"**不是公司口径，是为与 Anthropic 可比而做的总额法调整。调整幅度按绝对倍数点是 7.1 倍（25.0→17.9），大于两家 5.9 倍的差距；按相对幅度是 40%，小于 49% 的差距——两个基数给出的结论相反，所以不据此判断差距是否被口径放大。
3. **3** 月估值除 **9** 月收入会低估倍数约一半；唯一把两边都放在各自最近期的比较是 12.0×（Anthropic，10 月底 run-rate，公司披露）对 17.9×（OpenAI，年末预期），但两者不是同一时点，且后者的分母是预期而非已发生的用量。
4. 两家的经营亏损差一个数量级（−3.8% vs −109%），但成本归类不同，不能直接比较效率，只能说明 OpenAI 对持续融资的依赖更强。
5. 本报告不认为这两个数字足以判断"谁更贵"。 它们能说明的是：两家在 2026 年末的定价落在 12–18 倍年化收入的同一区间内，差异小于口径误差。

8.7 估值的边界：本报告不做什么

第一，不预测发行后的股价。 IPO 定价区间是承销商与公司商定的结果，受发行规模、锁定期、市场情绪影响，与内在价值的关系不明确。上市首日的价格由供需决定，不由本章的任何计算决定。

第二，不把 **1.4** 万亿美元的媒体说法当作基准。 按 S-1 披露的股本与价格区间，78 美元上限对应 1.02 万亿美元市值；即使按含全部期权与转股的约 155 亿完全稀释股本，也仅为 1.21 万亿美元（A-42）。要达到 **1.4** 万亿，股价需达 **107** 美元（基本股本）或 **90.3** 美元（完全稀释），均高于披露区间。 这个说法在本报告中只作为"媒体转述与原始文件偏差"的案例，不作为估值输入。

第三，不假设承诺会被完全履行，也不假设会被完全取消。 这两种假设各自导向完全不同的结论，而公开信息不足以判断哪一种更可能。本报告把"承诺的可执行性"列为最重要的待验证项（见第九章）。

第四，不评价发行定价是否"合理"。 "合理"取决于投资者的机会成本与风险偏好，这是提出者的事，不是本报告的事。

第九章 风险与局限

9.1 行业层面的五个风险

风险一：承诺的期限错配（本报告认为最高优先级）

9,009 亿美元采购承诺对应 4.25 年内年均约 1,063 亿美元的算力支出，而当前年化收入 680 亿美元（10 月底 run-rate 810 亿）、年化算力费用约 250 亿美元（A-32、A-07、A-55）。承诺的兑现要求收入在四年内达到 **Q3** 年化收入的 **2.8–4.5** 倍，或 **10** 月底 **run-rate** 的 **2.3–3.7** 倍（取决于届时的算力收入占比落在 56.6% 还是 35%；`scripts/runrate.py`）。

第八章已经论证：这个要求与 IPO 定价隐含的增速不兼容。两者必有一个会调整。若调整的是承诺，代价是容量受限；若调整的是收入预期，代价是估值重估。

观察指标：未来财报中"采购承诺"科目的余额变化；算力与基础设施费用的季度增速；经营性现金流是否转负。

风险二：毛利率的关联交易属性

85.5% 的毛利率是在云厂商同时作为股东、供应商、渠道的三重身份下形成的（A-16、A-19、A-22）。这个毛利率不由市场竞争决定，因此也无法用竞争逻辑预测其走向。

观察指标：若算力采购转向更多非股东供应商（S-1 明确说会这样做），毛利率会向市场价收敛——方向大概率是下降。

风险三：开源模型的价格压制

开源权重的边际成本为零，能力持续追赶。它不抢客户，但每一次客户续约都会变成一次降价谈判。

观察指标：单位任务成本的下降幅度是否开始超过算力成本的下降幅度。若是，说明降价已从"技术进步驱动"转为"竞争压力驱动"，毛利率将开始承压。

风险四：监管的合规成本（详见第七章）

欧盟 GPAI 义务、中国备案制、美国州层面立法，三套规则的合规成本不同，且都在演进。对模型层而言，监管的直接影响不是罚款，而是模型发布节奏——而发布节奏直接决定收入（半年内版本贡献 **80%** 收入，**A-29**）。

观察指标：新模型从训练完成到公开可用之间的时间间隔。

风险五：算力供给的物理约束

电力、变压器、先进封装、HBM 供给，任何一个环节的瓶颈都会传导为交付延迟。S-1 明确披露了多方可结构下"钱付了、算力拼不起来"的风险（A-35 同节）。

观察指标：承诺额与实际算力费用的差额（即"已承诺未交付"的部分）。

9.2 公司层面的四个风险

风险	具体机制	证据	影响量级
摊薄	完全稀释股本约 155 亿 vs 发行后 130.87 亿；公益基金可按股价里程碑再获 2%	A-42、A-52	约 18% 的即时摊薄，长期更高
控制权	Class F 1 股拥有 50.1% 投票权，Class N 无投票权	A-51	少数股东无法影响任何决策
客户集中	前五大占 25%，无单一客户 >10%	A-27	可控，但非零

风险	具体机制	证据	影响量级
渠道依赖	44% 收入经云厂商市场；CSP 应收占应收净额 52%	A-20、A-21	渠道中断将直接冲击收入与回款

摊薄与控制权这两项，本报告认为被市场低估。一家公司可以在业务上极其成功，而公开市场股东仍然得不到相应的回报——如果绝大部分经济收益通过股权结构、公益基金发行、以及持续融资流向其他方。这不需要预测业务，只需要读股本结构。

9.3 本报告的局限：未定、未测、已知缺陷

> 按本报告的写作规范，交付必须附局限清单，并区分"做不到"与"不存在"——两者混为一谈会误导对未来的判断。

未定（数据存在，但本报告未能取得或未能核实）

项目	状态	影响
Anthropic 收入的地域拆分	不存在于披露中（全文检索无结果）	无法判断中国/欧洲市场的实际贡献
Anthropic 分产品收入（API vs 应用 vs 订阅）	不存在于披露中	无法判断增长来自哪一层需求
云厂商通道费率	未单独披露，含于销售与营销 66.19 亿	无法精确计算渠道对毛利的侵蚀
算力采购的折扣率	未披露	无法判断毛利率中有多少来自关联交易定价
OpenAI 最新收入与估值的官方口径	待核实（见 8.6）	横向对照的可信度受限
第三方市场规模估算的一致口径	各机构口径差异极大	无法用 TAM 作为估值锚（见 8.3 约束三）

未测（本报告没有做的事）

项目	为什么没做	影响
模型能力的独立评测	需要统一的评测框架与算力，超出本报告范围	第五章的竞争分析不含能力排序，只含结构分析
客户访谈与渠道调研	无一手渠道	需求判断全部基于披露数据，缺少交叉验证
算力成本的独立估算	缺少 GPU/TPU 的实际采购价格与利用率数据	无法验证 56.6% 的算力费用率是否合理
承诺额的可执行性评估	需要合同原文，不在公开披露范围内	第八章的解释 A/B/C 无法进一步收敛

已知缺陷

1. 单公司样本。本报告的量化分析几乎全部基于 Anthropic 一家公司的披露。用一个最激进的增长样本代表整个模型层，会系统性高估行业的增长速度与毛利率水平。本报告在文中多处提示了这一点，但无法用数据修正。
2. **2026** 年第四季度是估计值。2026 年收入 502.4 亿美元中，第四季度按 Q3 的 170 亿美元持平处理。这是保守假设，但仍是假设。若按 10 月底 run-rate 折算（Q4 约 202.5 亿），全年收入为 534.9 亿，第八章的隐含增速由 31.5% 降至 30.3%——下移约 **1.2** 个百分点，不改变任何结论的方向（`scripts/runrate.py`）。
3. **DCF** 的终值依赖。三个情景的终值占比为 66%–78%，大部分价值来自 **2036** 年之后的假设，而那部分的可靠度最低。
4. 趋势外推的风险。本报告多处使用"当前增速""当前毛利率"作为分析起点。在一个一年增长 **12** 倍的行业里，任何"当前"都是瞬态。

9.4 什么会让我们改变判断

一份负责的报告必须写明自己的可证伪条件。以下五条中任何一条成立，本报告的核心判断就需要修正：

若观察到	则本报告的哪条判断被推翻
2027 年毛利率跌破 70%	第一章"85.5% 是周期顶部"被强化；第八章基准情景的利润率假设过于乐观
算力与基础设施费用增速连续两个季度超过收入增速	第三章"效率改进真实存在"不再成立
采购承诺余额大幅下调（超过 20%）	第八章解释 A 成立，但"容量受限"的代价需要重新评估
前五大客户占比升破 40%	第五章"分发关系是护城河"被削弱，增长重新变成大客户依赖
开源模型在主流基准上追平闭源前沿	第五章"开源是定价上限"从约束变成替代，模型层定价权消失

9.5 本章结论

1. 最高优先级的风险不是技术，是承诺的期限错配：四年内（到 2030 年底）需要把收入提高到当前 Q3 年化的 2.8–4.5 倍（按 10 月底 run-rate 为 2.3–3.7 倍）才能兑现承诺。
2. 毛利率由关联交易定价决定，无法用竞争逻辑预测；转向非股东供应商会使其向市场价收敛（下降）。
3. 摊薄（约 18% 即时、长期更高）与控制权（Class F 一股 50.1% 投票权）被市场低估。
4. 本报告的核心局限是单公司样本与终值依赖；两者都无法用现有数据修正，只能明确标注。
5. 五条可证伪条件已列明，其中"算力费用增速超过收入增速"是最灵敏的早期信号。

附录

附录 A 数据提取与校验方法

A.1 主源与提取路径

本报告的主源是 Anthropic, PBC 于 2026 年 11 月 10 日提交的 S-1/A 修正案（Amendment No.2，注册号 333-299633，CIK 0002130357，accession 0001628280-26-072484，文件 `anthropicpbc-sx1ax2.htm`），本地留档 14,798,492 字节、约 1,480 万字符。同时留存了 2026 年 10 月 26 日的 S-1 初版，仅用于比对"哪一版改了什么"，所有数字一律取自 S-1/A。

提取不是"让模型读全文然后复述"，而是四级可回溯的链条：

级别	产物	文件	作用
1	原始文件	`evidence/anthropic_s1a_20261110.htm`	唯一真值
2	分节纯文本	`evidence/s1a_is.txt`、`s1a_note8.txt`、`s1a_core.txt`、`s1a_shares.txt`、`s1a_cogs.txt` 等 8 个，另有 `s1a_runrate.txt`（2026-11-12 补建）	把 1,480 万字符切到可人工比对的粒度
3	结构化常量	`scripts/build_facts.py` 顶部字典	抄录结果的唯一入口
4	机读结果	`evidence/facts.json`	图表、表格、正文数字全部从这里取

为什么保留第 **2** 级： 报告里任何一个数字，都可以从正文 → `facts.json` → 常量字典 → 对应分节文本 → HTML 逐级回溯到原表。没有这一步，"可追溯"只是一句承诺。

一处已修复的提取缺口。 Run-Rate Revenue 的定义与数值（A-54 至 A-56）写在业务章节的正文段落里，不在任何一张表中；按表提取的 8 个文件均未覆盖，第一版报告因此完全没有使用这个数字。补建 `s1a_runrate.txt` 后，`build_facts.py` 增加了一条断言：代码里的 **run-rate** 常量必须能在证据件的原文中命中。这类缺口不可能靠内部一致性检查发现，只能靠"按主题而不是按表"回原文再搜一遍。

A.2 内部一致性检查（自加校验）

`scripts/build_facts.py` 对每个报告期（FY2023、FY2024、FY2025、9M2025、9M2026）跑四组自加检查，任一不符即中断：

表 A-1 四组自加检查及其排除能力

组	检查式	能发现的错误	不能发现的错误
1 利润表分项	算力 + 人员 + Bartz + 销售伙伴 + 其他管理 = 营业费用合计；收入 − 费用 = 营业亏损；营业亏损 + 利息 + 公允价值变动 + 其他 = 税前；税前 − 税 = 净利	抄错一位数、列错行、漏抄某一分项	整列抄自错误期间
2 附注 8 自加	收入成本 + 研发 + 销售 + Bartz + G&A = 附注合计；附注合计 = 利润表营业费用合计	功能口径与性质口径的抄录不一致（跨表检查）	两表同时抄错
3 Non-GAAP 调节	营业亏损 + SBC + 非现金捐赠 + Bartz = 调整后经营利润	调节项漏项、符号错	调节项定义理解错
4 股本	A 7,427,991,791 + N 5,658,680,800 + T 1 + F 1 + E 0 = 13,086,672,593	分项漏加、股本口径混用	股本本身不含某些类别（口径问题）

必须说清楚这四组是什么、不是什么。 它们是内部一致性检查：能发现抄录与加总错误，不能发现"原文的口径与我理解的不同"——因为检验用的就是同一份文本。要发现后一类错误，必须引入外部锚，见 A.3。

A.3 外部锚：四类"已知答案的输入"

锚	输入	已知答案	本报告得到	结论
公司自己给出的另一口径数字	TTM 收入 352.82 亿 ÷ 期末员工 5,300 人	S-1 原文 "over $6.6 million" 人均收入	665.7 万美元	吻合。两条独立路径（我的 TTM 拼装 vs 公司的口径）落在同一处，说明 TTM 拼装无系统性偏差——这是本报告最有效的一次外部校验
反解校验	Q3 2026 收入 170 亿、同比 +1,127%	Q3 2025 应约为 170 ÷ 12.27 ≈ 13.9 亿	13.86 亿美元；与 9M2025 的 25.45 亿相减得 H1 2025 ≈ 11.59 亿	自洽。若把"增长 1,127%"误算成"是去年的 1,127 倍"，会得到 1,920 亿这种量级即错的数，反解会立刻暴露
媒体说法作为负对照	媒体称估值 1.4 万亿美元	在 130.87 亿基本股本下需股价 107 美元；在 155 亿完全稀释股本下需 90.3 美元	披露区间为 70–78 美元	不成立。流程在给定输入下报告"不成立"，说明它不是只会说"成立"的流程（三条媒体核对见附录 B.4）
相邻业务的独立披露	xAI 的 "SI" 分部（SpaceX 10-Q）	若"规模摊薄"是真机制，另一家模型业务也应出现毛利率随规模上升	29.8% → 64.6%	机制层面吻合。这不能验证 Anthropic 的 85.5% 这个数，只能验证机制存在——两者的区别在第一章已写明

> 这四类锚各自能排除什么，必须分开记： 第一类排除 TTM 拼装错误；第二类排除增长率公式方向错误；第三类排除"核对流程永远说成立"；第四类只排除"机制不存在"。没有一类能排除"公司披露本身有误"，那需要审计，而本报告不是审计。

A.4 可比公司数据重建（SEC XBRL）

第八章的估值可比表（表 8-7、图 8-6）不是从行情软件抄的，而是从 SEC XBRL 原始提交重建：

- 数据源：`data.sec.gov/api/xbrl/companyfacts/CIK##########.json`，逐家落盘至 `evidence/facts/<TK>.json`（MSFT、GOOGL、AMZN、META、NVDA、ORCL、PLTR 共 7 份，单份约 4.9 MB，单次取数约 8.6 秒）。
- 单季还原：XBRL 中同一 fiscal-year 起点下并存多条 YTD 记录（3 个月、6 个月、9 个月、12 个月），单季值必须由同期 YTD 逐条相减得到。
- 去重与优先级：同一期末常同时出现"直接三月期"与"由 YTD 差分推得"两条记录，取优先级 = 2（直接三月期，时长 80–100 天）优于 1（推得）。
- 窗口连续性：TTM = 连续四季之和，要求首季期末与末季期末相隔 240–300 天；不满足则整窗丢弃，不拼接。
- 同期资产负债表：现金、短期投资、有息负债用同一期末，取数容差 45 天，避免用 3 月末的现金去除 9 月末的市值。
- 恒等式断言：`EV/收入 = EV/经营利润 × 经营利润率`，逐行校验 `|差| < 1e-6`。

表 A-2 重建过程中被检查出来的六个错误

#	症状	根因	修法
1	MSFT 的 TTM 序列为空，程序在 `rev["ttm"]` 处崩溃	同一季度既有直接三月期、又有 YTD 差分推得，两条期末相同但值不同，窗口被判为不连续	按期末去重，直接三月期优先
2	全部公司 TTM 为空	连续性判据写成"首尾相隔 < 300 天"，而四个季度首尾实际只跨约 274 天	判据改为 240–300 天
3	GOOGL 收入停在 2025-03-31、NVDA 停在 2022-01-30	取收入标签时返回"第一个能算出数的标签"，而不是"期末最新的标签"	改为在全部候选标签中取期末最晚者
4	ORCL 有息负债记为 0，EV 变成净现金	`LongTermDebtNoncurrent` 对该司 404；`LongTermDebt` 是 2022 年的陈旧零	增加 `LongTermNotesAndLoans` 等标签；容差由 100 天收到 45 天，强制与现金同日
5	PLTR 同比 +154.3%、META +46.9%	同比基数误用最近一个年度数，而非上年同期 TTM	在序列内部取 350–380 天前的 TTM 作基数
6	恒等式断言不成立	EV/收入与 EV/经营利润用了不同期末的资产负债表	每行统一期末，断言逐行跑

这里的外部锚是什么： 这七家公司的收入增速与利润率是公开事实（NVDA 约 +83%、MSFT 约 +18%、AMZN 约 +15%）。任何一家算出来差一个数量级，就说明拼装逻辑错了——表 A-2 的第 3、5 两行正是这样被发现的。这不是"再检查一遍"，是拿已知答案去跑。

A.5 方法层面的局限

1. 本报告不是审计。 S-1/A 的数据未经本报告核对原始凭证；四组自加校验只能证明"抄录与原文一致"，不能证明"原文本身正确"。
2. **XBRL** 重建只有一个实现。 没有用第二套独立实现算同一个量（V2 双方法交叉未做），因此第 3、4 行的标签选择错误只能靠"结果明显不合理"发现，不能靠交叉比对发现。
3. **DCF** 与敏感性分析的全部假设是本报告的选择，不是实测。 表 8-2、8-5、8-6 的参数在正文中逐条标注为"选择"，读者可替换后重跑 `scripts/dcf.py`。
4. 未取得的来源单列在附录 B.5，并区分"做不到"与"不存在"。

附录 B 来源分级与证据台账索引

B.1 分级定义

本报告采用行业研究报告来源分级的通行五级制：

级	定义	例子	使用规则
A 一手原文	发布者自己的原文	法规/公告/年报/招股书/官方统计/论文原文/公司官网	可直接作为结论依据；须记 URL、页码、发布日期
B 可复现数据	可重取的原始数据	数据库导出（附查询条件）、API 返回（附原始 JSON）、现场实测	须落盘 + 给重取方法；注明抓取时点
C 权威二手	有编辑把关的转述	券商研报、行业协会报告、权威媒体原创报道	尽量回溯 A；不能回溯时注明"转述，未核原文"
D 一般二手	无编辑把关或易失真	自媒体、百科、聚合站、AI 摘要、转载	只用于发现线索，不作结论依据
X 不可得	本次拿不到	付费终端、未披露数据、登录墙	正文标"不可得"，并说明是"做不到"还是"不存在"

三条硬规则，本报告全部适用：

1. 多家媒体转载同一来源 ≠ 多个来源。 第七章的 IDC 数据被 5 家以上媒体报道且数字一致，但它们是同一个来源的转述，因此按 C 级处理，不因"多家一致"升为 A。
2. **D** 级来源不单独支撑任何结论。 表 1-3 的 OpenAI 毛利率、表 8-12 的 OpenAI 经营亏损均为 D 级，正文均标明"未审计""上限而非实际值"。
3. **A** 级也可能错。 欧盟 AI 法案的原始时间表就被 Regulation (EU) 2026/1744 改掉了——第七章 7.2.1 专门处理这一点。

B.2 来源登记表

表 B-1 本报告实际使用的来源（按首次出现顺序）

#	来源	发布者	时点	级别	取得状态	支撑的内容
1	Form S-1/A （Amendment No.2）	Anthropic, PBC / SEC EDGAR	2026-11-10	**A**	原文已核，本地留档	第一至九章全部 Anthropic 数据（A-01 至 A-57）

#	来源	发布者	时点	级别	取得状态	支撑的内容
2	Form 10-Q，"SI" 分部	SpaceX / SEC EDGAR（CIK 0001181412）	2026-09-30 期	**A**	原文已核	xAI 分部收入与毛利（表 1-3）
3	XBRL companyfacts	SEC	2026-11-12 抓取	**B**	原始 JSON 逐家落盘	七家可比公司 TTM 与资产负债表（表 8-7）
4	收盘价	Yahoo Finance	2026-11-12	C	已取	七家可比公司市值
5	Regulation (EU) 2024/1689	EU Publications Office	2024-07-12	**A**	原文已核	欧盟 AI 法案第 51、53、55、99、101、111、113 条
6	Regulation (EU) 2026/1744（Digital Omnibus on AI）	EU Publications Office（CELEX 32026R1744）	2026-07-24 刊登	**A**	修订条款原文已核	高风险条款推迟至 2027-12-02 / 2028-08-02（表 7-1）
7	《生成式人工智能服务已备案信息的公告（2026年7–8月）》	国家互联网信息办公室	2026-09-14	**A**	原文已核	备案 1,112 款 / 登记 731 款（表 7-2）
8	《生成式人工智能服务管理暂行办法》（七部门令第 15 号）	七部门	2023-07-10 公布	**A**	条文已核	准入与训练数据合规（7.3.3）
9	《人工智能生成合成内容标识办法》（国信办通字〔2025〕2 号）	国家网信办等	2025 年发布，2025-09-01 施行	**A**	条文已核	显式 / 隐式标识义务（7.3.3）
10	备案系统合并通告	国家网信办（beian.cac.gov.cn）	2026-11-09	**A**	原文已核	四类备案合并（表 7-2 备注）
11	商务部、科技部 2026 年第 48 号公告	商务部	2026-10-16	**A**	公告已核	23 项技术出口目录（征求意见至 2026-11-15）
12	TAKE IT DOWN Act，Public Law 119-12	美国国会 / govinfo	2025-05-19	**A**	原文已核	美国联邦专门法（表 7-3）
13	EO 14365，90 FR 58499	白宫 / Federal Register	2025-12-11	**A**	原文已核	行政令设立 AI 诉讼工作组，但不废止州法
14	加州 SB 53（Chapter 138, Statutes of 2025）	California LegInfo	2025-09-29 签署	**A**	条文已核	10^{26} FLOP / 5 亿美元双门槛
15	科罗拉多 SB 26-189	leg.colorado.gov	2026-05-14 签署	**A**	原文已核	废止并重立 SB 24-205（表 7-3）
16	韩国《AI 基本法》	law.go.kr	2026-01-22 施行	**A**	法条页已核	罚则 3,000 万韩元 + 一年缓期（表 7-4）
17	《2026: The State of Enterprise AI》	Menlo Ventures / Morning Consult 执行	2026-11-03	C	报告页已取	企业 LLM API 支出份额（表 5-1）
18	《The Plunging Price of Thought》	Epoch AI	2026	C	已取	推理价格下降速度（表 6-1）
19	平台用量与实付均价	OpenRouter	2026-10-22	**B**	公开帖文与公开端点	实付均价 $0.77→$0.14、agent 占比（表 6-1）
20	《中国 AI 软件市场半年度追踪，26H1》	IDC	2026 年 10 月下旬	C	仅媒体转述（付费墙）	MaaS 106.3 亿元、份额两口径（表 7-6）
21	旗舰模型 API 定价页	阿里云百炼、火山方舟、DeepSeek、智谱开放平台、Kimi 开放平台、MiniMax 开放平台、腾讯云 TokenHub	2026-11-12 查询	**A**	页面已取	表 7-7 全部价格
22	三份新闻稿	Gartner	2026-07-20 / 2026-11-12	C	新闻稿已取，报告原文未取	"AI 市场规模"三口径（表 8-4）
23	市场规模网页	Precedence Research、Grand View Research	2026	C	网页已取	表 8-4 第三、四口径

#	来源	发布者	时点	级别	取得状态	支撑的内容
24	OpenAI FY2025 财务文件	Financial Times 核实（原始文件匿名来源）	2025	**D**	转述	OpenAI 毛利率上限 53.7%（表 1-3）
25	OpenAI 投资人材料	Wall Street Journal 报道	2026	**D**	转述	OpenAI 经营亏损（表 8-12）
26	OpenAI 融资与估值公告	OpenAI 官方博客	2026	**B**	已取	隐含估值（8.6）

> 级别与可靠性的关系不是单向的。 第 20 行（IDC）是被 5 家以上媒体转述的一致数据，仍只能是 C 级，因为它仍是一个来源；第 24、25 行（OpenAI）虽是 D 级，但本报告只用它们做上限与量级判断，不用它们支撑任何结论。

B.3 证据台账索引（A-01 至 A-57）

台账全文在 `evidence/EVIDENCE_LEDGER.md`，逐行给出数值、口径、时点、出处与级别。按主题分组：

组	编号	内容
文件与期间	A-01 至 A-06	注册文件版本、各期收入、单季收入与同比
利润表与费用	A-07 至 A-18	按性质/按功能的费用、营业亏损、净亏损、SBC、调整后利润、毛利率倒算
收入质量	A-19 至 A-24	总额法确认、渠道占比、应收集中度、CSP 股东身份、收入模式、地域拆分未披露
客户	A-25 至 A-31	NRR、客户分层、集中度、版本贡献、员工数、人均收入
承诺与合同	A-32 至 A-34	9,009 亿采购承诺、RPO 138 亿、递延收入
资产负债与股本	A-35 至 A-43	可转债本金与公允价值、累计融资、现金、资本化表、发行条款、股本、净募资
产品与治理	A-44 至 A-53	模型族、单位成本、芯片组合、客户交叉使用、现金流、资本开支、控制权、公益股、认购意向
更新口径（run-rate）	A-54 至 A-57	Run-Rate Revenue 的定义、2026-10-31 的 810 亿（710 亿为推导）、口径警告三项、未经第三方独立验证

B.4 已发现的媒体与原文不符（三处）

#	媒体说法	原文	差额	处理
1	采购承诺 8,954 亿美元	**9,009** 亿美元（A-32）	55 亿美元	全文用 9,009 亿
2	估值 1.4 万亿美元	78 美元上限 × 130.87 亿股 = **1.02** 万亿；155 亿完全稀释股本下 1.21 万亿	需股价 107 / 90.3 美元，均超出披露区间	在核心判断六与 8.2 中明确写出该说法不成立
3	调整后经营利润 40.8 亿美元（易被读成单季）	属 **9M2026** 累计（A-12）	口径错，非数值错	表 8-1 与 8.5 均标"前三季度年化"

B.5 未取得的来源（X 级）

来源	状态	属于	对报告的影响
Gartner 报告原文	未取得（付费墙），仅取新闻稿	做不到	表 8-4 只能用新闻稿口径，段间是否重叠无法核实
IDC 报告原文	未取得（付费墙），仅取媒体转述	做不到	表 7-6 按 C 级处理，份额两口径并列
Precedence / Grand View 报告原文	未取得，仅取网页摘要	做不到	同上
商务部答记者问原文	未取得，仅二手转述	做不到（截至 2026-11-12 未检索到原文）	7.3.4 明确写"据商务部答记者问"

来源	状态	属于	对报告的影响
科罗拉多 SB 24-205 诉讼的法院文书	未取得	做不到	7.4 改为"是否有联邦法院禁执令，未核实"
Anthropic 收入地域拆分	S-1 未披露	不存在	不填补，正文标注（A-24）
Anthropic 分产品收入	S-1 未披露	不存在	不填补
Anthropic 算力采购折扣率	未披露	不存在	8.5.6 明确"用公开信息无法判定"
Anthropic TAM 估算	S-1 未披露	不存在	不用行业总量倒推公司份额

附录 C 图表索引

C.1 图（共 18 张）

编号	标题	章节	数据来源
图 1-1	Anthropic 收入：从 4,100 万美元到单季 170 亿美元	1.2	A-02、A-03（`fig01`）
图 1-2	营业费用结构：算力始终是第一大项	1.3	A-07（`fig04`）
图 1-3	毛利率与调整后经营利润率的两条路径	1.4	A-18、A-12（`fig03`）
图 1-4	1,040 亿美元净亏损中，943 亿是非现金的公允价值变动	1.4	A-09、A-10（`fig06`）
图 2-1	单季收入的加速	2.1	A-04、A-05、A-06（`fig02`）
图 2-2	客户数与客单价的四个台阶	2.2	A-26（`fig08`）
图 3-1	五个分母下的承诺倍数	3.2	A-32、A-07、A-37、A-39、A-55（`fig07`）
图 3-2	9,009 亿美元承诺的期限结构	3.2	A-32（`fig13`）
图 4-1	渠道结构的演进	4.1	A-20（`fig09`）
图 4-2	收入的去向	4.2	A-07（`fig05`）
图 7-1	三条监管路径，三种传导机制	7.1	欧盟 2024/1689 与 2026/1744、网信办公告、加州 SB 53、EO 14365（`fig14`）
图 7-2	中国 API 定价是分层的，不是单一价格市场	7.7.2	七家厂商官方定价页，2026-11-12（`fig15`）
图 8-1	S-1 定价区间对应的市值与企业价值	8.2	A-40、A-41、A-39（`fig10`）
图 8-2	支撑当前估值所需的前五年收入增速	8.3	`scripts/dcf.py`（`fig12`）
图 8-3	同一个 2030 年：承诺要求比估值隐含高 9%–102%	8.3	A-32 + 本报告计算（`fig11`）
图 8-4	第三方市场规模估算与单公司实测收入	8.3	表 8-4 四口径 + A-05（`fig17`）
图 8-5	三个情景的企业价值跨度 20 倍	8.4	`scripts/dcf.py`（`fig16`）
图 8-6	EV/收入与经营利润率：市场在按利润率定价	8.5	SEC XBRL + Yahoo Finance（`fig18`）

分布说明：第五章（竞争格局）与第六章（行业阶段）不含图，因为这两章的核心论证是口径辨析与机制判断（同一份调查的两个分母、成本下降速度的三种测法），用表格比用图更准确；图集中在第一、八章，因为这两章的核心是数量关系。总图数 18 张，落在交付要求的 15–20 张区间内。

C.2 表（编号表 32 张，另有 32 张行内对照表不编号）

章节	表号	标题
第一章	1-1 至 1-3	收入模式的三个层次；两种成本口径的映射；外部对照毛利率
第二章	2-1	企业需求的关键指标

章节	表号	标题
第四章	4-1、4-2	闭环的四个环节与可验证的数字；收入的去向
第五章	5-1、5-2	企业 LLM API 支出的份额；护城河检验汇总
第六章	6-1	独立测得的 AI 成本下降速度
第七章	7-1 至 7-8	欧盟时间表；备案累计数；美国监管结构；其他辖区；监管→财务变量；中国 MaaS；中国 API 定价；2026 年资本事件
第八章	8-1 至 8-12	IPO 倍数；隐含增速；承诺要求；市场规模四口径；敏感性；三情景；可比公司；EV/经营利润三口径；隐含 EV；OpenAI 倍数；各自最近期对照；经营损益对照
附录	A-1、A-2、B-1	四组自加检查；XBRL 重建的六个错误；来源登记表

编号规则： 只有承载独立论据、会被正文反复引用的表才编号；用于在行文中做两列对照的小表（例如 8.5.4 的 NVIDIA/Palantir 对照）不编号，随文附注来源。按此规则，全文共 66 张表（含本附录的图、表索引两张），其中编号表 32 张、行内对照表 32 张；编号表已按文档顺序排列。

全部图表由 `scripts/charts_s1.py`、`scripts/charts_more.py` 生成，数据来自 `evidence/facts.json`、`evidence/dcf.json`、`evidence/runrate.json`、`evidence/comps_final.json`，正文中的图表数字不由人工二次录入。
