EXPONENTIAL VIEW

# 人工智能经济现状

2026年6月25日

阿兹米尔·阿扎尔，威廉·吉尔德，汉娜·佩特罗维奇博士，内森·沃伦与玛丽亚·加夫里洛夫

指数视角
www.exponentialview.co

# 我们为什么要这样做

人工智能经济存在一个可见性问题。到目前为止，一直无法解构真实的客户需求。

人工智能经济的供给侧已有详尽记载。大多数半导体公司和超大规模企业都是上市公司，并会详细披露其业务活动。卖方分析师在剖析其业绩方面做得非常出色。

客户实际购买的是什么，以及这些收入是否真实，一直很模糊。

需求端，最大的实验室多为私营，即便是公立公司也将其人工智能收入隐藏在部门总收入之中。

若不理解真实需求，便无法判断支撑着2.27万亿美元股市估值、并在过去六个季度推动美国GDP增长的人工智能经济的健康状况。

我们希望这份报告能成为一份客观反映当前形势的参考资料，不受浮夸和恐慌情绪的影响，同时帮助我们所有人就人工智能对经济乃至整个世界产生的引力进行更明智的探讨。

特别感谢 Alex Imas、Shanu Mathew、Patrick Rutherford、Jaime Sevilla 和 Amy Sutter 温情审阅了本次演示的初稿，并给予了我们反馈。

阿兹姆与指数观点团队

# 一种专有的按层级计费的营收模式：采购、评分、三角验证、去重

## 1 来源

**Bottom-up, 1,000+ firms**

**每一笔收入追溯到原始申报** 审计账目、记录文本及可信报告；此外还有云服务归属认定，即当私营公司的收入出现在公有公司账目中时（例如：OpenAI通过Azure，Anthropic通过Bedrock）。

额外 **我们使用的软信号包括非官方来源**
例如：

- 高管及相关方公开发表的意见。
- 代理和样本指标。
- 传统媒体、新媒体及社交媒体上的评论、未经证实的估计和泄露信息。

我们标记、调查和维护我们的数据集。
**分析师研究，增强** 通过一个专有系统扫描、爬取和整合见解。

## 2 置信度

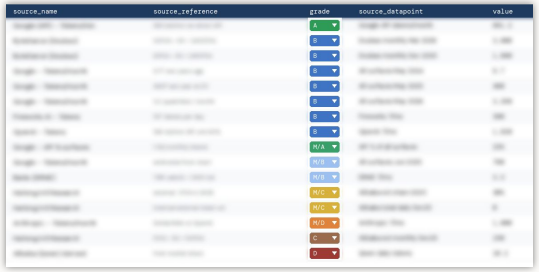
信心评分前
它算数。

每行都承载着严谨
**置信度** 在进入任何模型之前，以便弱输入不能虚增数量。

We
**等级已填数字最高**
其他主要资料之上，证实了3
第三方估计和单源声明。

所有派生数据
**继承最低评级**
从输入来源。

**审计追踪：示例源表**



## 3 建模和三角测量

公司财务模型进行自上而下验证

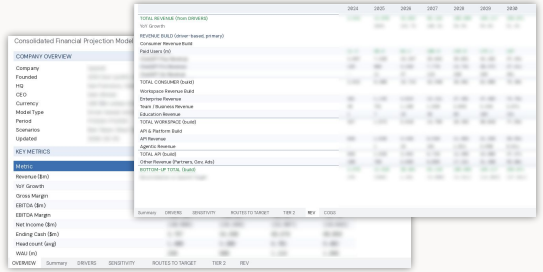
我们建造
**全公司模型**

专门针对生成式AI金融业务（从总收入报告中分离出来），涵盖收入的关键驱动因素、盈利能力和成本。

the
**损益表、现金流量表和资产负债表**

**独立代理**
硅片（芯片制造商收入），建造成本，产品组合，行业研究，交通与产能。

**审计追踪：样本公司收入模型**



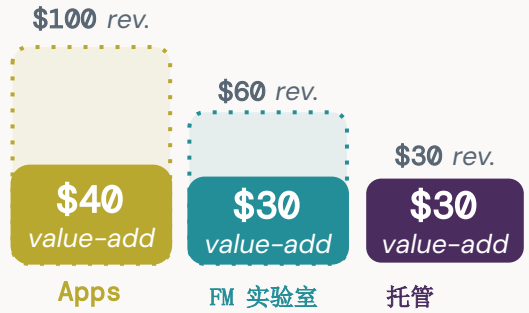
## 4 去重

仅计算一次

收入在每个层级计算，但从未跨层级汇总：通过增值额进行归属，因此相同的一美元不会被重复计算。

或重复计算三次。
e.g.
**100美元应用内消费**

**$100, not $190 :**



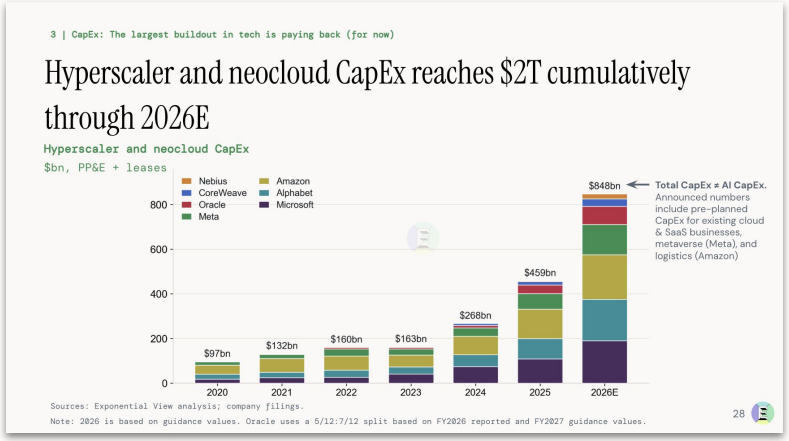

范围：全球除中国外 • 应用、模型及基础设施收入计入 • 不包括芯片、人工智能广告提升、传统软件功能和融资。

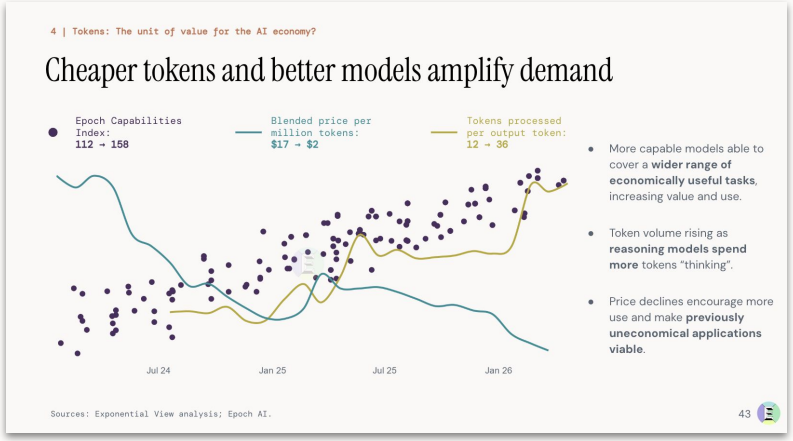
# 顶行

人工智能的需求已比以往的平台转型更清晰地得到了实际收入的验证。生成式人工智能生态系统的收入已经超过了 1750亿年化
（在从供应商收入中消除重复计算之后）

资本支出强度已远超历史大型科技股的常规水平，以支持人工智能的规模化部署，而第三方融资正日益成为融资组合的一部分。

降低人工智能的成本能否产生足够的规模和利润来支撑其扩张。

# 内容

# 是真需求大且快

收入由真实外部客户驱动。该领域的发展速度比之前的任何IT浪潮都快三倍。这种需求催生了一个计算超级周期：计算能力增加了10倍，新能源发电，数据中心规模扩大，以及因供应无法满足而不断累积的积压订单。

# 110亿美元12个月的累计收入——现已达到175亿美元的步伐

## 生成式人工智能经济收入，去重后

$bn/year, Jan 2023 - Jun 2026


指数观分析。
注意：全球除中国大陆地区外。去重后的应用程序、基础模型和基础设施托管收入。不包括芯片制造。

# 真实的外部需求驱动着人工智能收入

。

展示我们在不同服务商之间进行收入建模和去重的示意图

客户

*支付应用程序许可证/费用*
*，或者*
*直接购买代币*

收入流下栈

**应用层**

**模型实验室的应用程序**

*克劳德代码，科德克斯*

增值：收入减去代币成本

**独立生成式AI应用**

*光标，OpenRouter，哈维*

增值：收入减去代币成本

**基础模型层**

**封闭权重模型**

*e.g. Opus 4.8, GPT-5.5*

增值：收入减去推算成本

**开放式重量模型**

*Deepseek-v4, MiniMax M3*

增值：0美元（收入包含在托管服务中）

**托管层**

*Azure：微软AzureCoreWeave：CoreWeaveNebius：Nebius*

AI收入（完整）

**不计入**

*非AI母语应用* *（按消耗的代币数计算）*

*芯片销售* *（来自承载层的资本支出）*

*广告提升* *（谷歌/Meta AI广告收入）*

**我们采购、交叉验证、建模&审计以验证和去重**

*根据官方文件及第一方披露泄露，政府统计数据，第三方分析师，以及代理指标；所有来源均已按质量分级。*

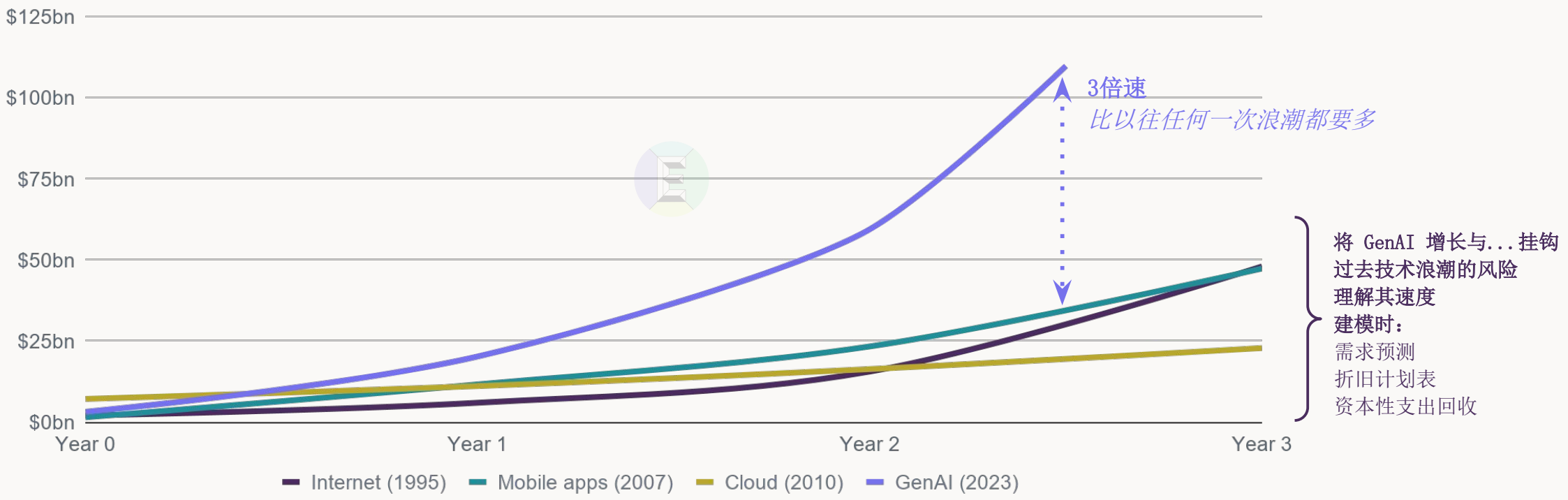
*公司层面的严格财务建模构建自下而上的去重收入模式。*

*持续扫描和爬取跨越数百维护和调整数据集所需的资源。*

# 人工智能的发展速度是任何IT浪潮的三倍。

**实现收入轨迹与零年时间对齐**

每年数十亿美元，经通货膨胀调整

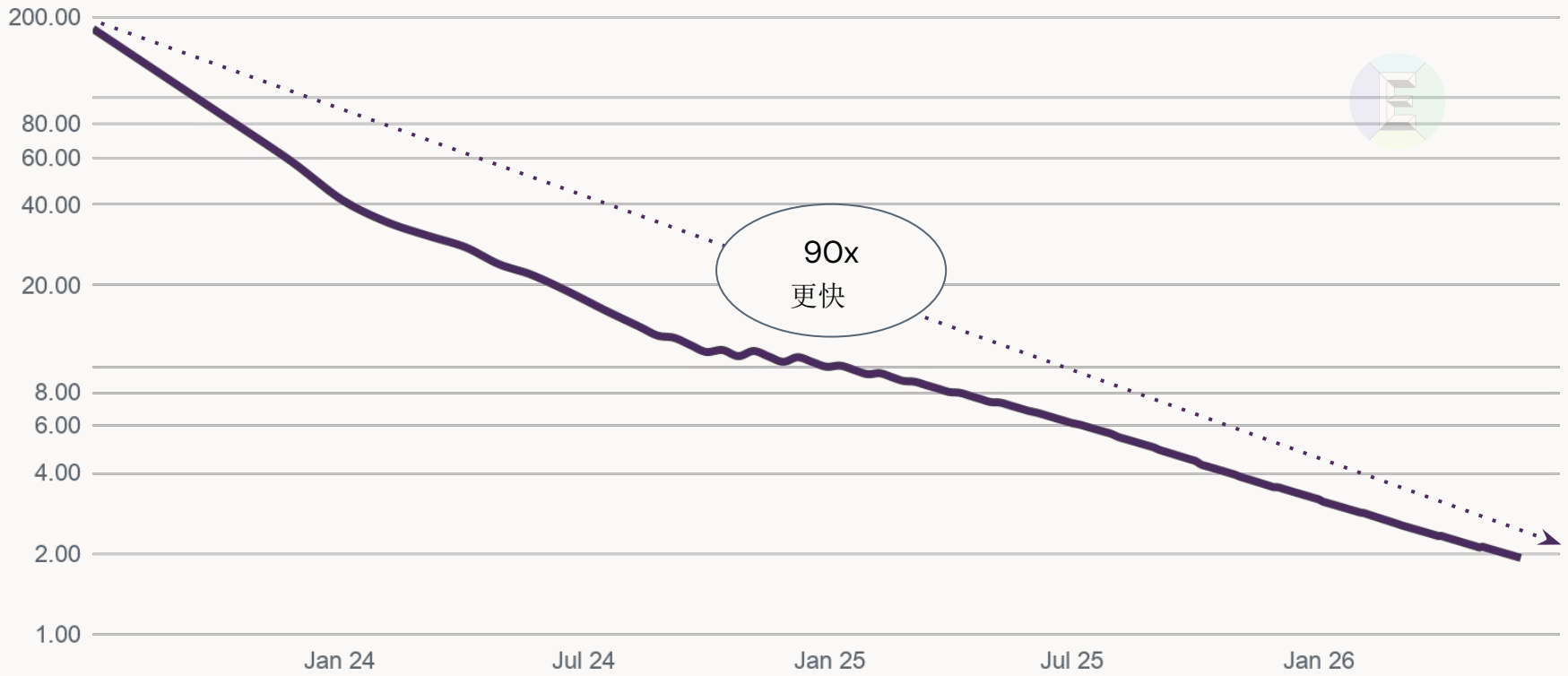



来源：指数视图分析；美国商务部；公司申报；UBS；美国劳工统计局。注：我们使用收入的第一完整年，因此从2023年1月开始进行生成式人工智能的衡量。

# 每一新增的10亿美元收入，其到账速度都更快。

**增加10亿美元额外累计收入的时间到了**

天，对数刻度




2023年，人工智能产业累计营收达10亿美元。

需要180天来添加

现在需要不到两天。

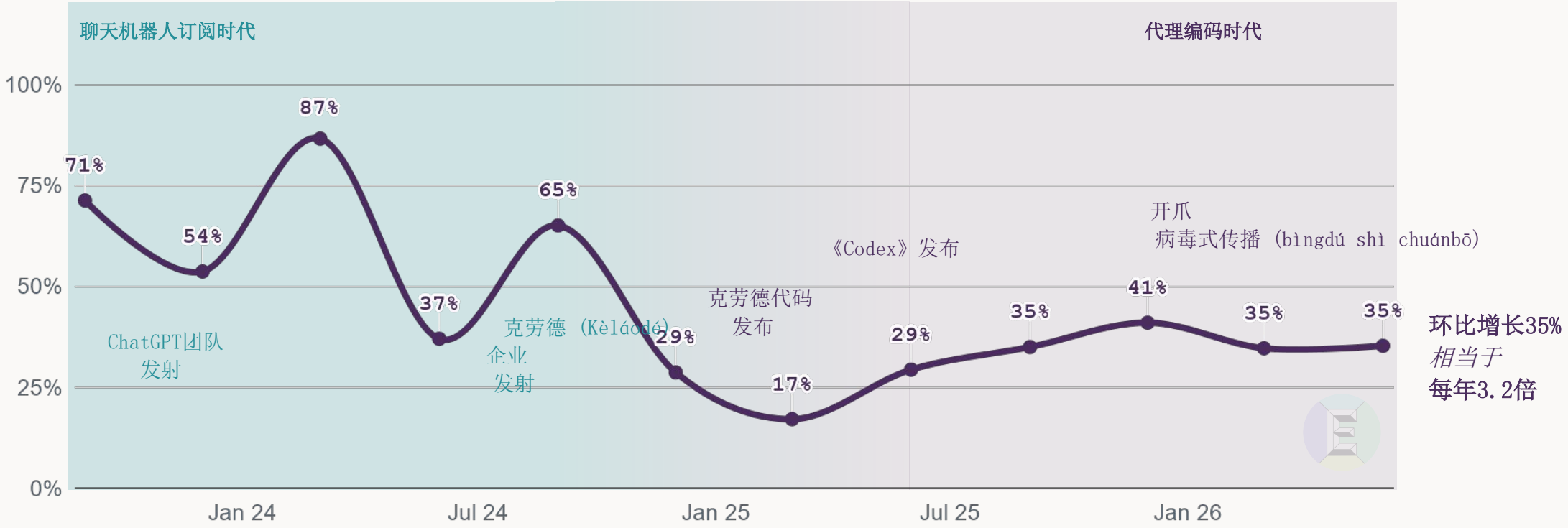
指数观分析。

注意：全球除中国大陆地区外。去重后的应用程序、基础模型和基础设施托管收入。不包括芯片制造。

# 增长在每个采用阶段都得以保持

**环比收入增长**

环比变化率




**环比增长35%**

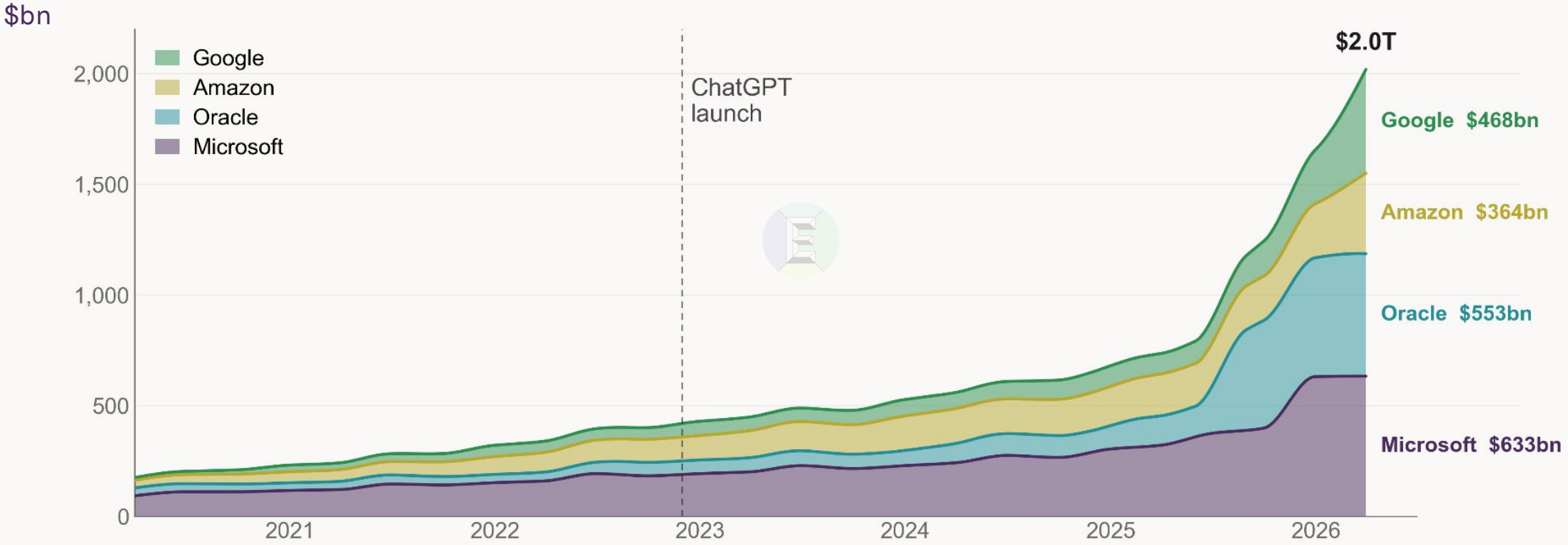
*相当于*

**每年3.2倍**

指数观分析。

# 这种快速增长的需求正表现为合同积压。对于超大规模企业

组合超大规模企业积压订单（剩余履约义务）



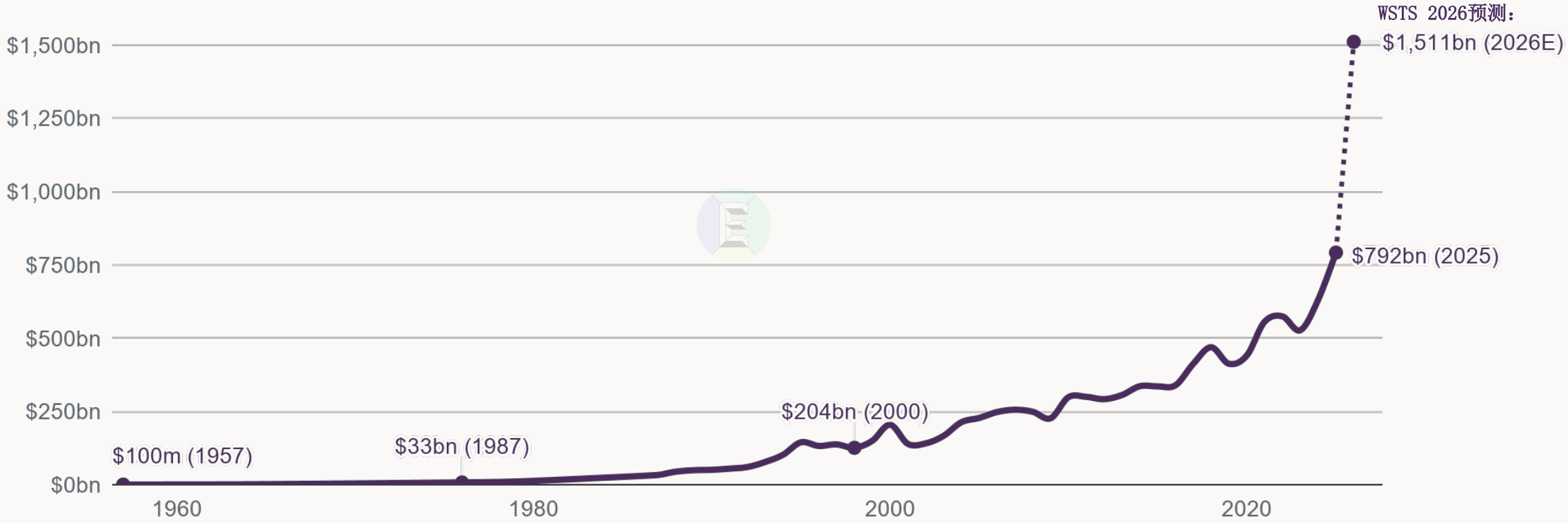

来源：指数观点分析；公司文件。

注意：微软=总RPO（含M365/Dynamics）；亚马逊=公司总RPO（主要为AWS）；谷歌=收入积压（主要为云业务）。甲骨文季度结束时间早一个月。

# 需求已启动计算超级周期。

全球半导体市场收入

$bn/year



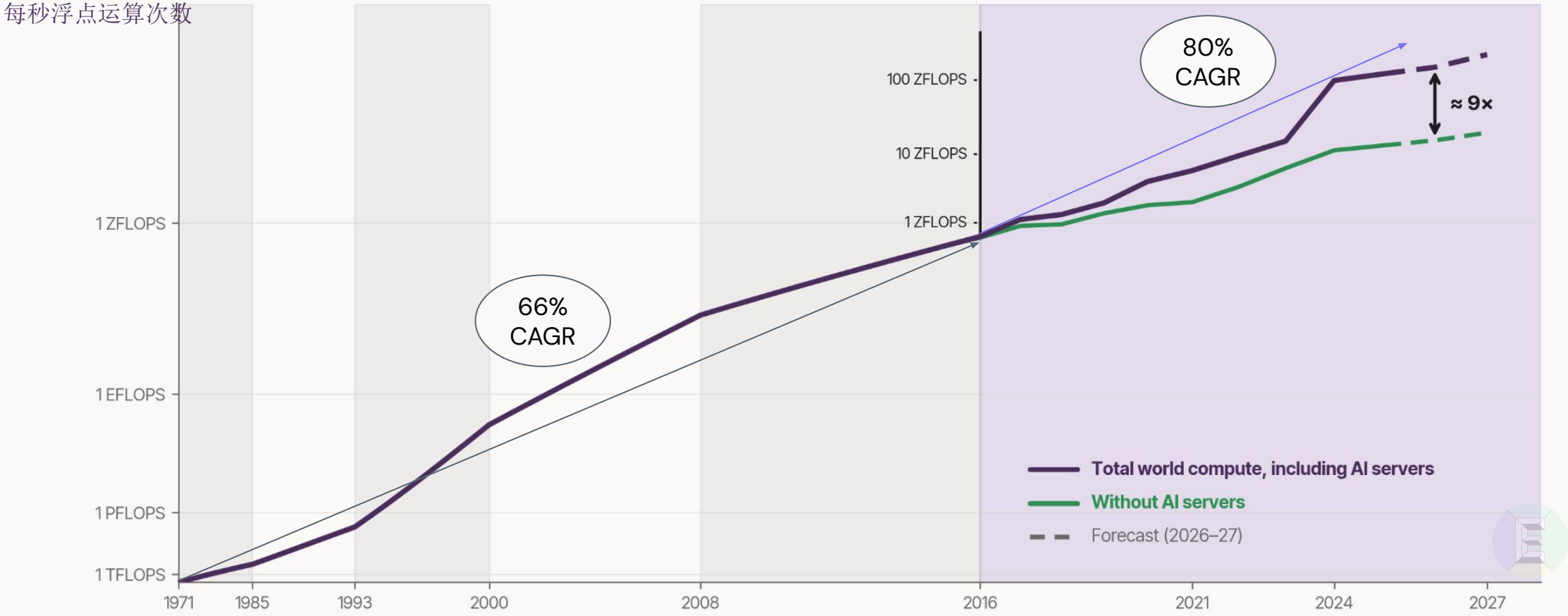

来源：指数观分析；世界半导体贸易统计；日本半导体历史博物馆。

注意：名义上美元

# 人工智能推动了计算增长长达50年的趋势的回升。

自1971年以来的全球计算

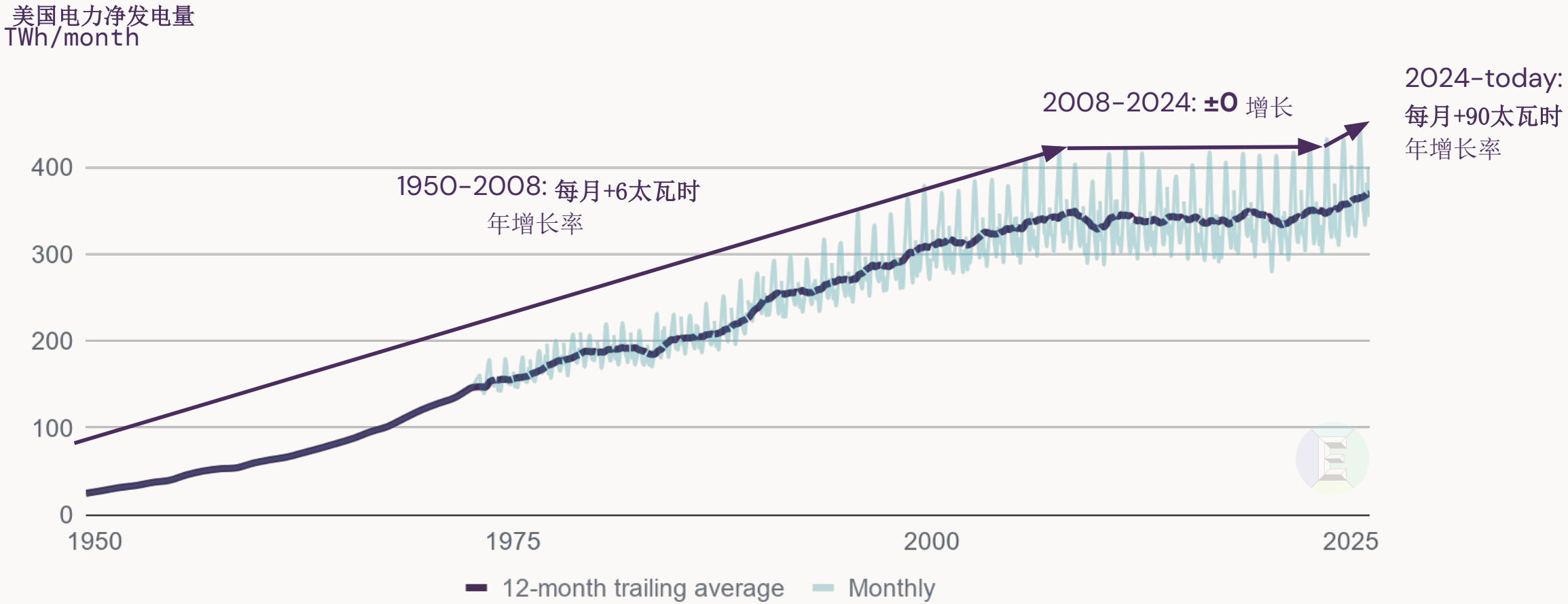
每秒浮点运算次数




指数观分析。

注意：包括大型主机、小型计算机、个人电脑、服务器、智能手机、物联网和人工智能计算。人工智能服务器FLOPS根据英伟达GPU的安装基数按代别计算，从Hopper开始采用FP8。

# 人工智能需求正重燃美国衰落的电力部门

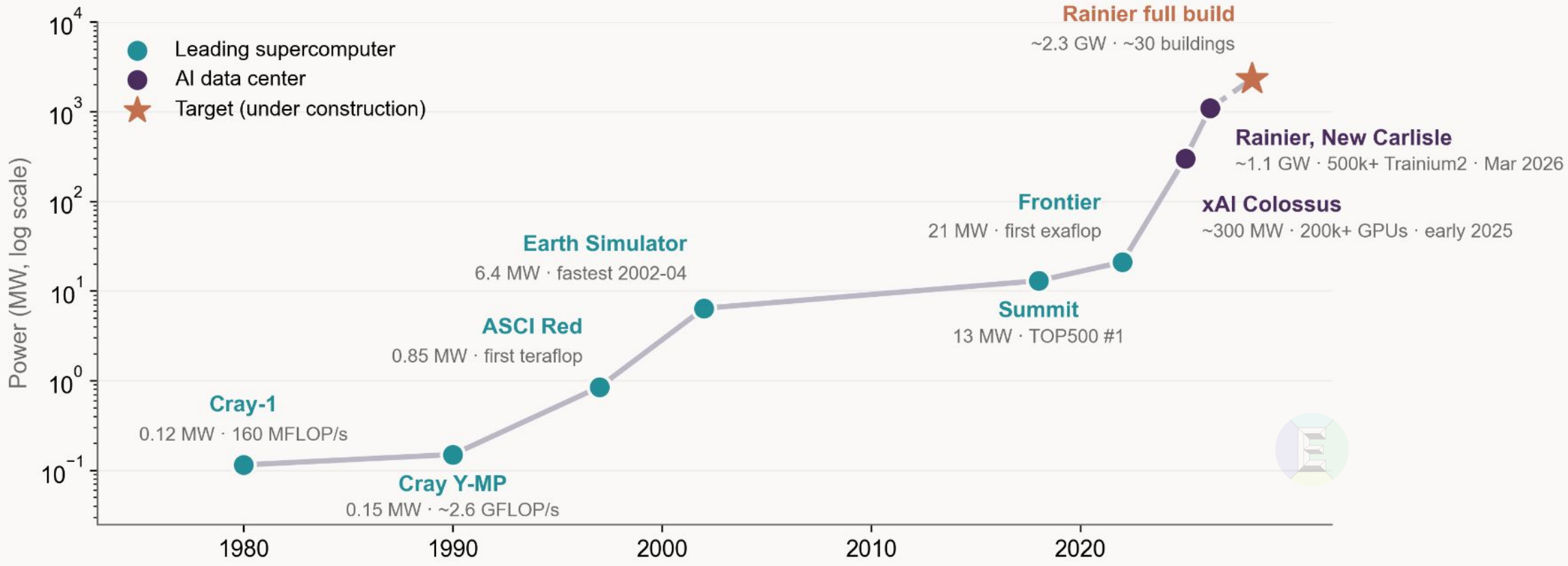



来源：指数观点分析；美国能源信息署。

# 最大数据中心的规模在四年内增长了50倍。

**最强大计算机随时间的力量**

MW，对数刻度



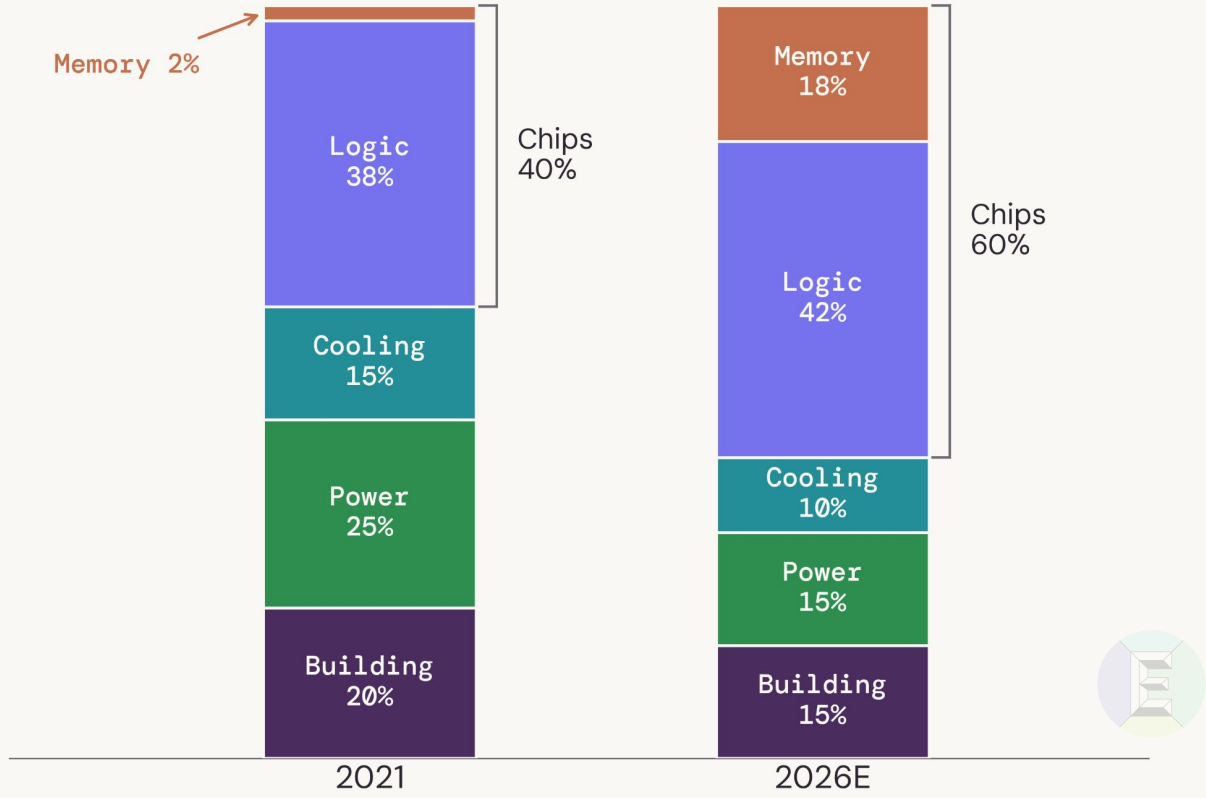

来源：Exponential View分析；TOP500 / Green500；Epoch AI；OpenAI；Oracle。

注意：Rainier完全构建是一个目标。

# 内存和计算现在占据了数据中心建设所花费的每一美元中的大部分。

按组件划分的数据中心建设总成本占比，2021年与2026年预测值

%




- 每一美元的新购买力 **更多硅，更少混凝土**

薯片所占的支出份额增加了50%（从40%上升到60%）。

- 记忆

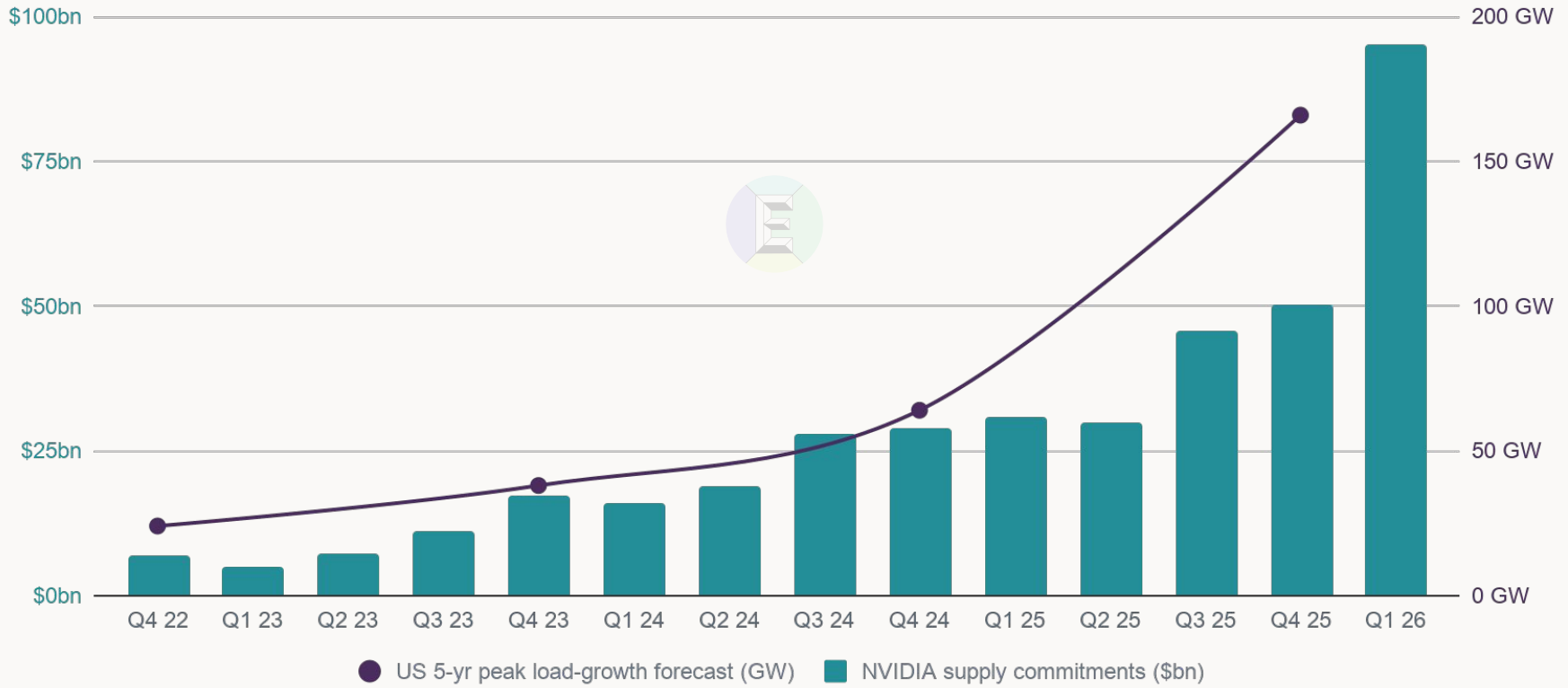
这是最大的变动因素，从2%的舍入误差到~18%。

来源：Exponential View分析；高盛；Epoch AI；Semi Analysis
注意：由于四舍五入，数字可能无法加总至100%。

# 由此带来对计算能力和能源的日益增长的承诺

**计算与电力承诺**

英伟达供应承诺（十亿美元，左轴）与美国家用负荷增长（吉瓦，右轴）




- 英伟达的供应承诺已增长至 **$95bn**，过去一年 **$31bn to**
- 到2030年，美国电网预计所需的额外电力 **自2022年以来增长了约7倍**（24吉瓦→166吉瓦），其中数据中心约占这一增长额的55%。

来源：Exponential View分析；Grid Strategies；Nvidia提交的文件。

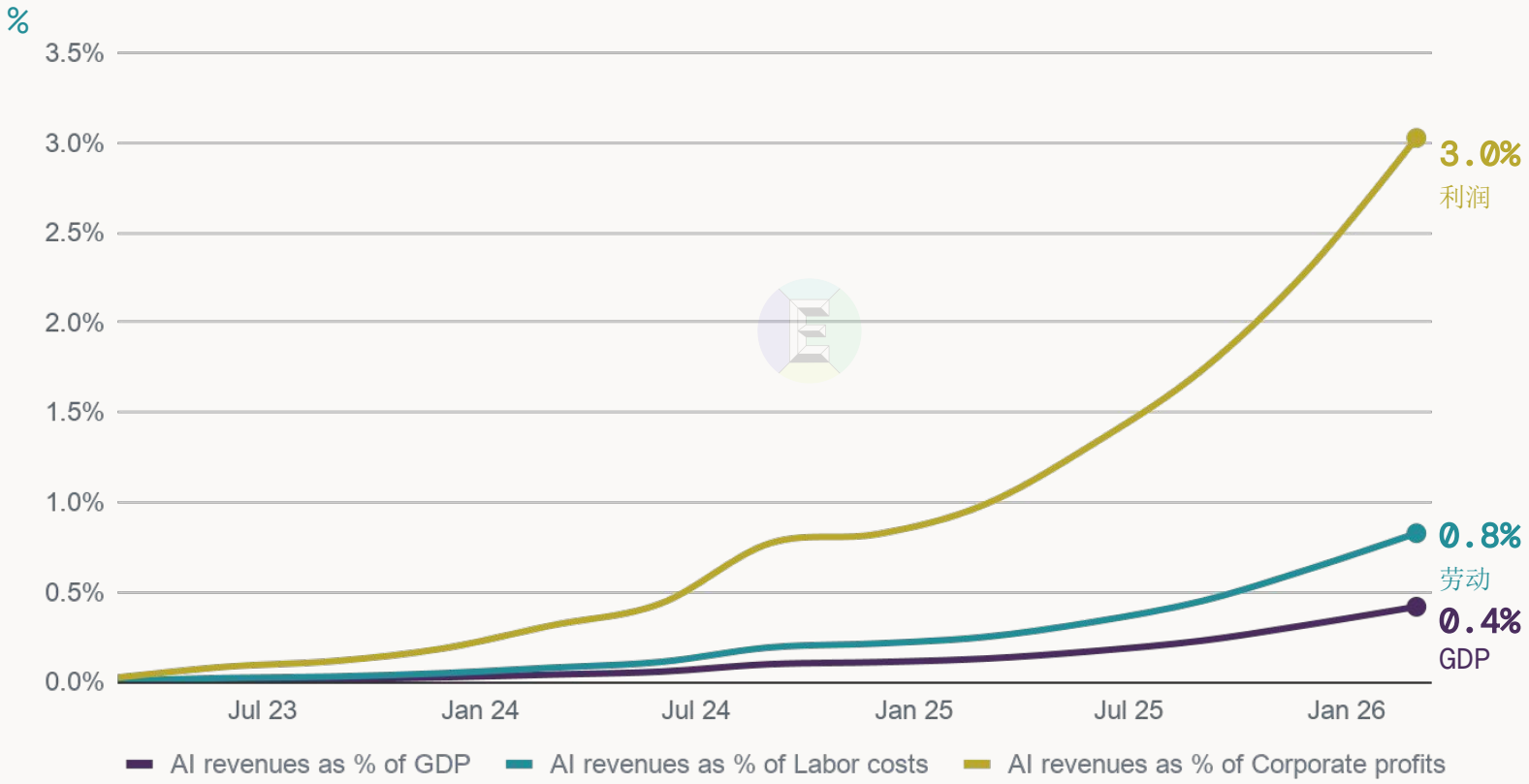
# 2. 仍早且早

即使是最大的企业支出者，人工智能在损益表上也是微不足道的零头。

它看起来仍然很早。虽然组合正在变化，但举措一直聚焦于效率与成本节约。而且，测量的收入可能低估了社会收益，因为消费者报告了那些尚未在数据中体现出来的好处。

# 相对于GDP，人工智能的收入仍然只是个微不足道的零头。

全球人工智能收入（例如中国），相对于美国GDP、劳动力成本和公司利润



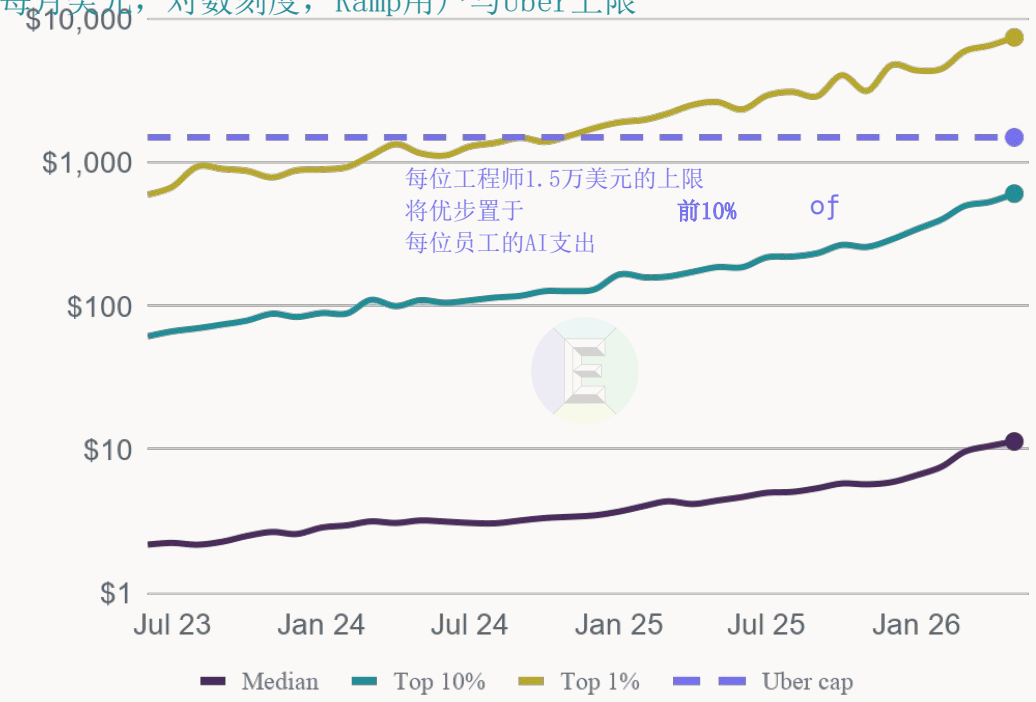

- 依然渺小：人工智能产业收入相当于美国GDP的0.42%（而信息技术产业为9.4%）。
- 即使是慷慨的衡量标准（公司利润）也32倍大 超过所有生成式人工智能收入。
- 还早：人工智能收入相对于GDP有所增长 3倍 vs 2025年第一季度(0.13%)，以及10倍 vs 2024年第一季度(0.04%)。

来源：指数观分析；圣路易斯联邦储备银行

# 在公司层面，人工智能支出仍然相对较小：例如，优步每名工程师1.5万美元的支出仅对损益表造成微不足道的影响。

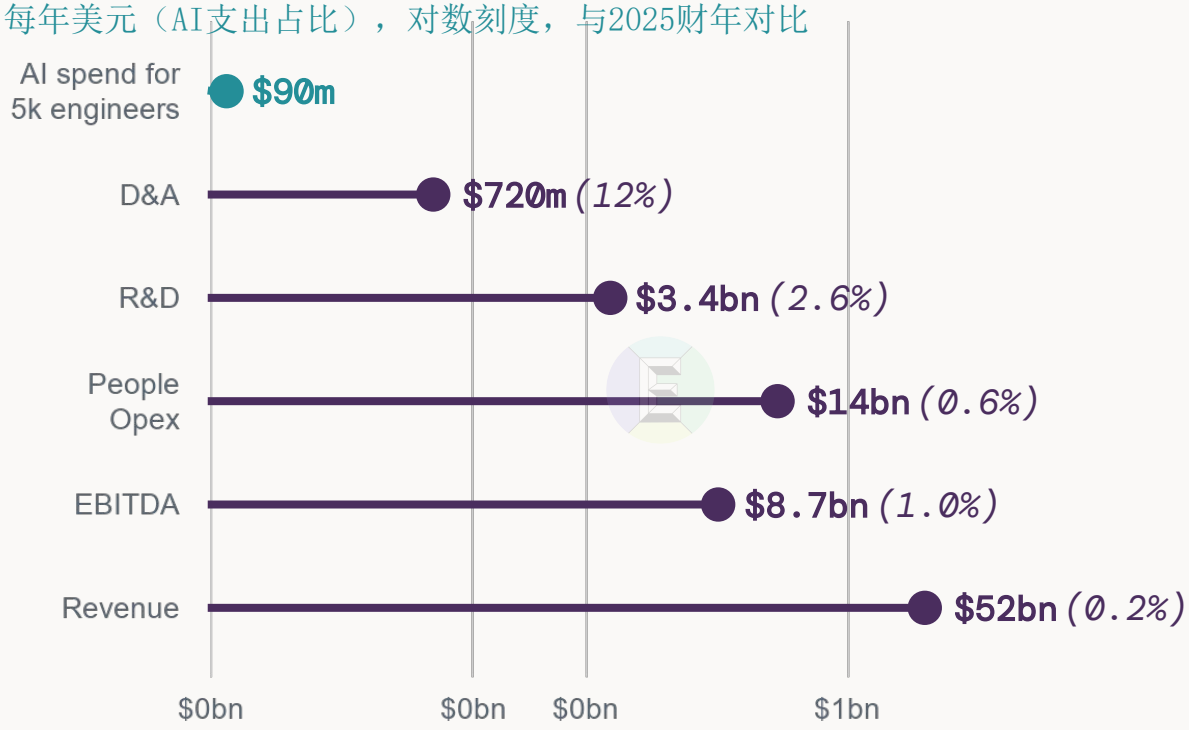
## 每位员工的AI支出

每月美元，对数刻度，Ramp用户与Uber上限




## 优步人工智能支出（上限）与损益项目

每年美元（AI支出占比），对数刻度，与2025财年对比




来源：指数观分析；Ramp经济学实验室（n=7万美国企业），优步（Uber）提交的文件。注：前1% / 前10% / 中位数是根据AI支出水平相对于Ramp客户群进行比较来定义的。优步的数字是每位工程师的最高限额，与Ramp每位员工的AI支出基准进行比较。

# 与以往通用技术类似，一些收益可能无法被GDP衡量。

## 人工智能的影响

### 消费者剩余

直接惠及大众且近乎零成本的价 tri。销售量甚微，因此GDP未能充分反映消费者利益，例如：

● 软件与服务购买免费更换 ● 学习、休闲与便利

### 生产者剩余

已售商品和服务中蕴含的价值。交易额更多。 , e.g:已记录在GDP中

● 具备AI功能以驱动收入● 更快（有价值）的发布● 服务型公司利润

## 历史案例

### 1880-1920: Electric lighting

直接GDP影响 ≈$0

~99.97% cheaper

~40,000x more

光变成了一小时的工资可以买到价格并未录得这一涨幅。

Nordhaus (1996)

### 2000-2020: Free digital goods

直接GDP影响 ≈$0

免费搜索、百科全书和地图取代了付费服务。仅搜索本身就值~$17.5k/yr/person.

Brynjolfsson et al. (2019)

### 1850-1870: Steam

+0.4 年度人均美元GDP年增长率

机械化工厂和铁路，使得一个工人能够生产并运往市场的东西远超以往。

Crafts (2004)

### 1980-2000: Automation

+0.37 年度人均美元GDP年增长率

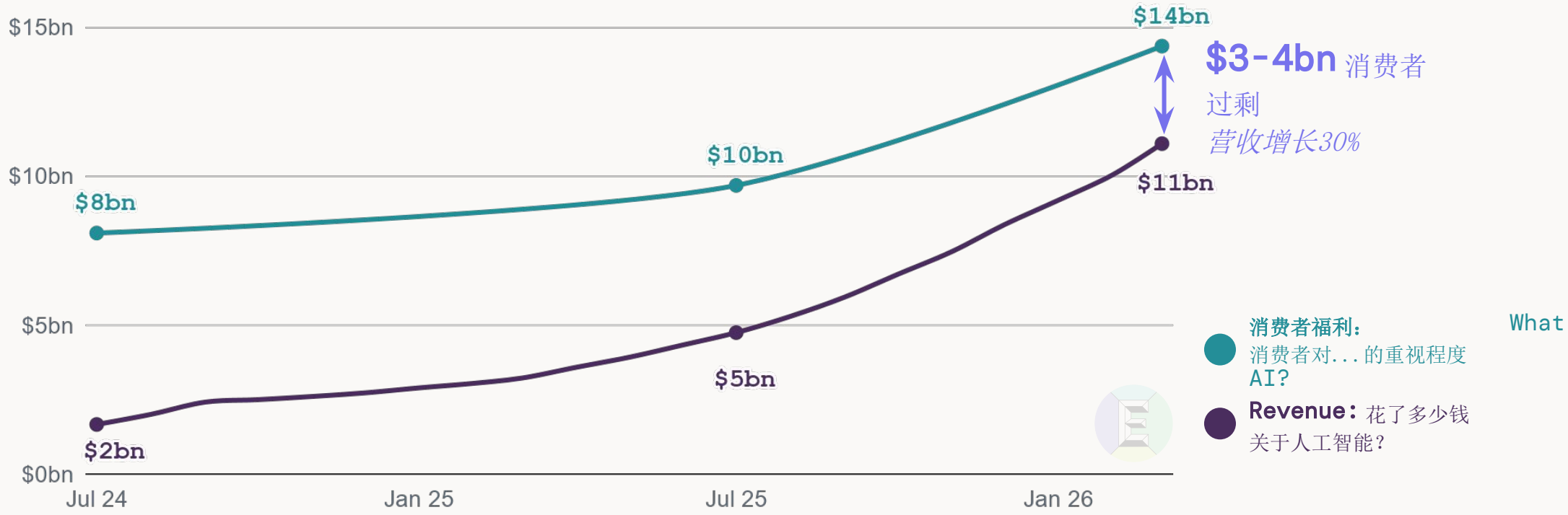
可编程机床能降低每个单位的劳动成本，从而提高每个工人的产量。

Graetz & Michaels (2018)

# GDP衡量一切价格，却不知任何价值。AI的经济价值超过衡量到的收入。

日度生成式人工智能收入与美国家庭福利

$bn/month, Jul 2024 – Mar 2026



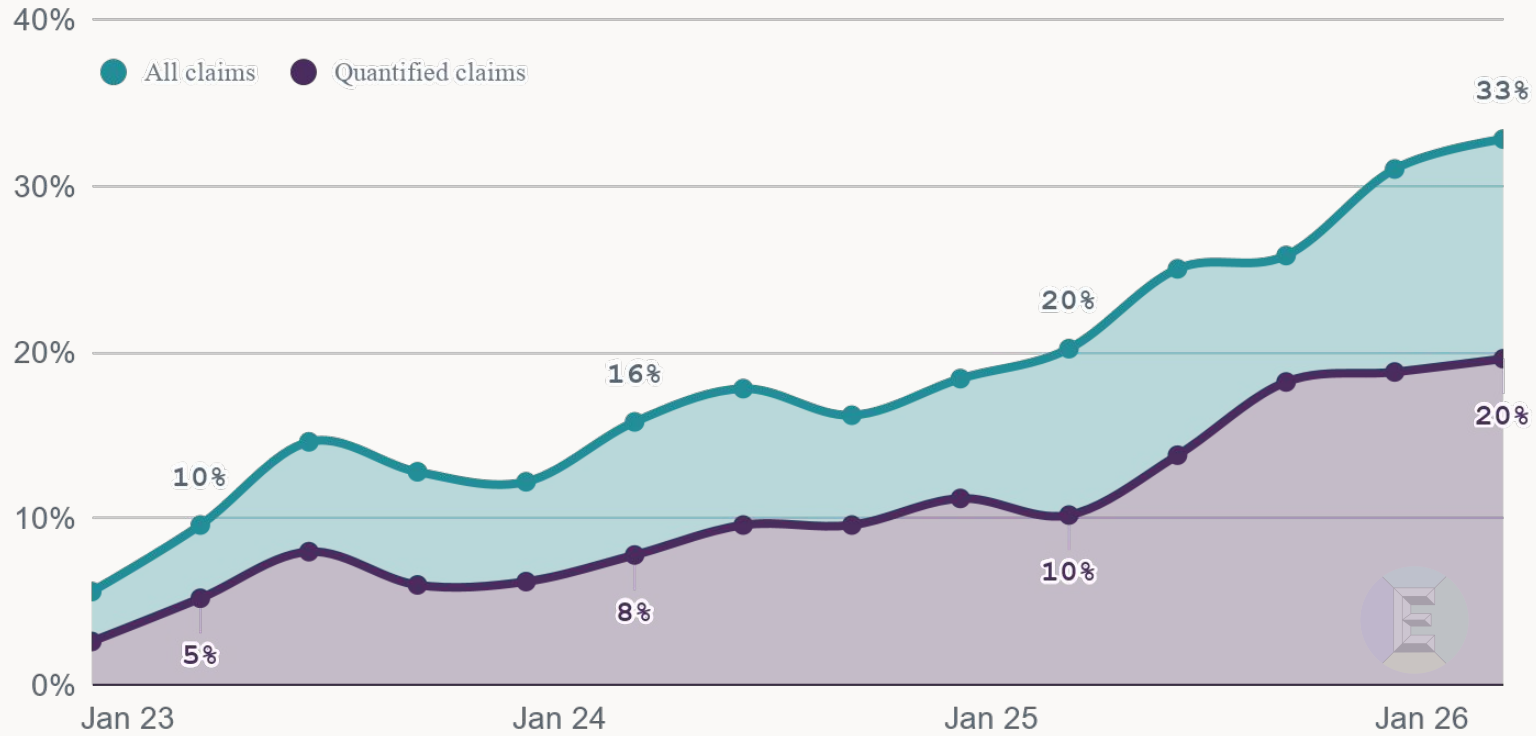

What

来源：指数观分析；斯坦福数字经济实验室

注意：福利价值根据对“你是否愿意放弃使用所有AI工具，如ChatGPT、Gemini、Claude或Copilot”这一问题的回答确定。从明天早上开始，为期一个月，用以换取[美元]？” 收入包括全球（不含中国）消费者和企业支出。

# 上市公司正报告称，生成式人工智能的影响日益增加。

在收益电话会议中声称人工智能对盈利有影响的公司
S&P 500, Q4 2022 – Q1 2026

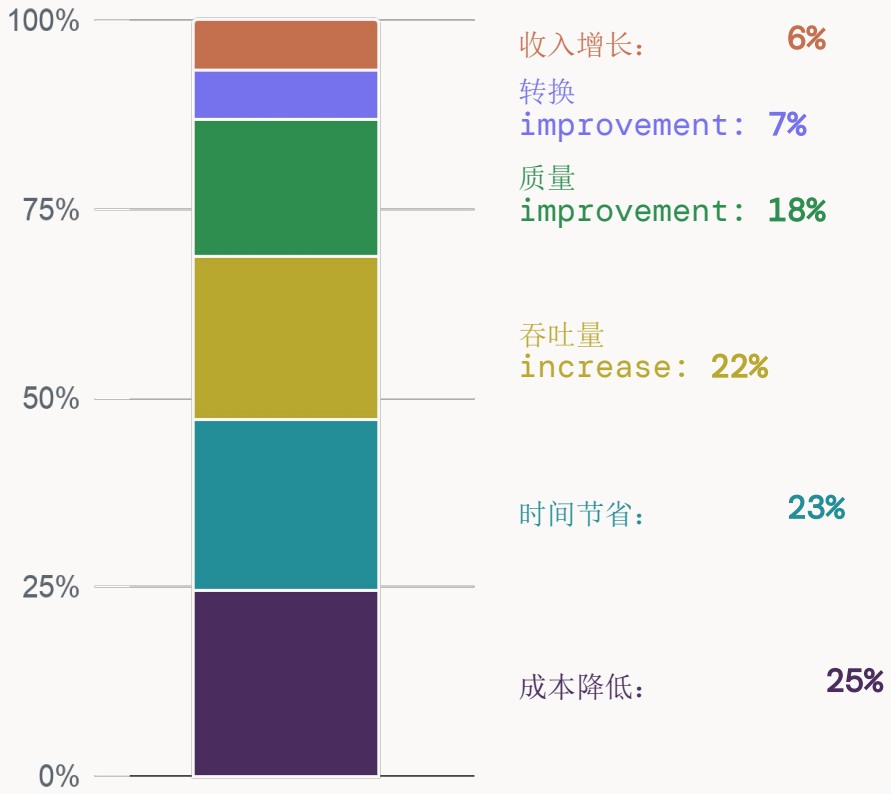



- **越来越多的关注：** 企业将人工智能视为提升盈利能力的机遇。我们追踪到自2023年以来，标普500指数成分股中提及人工智能影响的频率增长了3-4倍。
- **给它编号：** 目前有50-60%的索赔已被量化，但尚待确定这些索赔对公司最终财务状况而言有多大影响以及有多重要。
- **仍然是少数** 大多数公司尚未报告其使用人工智能所带来的量化影响。

来源：指数观点分析；财报电话会议。

# 十分之七的生成式人工智能（GenAI）主张集中于节省成本

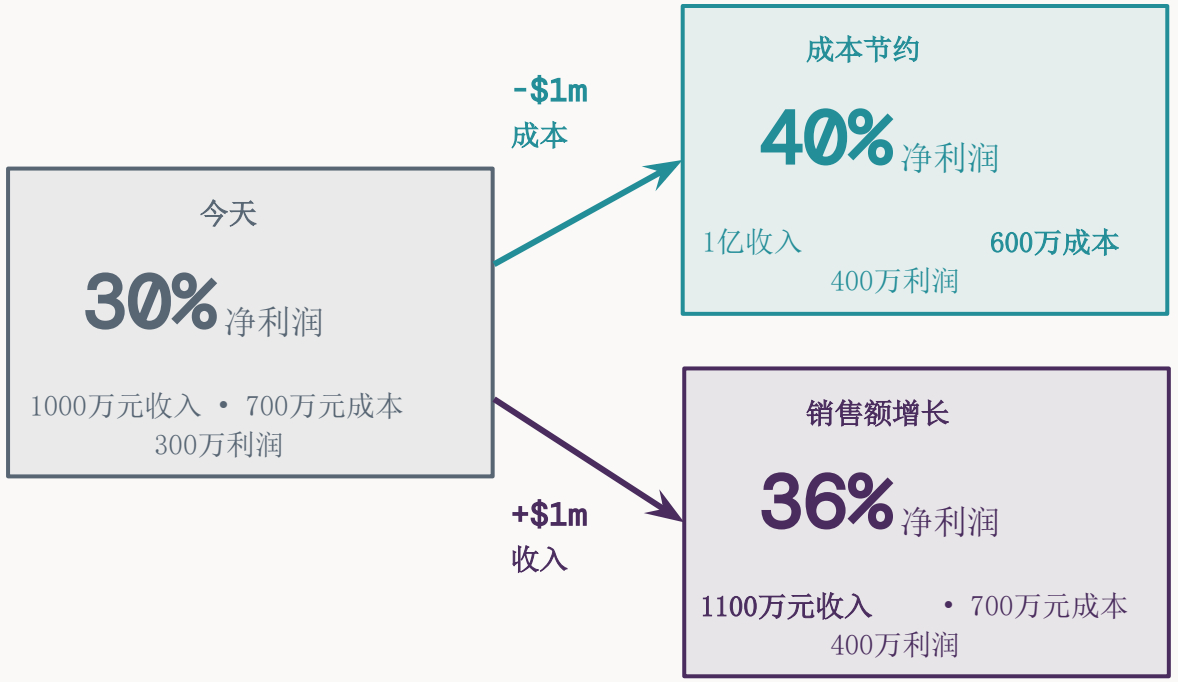
声称的人工智能结果
S&P 500, Q4 2022 - Q1 2026




来源：指数观点分析；财报电话会议。
注意：由于四舍五入，数字可能无法加总至100%。

为什么初始项目优先考虑效率，一个说明性的例子：

相同的一百万美元影响同样具有
4页 通过节省获得更高的利润率 与销售增长

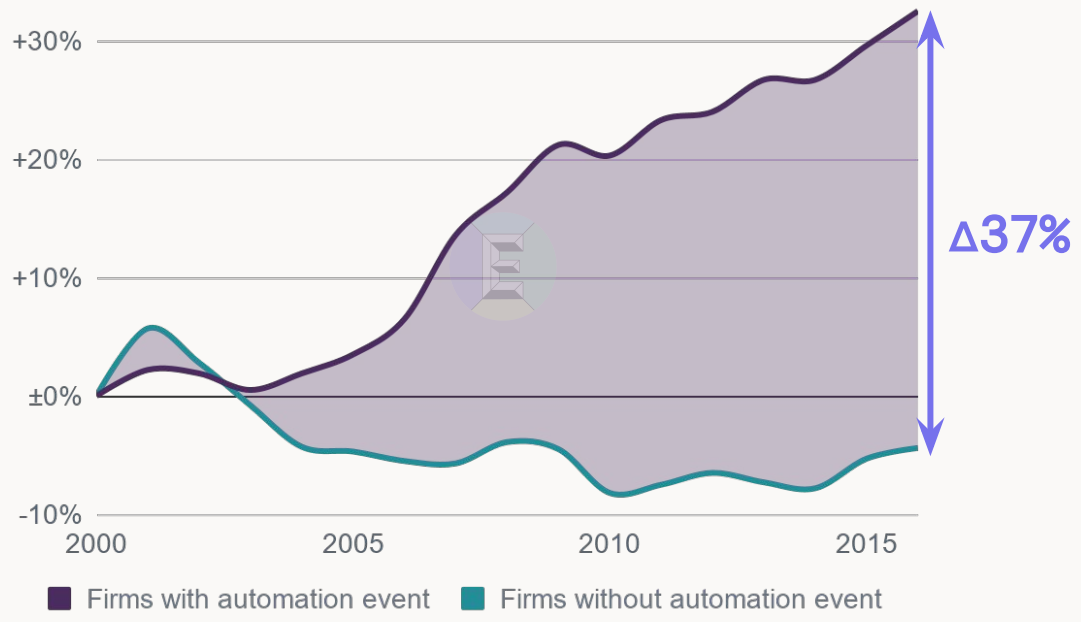



注：为说明起见，收入以0成本加入。若加入销售成本，将进一步抑制利润率的增长。

# 与以往几波类似，早期采用者正逐渐超越他们的同龄人。

**历史案例研究：**

企业层面就业（含/不含自动化）百分比变化 2000-2016

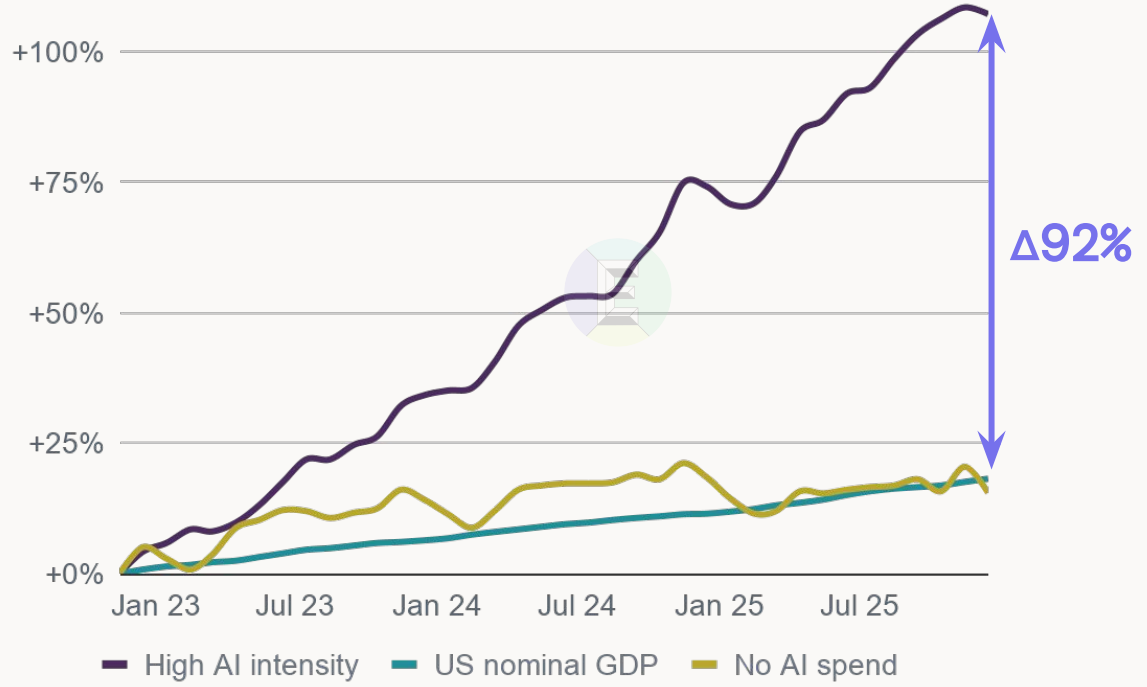



来源：指数观点分析；Bessen, Goos, Salomons & Van den Berge（2020）

自动化公司里的员工会怎样？

**今日的AI经济：**

2022年11月以来，高比例使用AI与未使用AI的营收增长率对比变化




来源：指数观分析；斜坡经济学实验室（n=7万美国企业）。注：高强度=按收入占比排名前25%的AI支出者。

# 3 | 资本支出

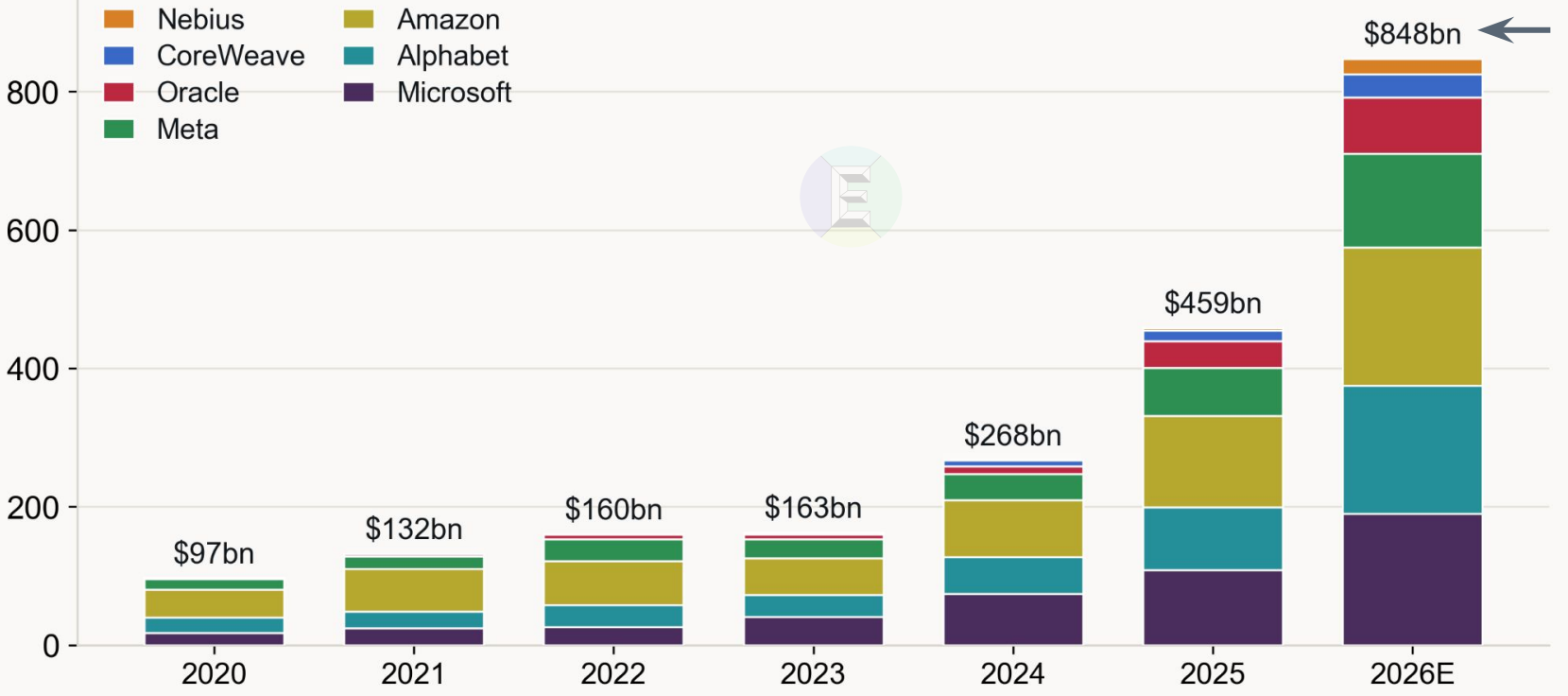
## 科技要企最共规模的建设正在（暂时）获得回报

超大规模云服务商和新型云服务商已承诺到2026年累计投入2万亿美元资本支出，这给不断增长的营收带来还本付息的压力，尤其是随着更多资金由外部资本提供。这些经济规律为数据中心和代币生产的财务状况定下了基调。

# 超大规模云服务商和新型云的资本支出到2026年预计累计达到2万亿美元。

**超大规模云服务商和新型云资本支出**

百亿美元，固定资产加租赁




**总资本支出 ≠ 人工智能资本支出**

宣布的数字
包含预计划
现有云的资本支出
SaaS企业，
元宇宙（Meta），以及
物流（亚马逊）

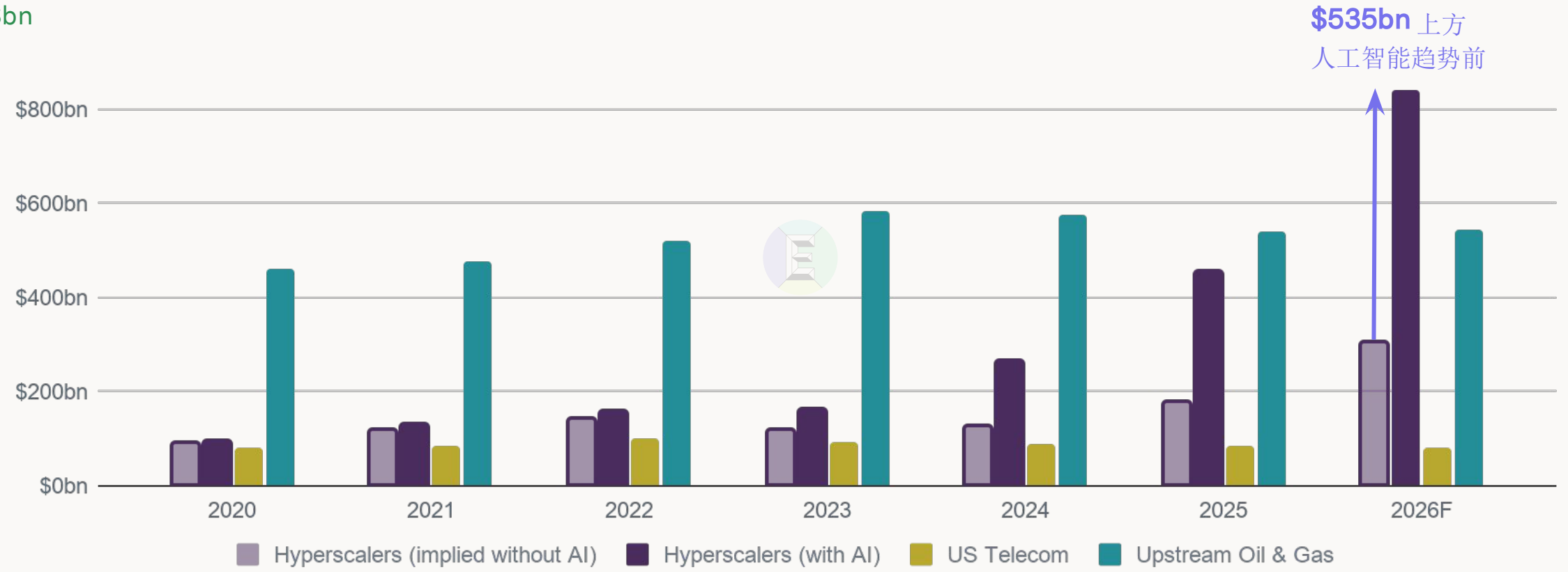
来源：指数观点分析；公司文件。

注意：2026年数据基于指导值。Oracle根据2026财年报告值和2027财年指导值，采用5/12:7/12的分割比例。

# 到2026年，与AI相关的资本支出将比AI出现前的趋势高出5

每个行业的年度资本支出

$bn

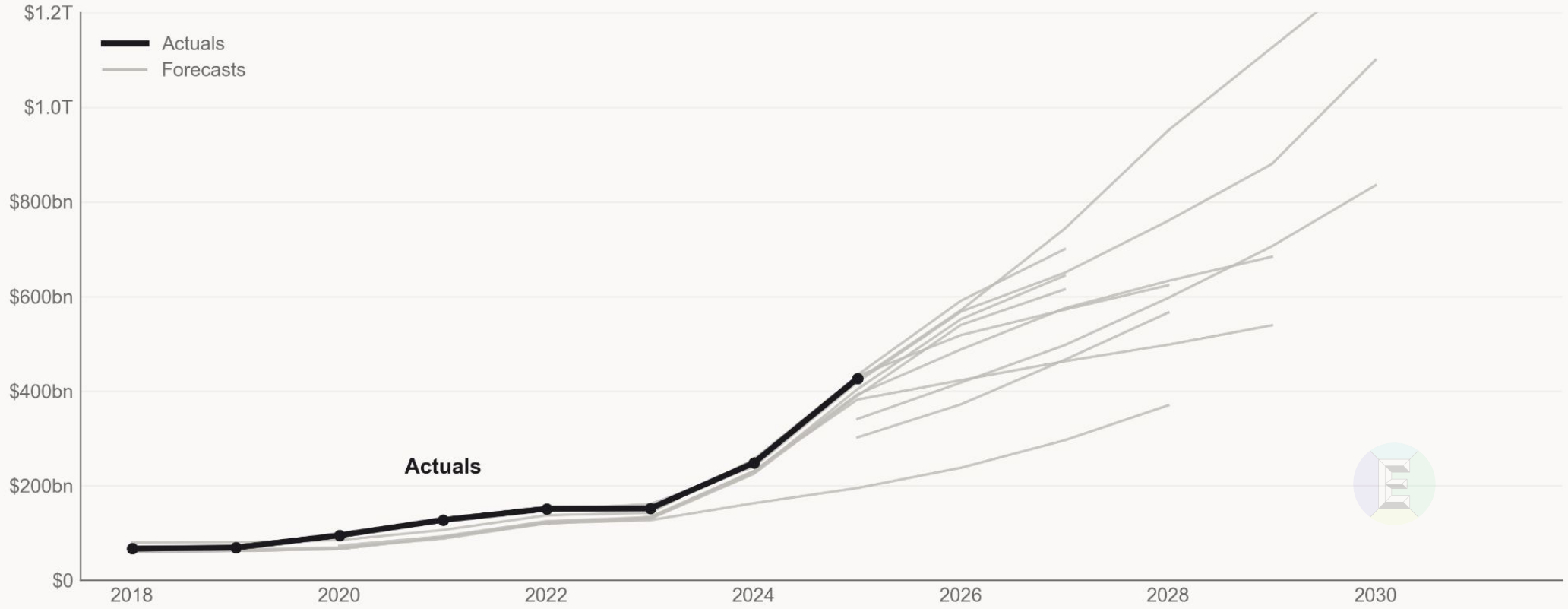



来源：Exponential View分析；超大规模企业盈利与指引；美国电信运营商及 carriers 指引；国际能源署。

# 预测已推动资本支出曲线走高。

分析师预测的超级云/AI基础设施资本支出及预测日期

$/year

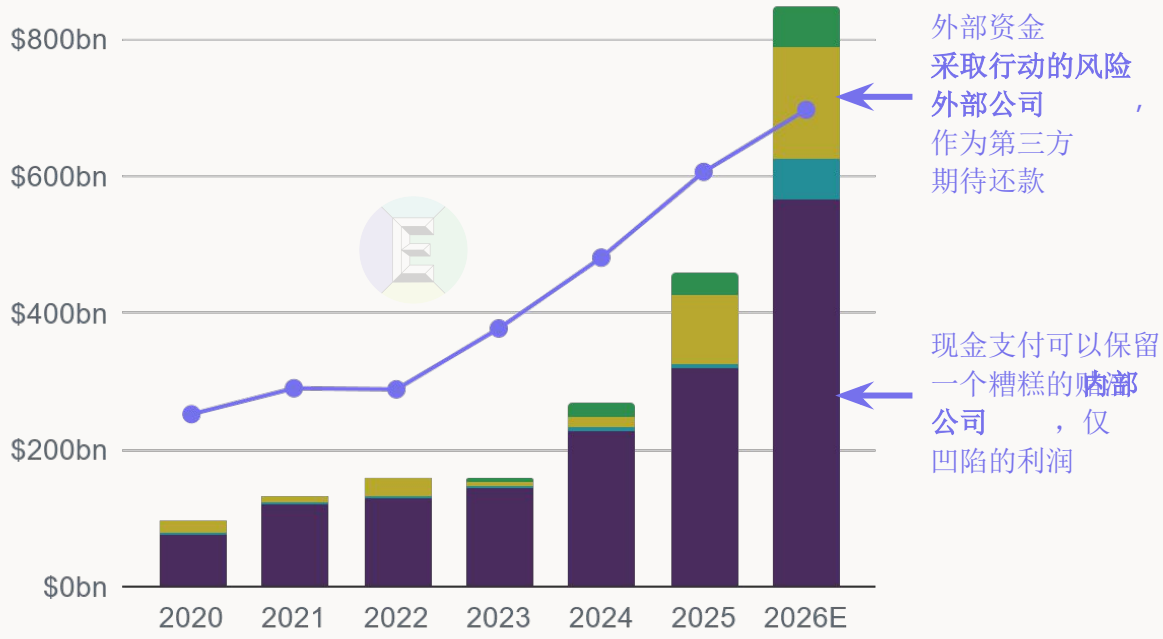



来源：Exponential View分析；巴克莱、花旗、高盛、摩根大通、摩根士丹利、新街研究、SemiAnalysis、UBS；公司文件。

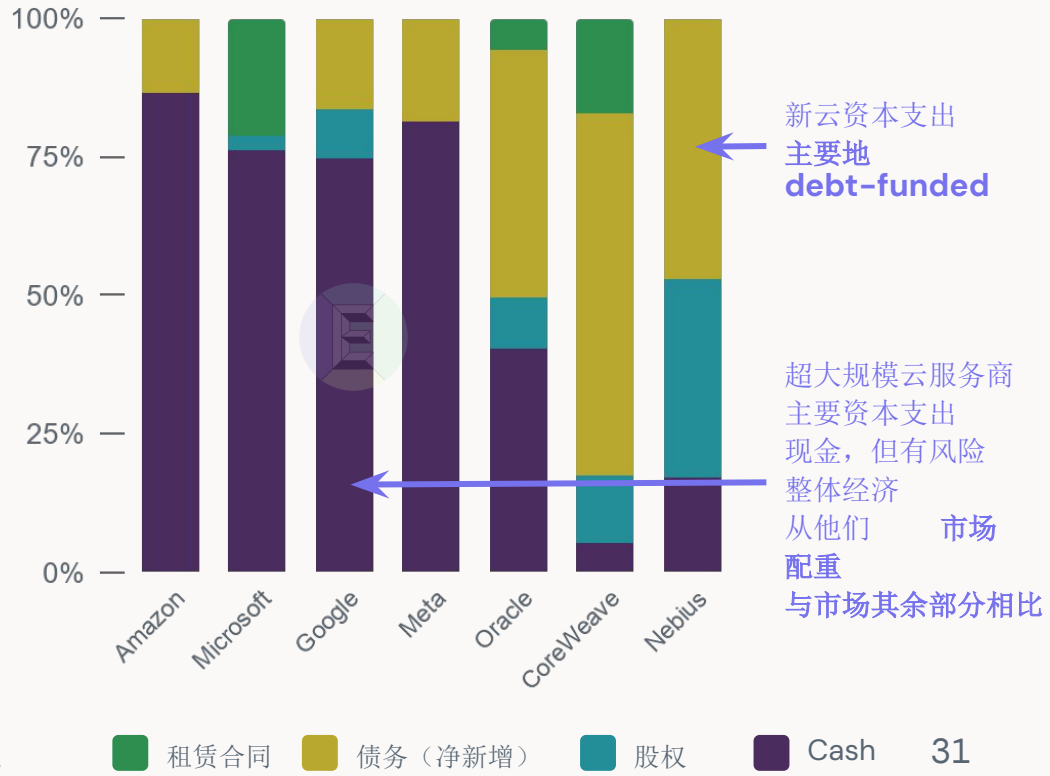
注意：预测者使用略有不同的范围（大型五家超大规模计算公司 vs 更广泛的AI基础设施）。此处“实际值”对应微软、谷歌、亚马逊、Meta、甲骨文、CoreWeave和Nebius的现金资本支出（房产和设备的购

# 边际人工智能基础设施美元正越来越多地依赖外部融资。

由资金来源划分的超大规模云服务商和新云资本支出
$, 2020-2026E



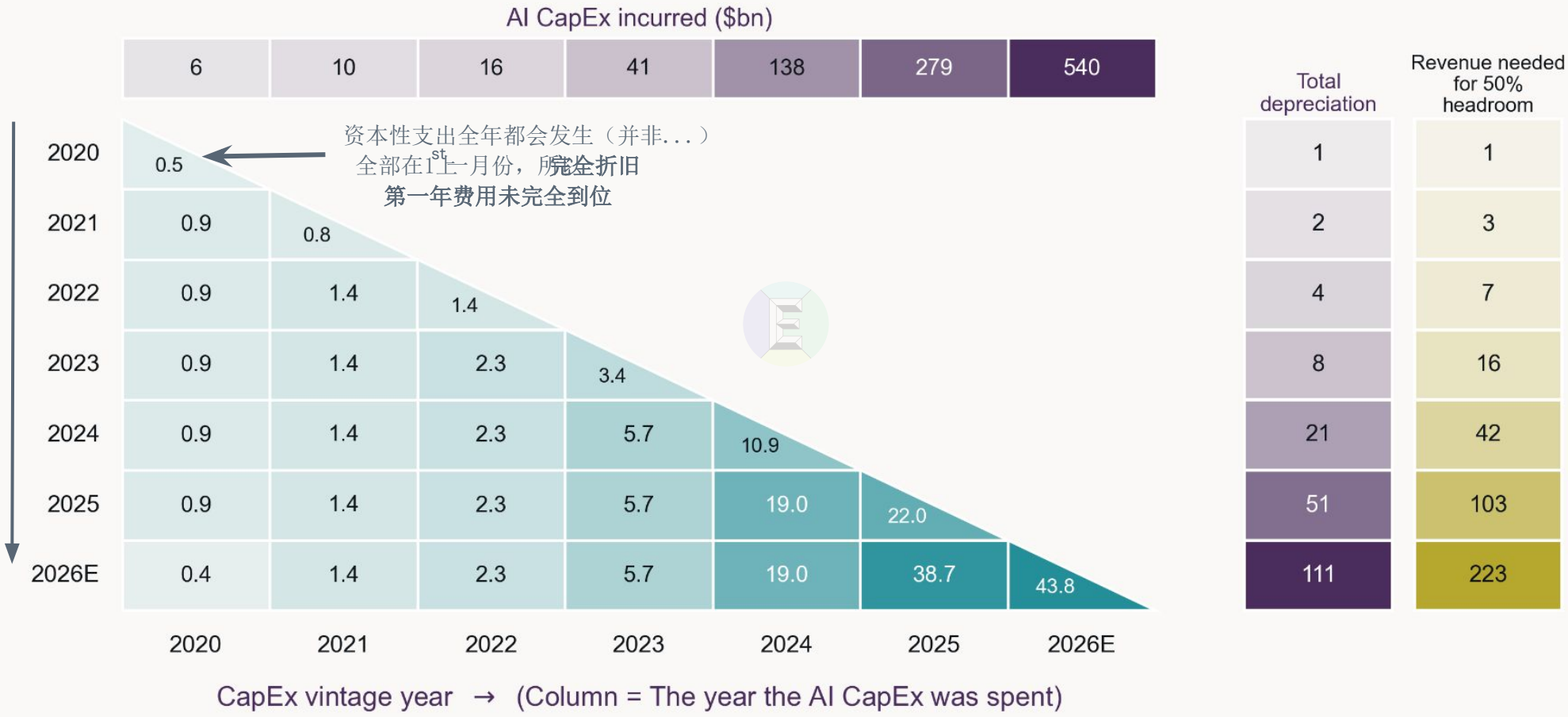

按资金来源划分的资本支出
%, 2020-2026E total




来源：指数观点分析；公司文件。
注意：债务已扣除偿还额（非发行总额）且包括所有债务工具（债券、商业票据等）。

# 2026年折旧费用接近1110亿美元

资本性支出计入费用通过折旧资产的使用年限生命。所**成本是传播**并且不需要得到认可瞬间



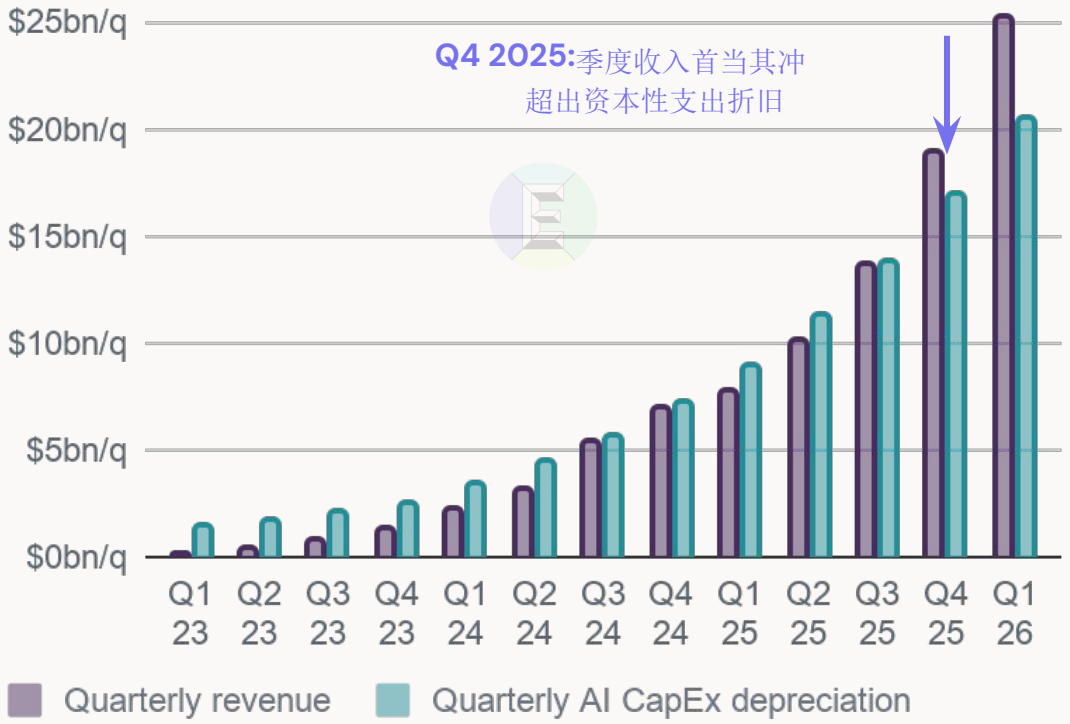

| Total depreciation | Revenue needed for 50% headroom |
|---|---|
| 1 | 1 |
| 2 | 3 |
| 4 | 7 |
| 8 | 16 |
| 21 | 42 |
| 51 | 103 |
| 111 | 223 |

指数观分析。注：IT设备按6年折旧，建筑物按14年折旧。所需收入值不包括运营支出。剩余空间是指超出满足折旧费用所需收入的部分。

# 收入覆盖了持续性的开支，但尚未覆盖累积的账单。

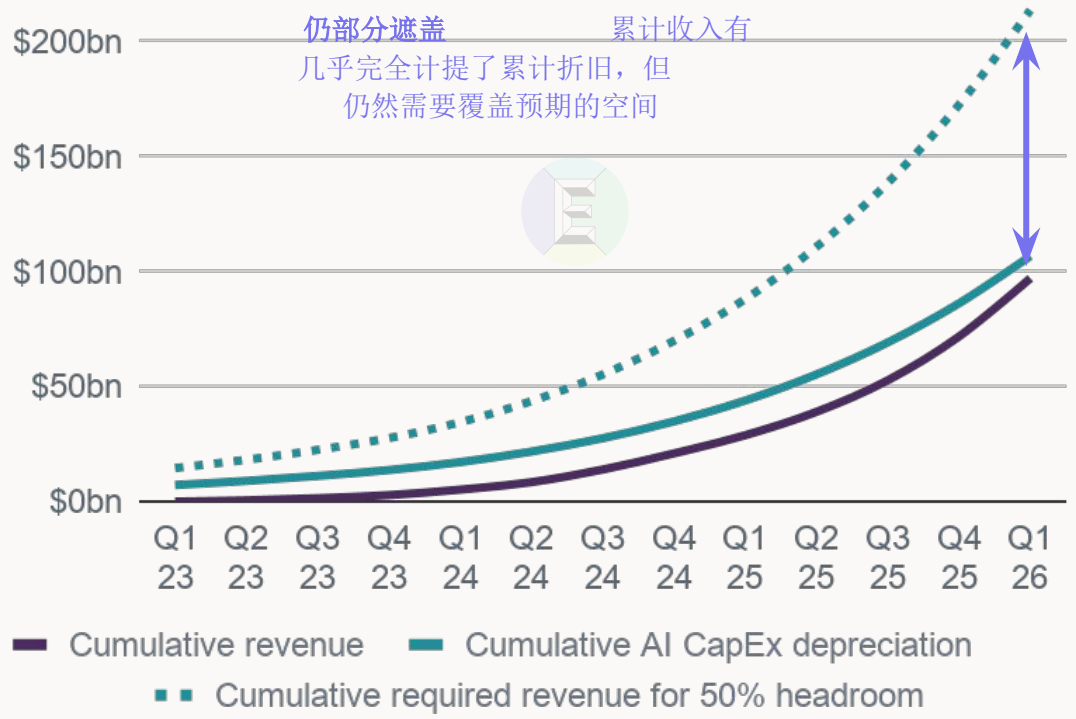
## 季度人工智能收入与资本支出折旧

数十亿美元/季度，仅限超大规模企业及新云服务商




## 累计人工智能收入与资本性支出折旧

百亿美元级，仅限于超大规模企业及新云平台



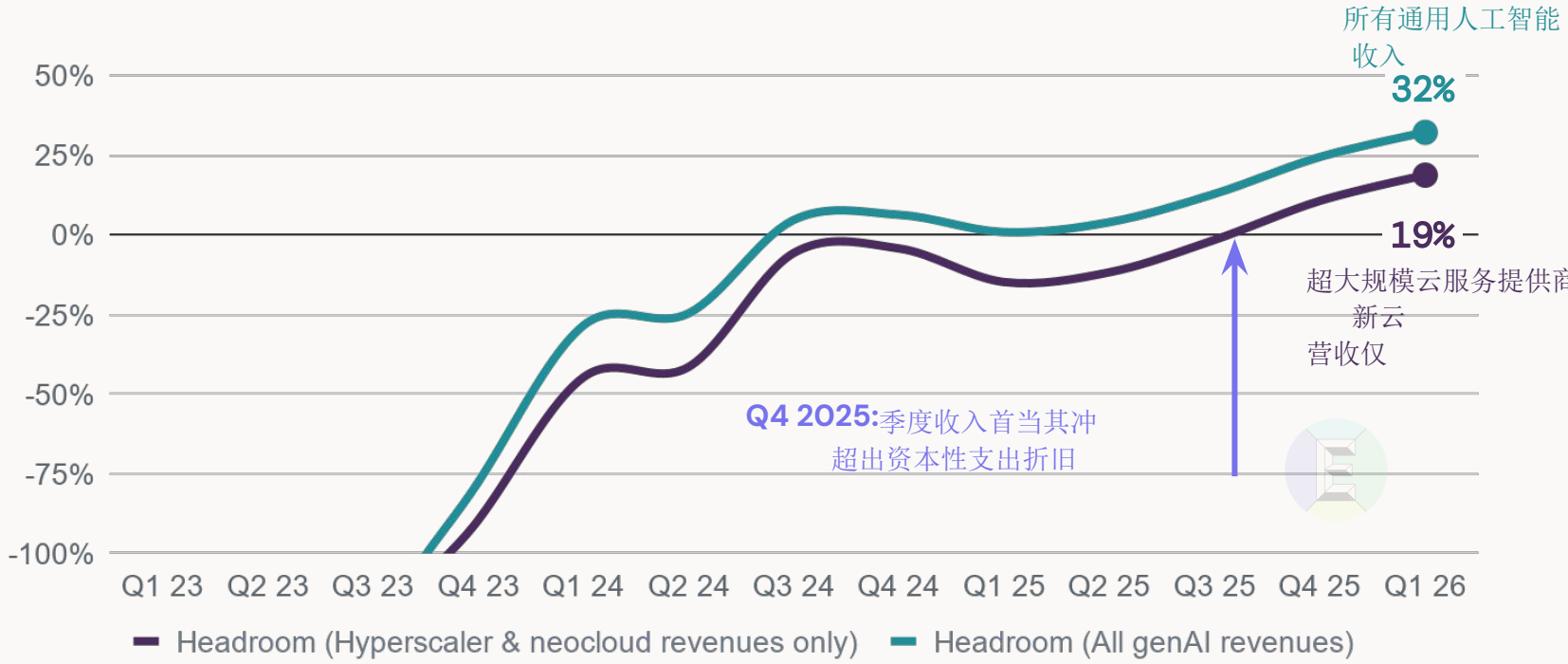

来源：指数观点分析；公司文件。

注意：Meta 对行业资本支出（CapEx）有贡献，但其举措专注于广告提升，因此不被视为纯粹的生成式 AI 收入，或者目前直接货币化程度极低（例如 Meta AI 助手、Muse Spark）。

# 目前，AI基础设施收入刚刚超过今天的折旧门槛。

## 季度资本性支出折旧后的空间

% = (收入 - 折旧) ÷ 收入

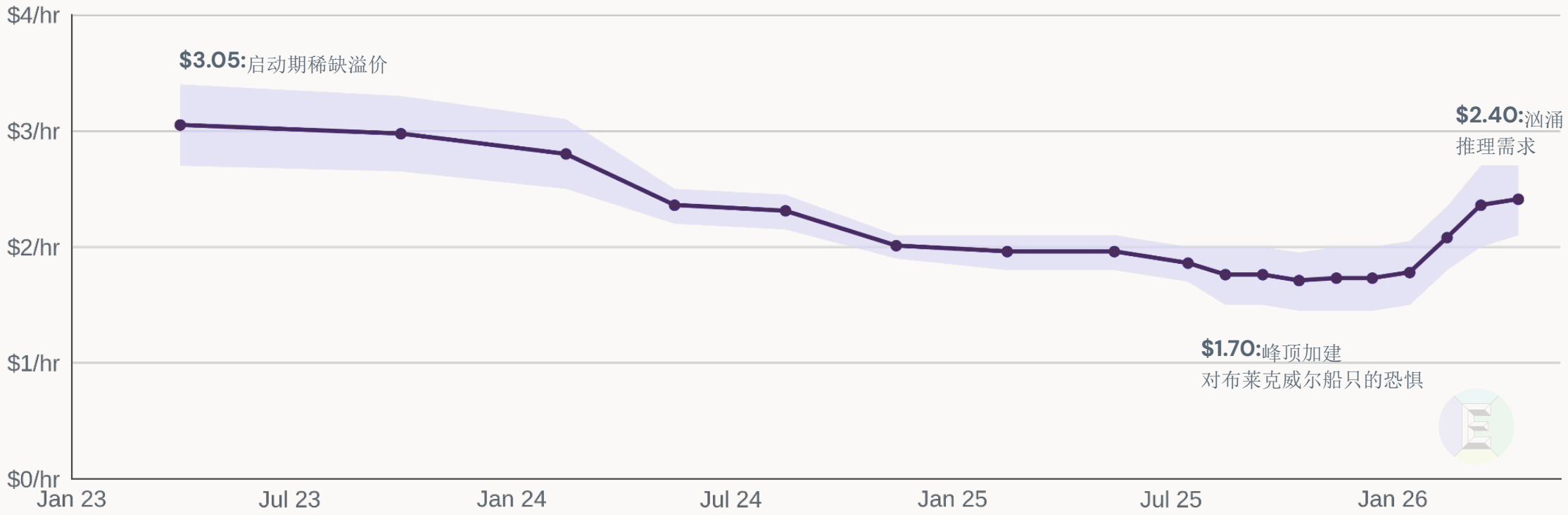



- **目前，生成式AI的收入已足以覆盖其人工智能基础设施的季度折旧。**
  Q1 26，超大规模云服务商/新云收入的空间利用率达到19%，而所有生成式人工智能收入的空间利用率则为32%。
- **覆盖范围仍然有限。**
  折旧费用大致吸收了超大规模/新云GenAI收入的8%，以及（在考虑额外成本前）GenAI总收入的68%。
- **下一次测试是增量覆盖率测试。**
  随着承诺的AI资本支出投入运营，折旧基数将上升。收入增长、利用率和定价必须持续复合，否则空间将被再次压缩。

# 租金水平表明，现有供应正在被需求所吸收。

## H100 一年期租赁合同价格

每小时/GPU美元。H100是场外交易市场中最活跃、交易量最大的GPU：这是一个有用的信号。




来源：指数观分析；半分析

# 数据中心经济设定了代币定价的门槛。

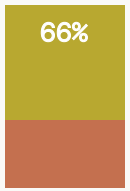
## $7.9bn

每年拥有和运营1吉瓦人工智能能力的成本

**资本成本** $7.0bn/year · 89%


- **服务器** $4.6bn
  48万块GPU，分布于6700套系统中 • 6年寿命
- **设施** $1.4bn
  外壳，电源与冷却 • 14年寿命
- **直流网络** $1.0bn
  交换与互联网络 • 6年寿命
- **土地+公用事业** $33m
  按资本成本计提的土地使用成本 • 效用14年寿命

**OpEx** $900m/year · 11%




- **能量** $594m
  用于运营车队电力 • 66%的运营支出
- **其他运营支出** $308m
  员工、维护及管理费用

● Kimi K2.5 (1T) ●封闭授权下的Kimi级模型（示意）

**将每年7.9亿美元除以代币产出：**

每吉瓦/年 ТОКЕНОВ

75-85万亿

**推理服务提供商成本**

| | 每1M个token |
|---|---|
| $0.10 | $0.42 |
| | 0.32美元的许可费（约占1.29美元的加权平均售价的25%） |

**50-75% 毛利率所需价格**
**MARGIN:**

| | 每1M个token |
|---|---|
| $0.20 – $0.40 | $0.84 – $1.68 |

**25%投资回报率所需客户价值**

| | 每1M个token |
|---|---|
| $0.25 – $0.50 | $1.05– $2.10 |

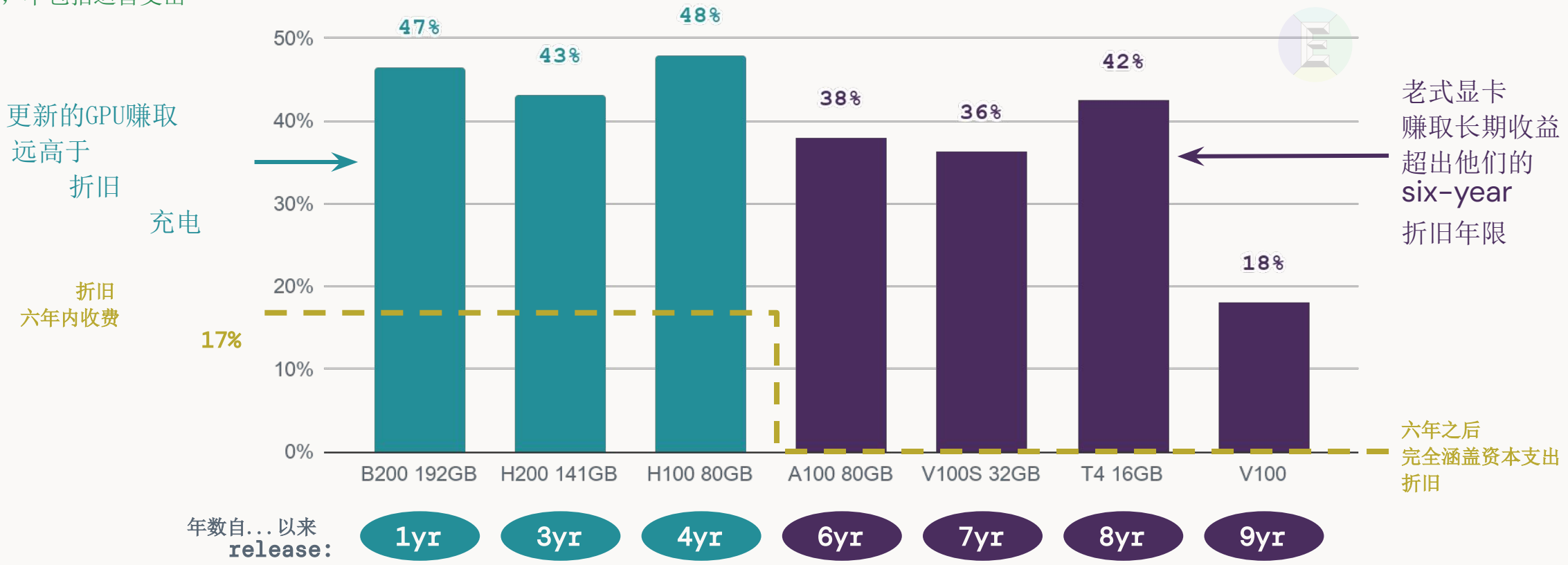
来源：Exponential View分析；Epoch AI；SemiAnalysis

注意：示例模型。1吉瓦IT算力（6.7千GB200 NVL72系统，48万GPU），每周期AI拥有成本（2026年5月），年化包含资本成本，6年IT使用寿命。来自SemiAnalysis InferenceX的Token输出：FP4，8k输入/1k输出，每用户50 tokens/秒，65%利用率
（2026年4月）。令牌输出范围反映了相对于推测解码的+10-25%的吞吐量提升。开放权重模型无需模型许可费。封闭权重列增加25%的许可费，基于Kimi的1.29美元综合价格。

# 租金回报率表明其使用寿命超过六年。

GPU良率在50%利用率下

%，不包括运营支出




来源：指数观分析；硅数据。注：良率 = （按需率 x 50% 利用率 x 8760 小时）÷ 原始清单价格。

# 更长的GPU使用寿命为未来发展提供了空间

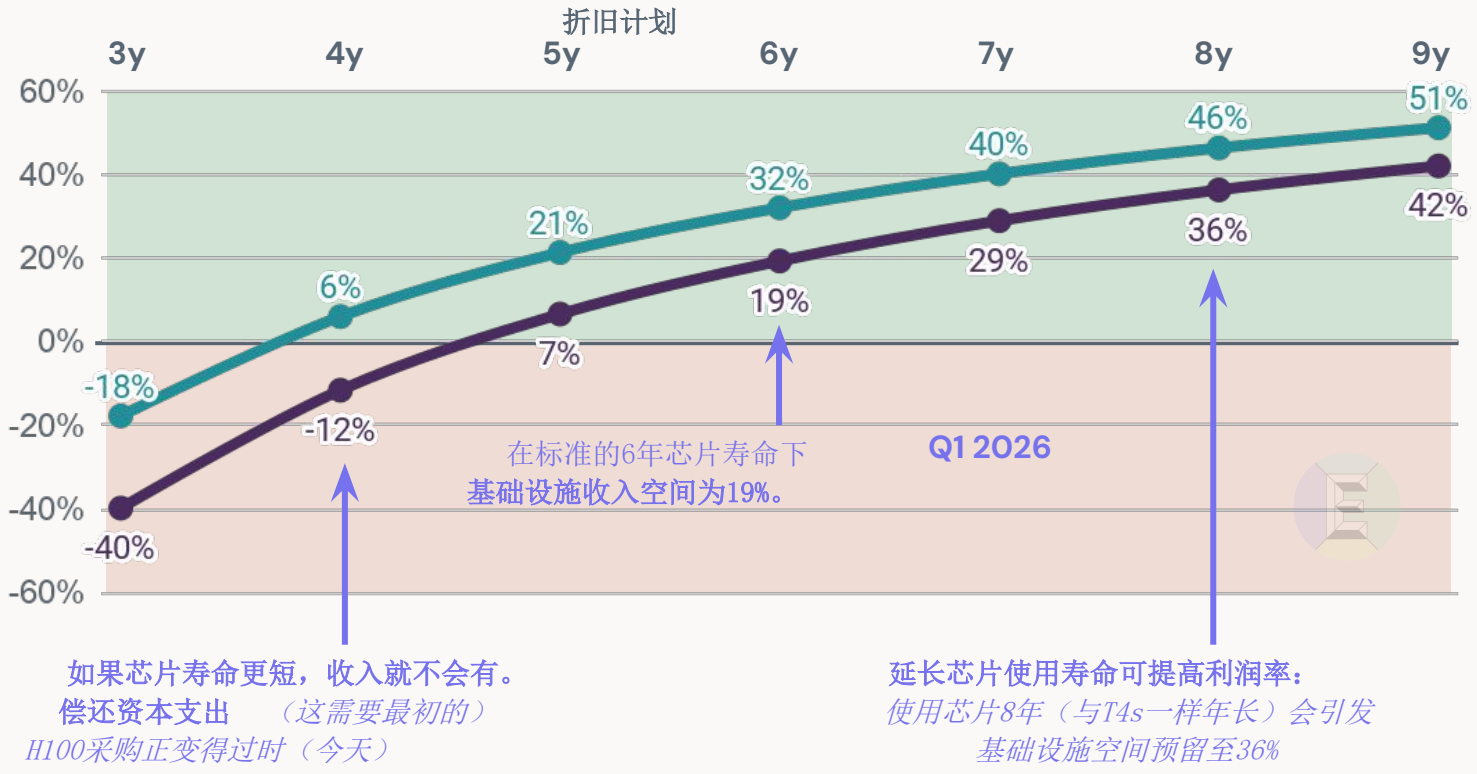
**马克·扎克伯格**
Meta Q3 2025 财报电话会议

*……最糟糕的情况就是，我们实际上已经提前几年就建好了，在这种情况下，当然会有一些损失和折旧，但我们最终会适应并利用它。*

**过度建设可能是一种押注于更长的芯片折旧。**

**单颗芯片资本支出折旧计划后的可用空间范围**

Q1 2026，3-9年计划，% = （收入 - 折旧）÷收入




来源：指数观点分析；Meta
请注意：建筑物在14年内按固定折旧率计算。

# 4 | 令牌AI经济的价值单位是什么？

代币的年增长率高达14倍，这主要是由代理型工作负载和高弹性需求所推动。基于代币的定价模式使这一情况尤为突出，但它也为行业提供了将代币消耗产生的产出进行归因和评估的机会。

> *输入是电子，输出是代币。中间是英伟达。*
>
> 黄仁勋

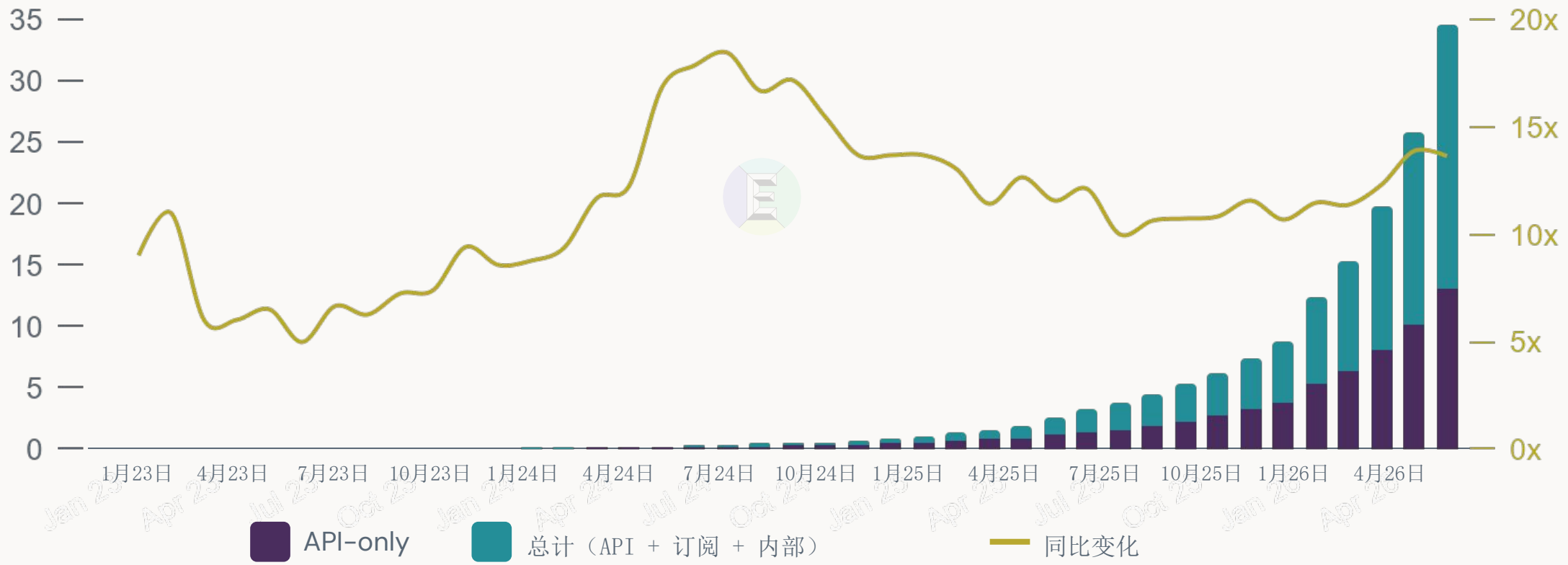
> *“令牌，我们模型处理的数据的基本单元……”*
>
> Sundar Pichai

# 生成式人工智能经济是一种代币经济吗？有点像。

# 全球代币总量每月超过30亿枚，同比增长14倍。

**推理令牌处理**

每月万亿代币（左轴），增长率倍数（右轴）

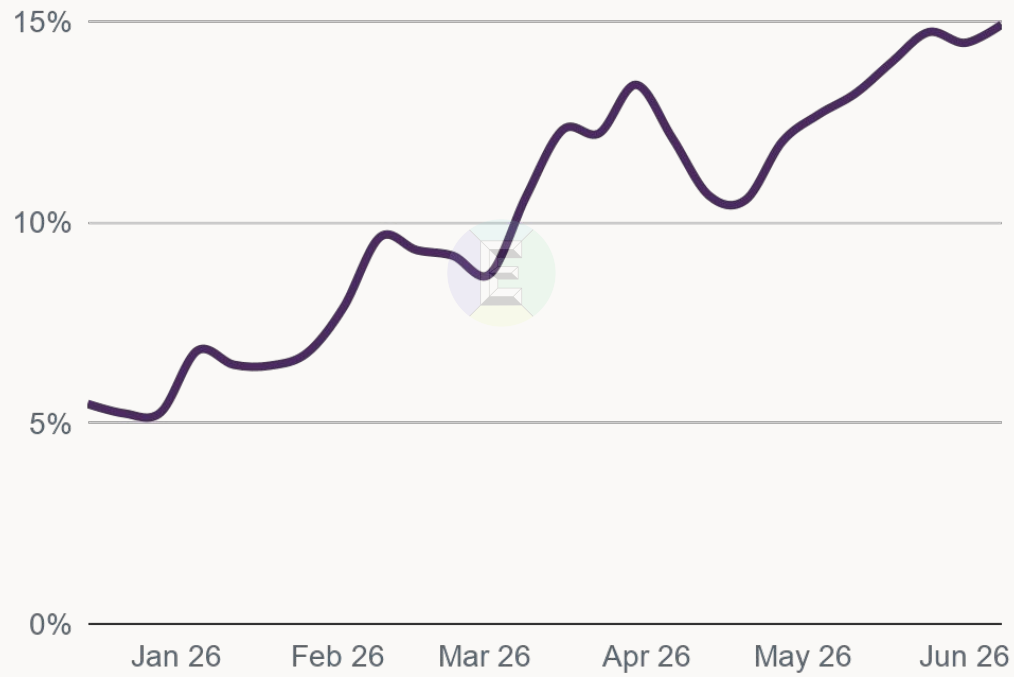



来源：指数观点分析
注意：全球公司 中国

# 从聊天转向代理正在增加代币的使用量

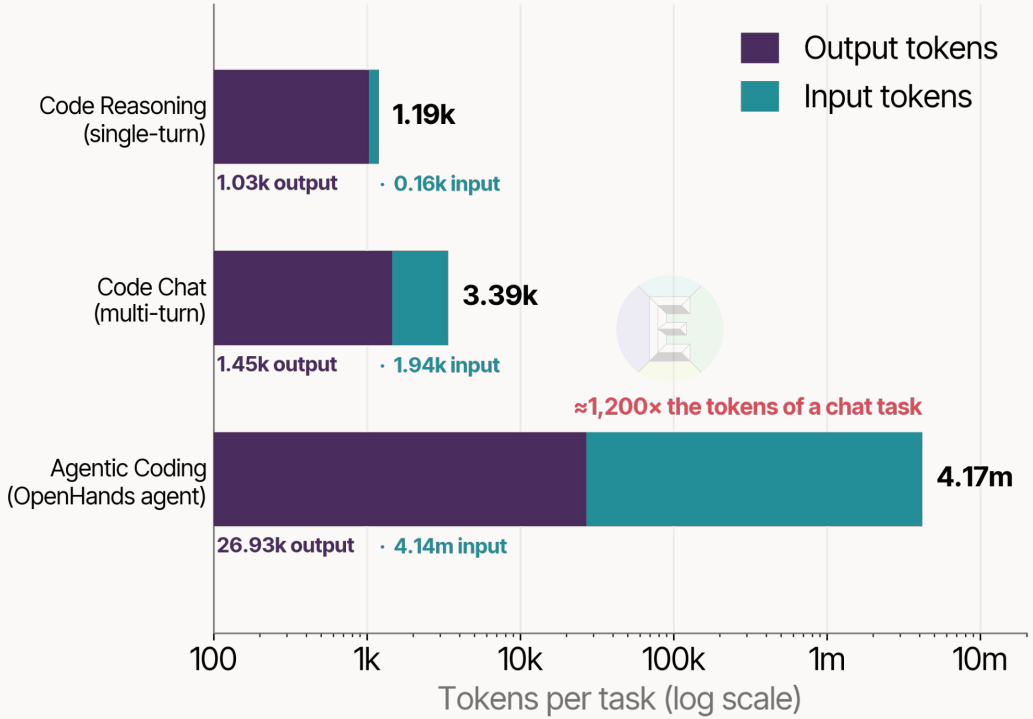
## 代理协调密度

每个提示的%工具使用率，OpenRouter



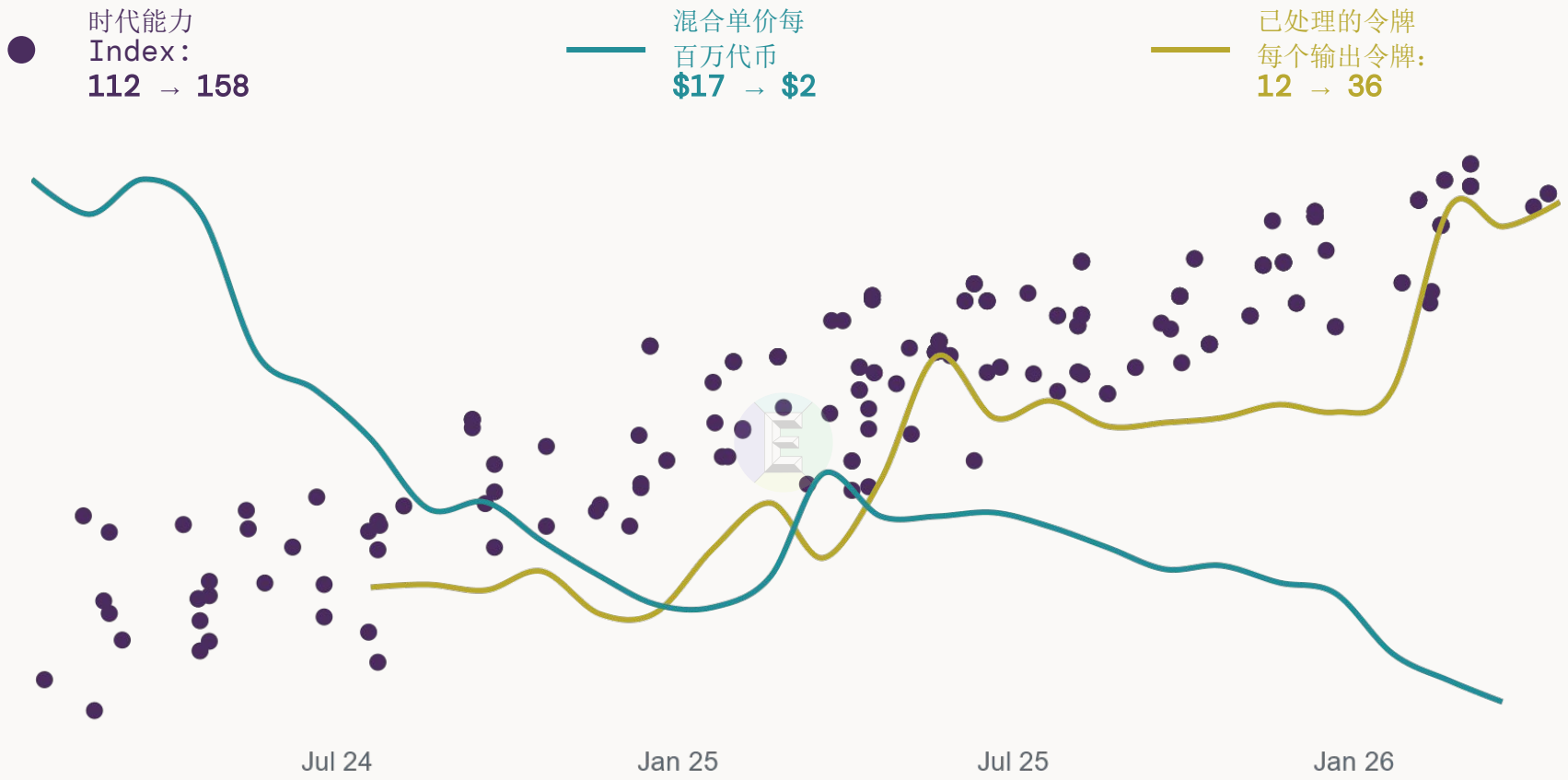

## 每项任务的代币消耗量

每项任务平均消耗的令牌数，对数刻度




来源：指数观点分析；OpenRouter；白、黄、王、孙、米哈尔察、布赖恩约夫森和彭特兰 2026。

# 更便宜的代币和更好的模型会刺激需求。




- 更强大的模型能够覆盖 更广泛的具有经济利用价值的任务

提升价值与使用。

- 代币数量上升中

推理模型花费更多“思考”的代币

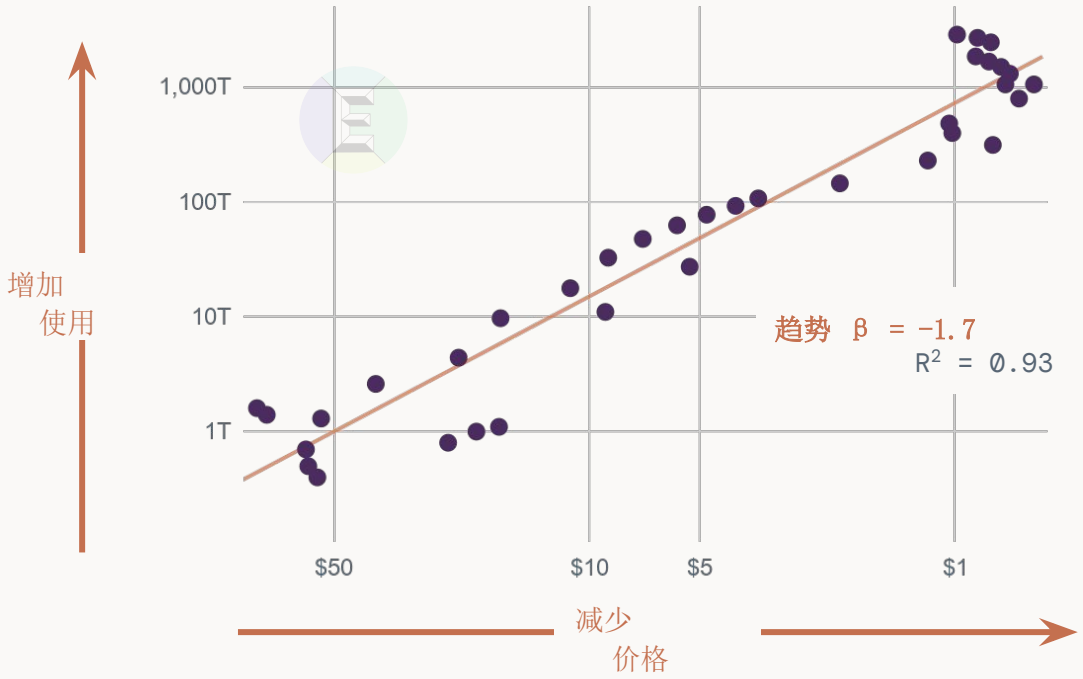
- 价格下降鼓励更多使用，并使先前不经济的应用变得可行

来源：指数观点分析；纪元人工智能。

# 需求对价格弹性明显：随着价格下跌，使用量增长更快

## 谷歌价格弹性（平均价格与销量）

百万级 token 每月 vs 万亿级 token 每月，双对数坐标，2023-2026




**Sundar Pichai** Google: I/O 2025

*……我们当时每月处理97万亿个token。现在超过480万亿——是原来的50倍。*

**Price -97% | Volume 50x**

**山姆·奥特曼** OpenAI：“三点观察”，2025

*使用特定级别AI的成本大约每12个月下降10倍，而更低的价位则带来了更多的使用。*

**价格 -90% | 数量 ↑**

**单大卫** Volcengine / ByteDance, 2025

*豆包本月日活量使用量突破500万亿，较2024年12月的40万亿有所增长。*

**Price -50% | Volume 12x**

**elasticity ≈ 1.2-1.8**

跨服务提供商，规模

每降价10% → 代币数量增加12-18% → 代币总支出仍然上升

来源：指数观分析；谷歌；OpenAI；字节跳动。注：时间序列，非横截面：价格和使用量均随时间趋势变化，因此 β 可能高估了纯价格弹性。

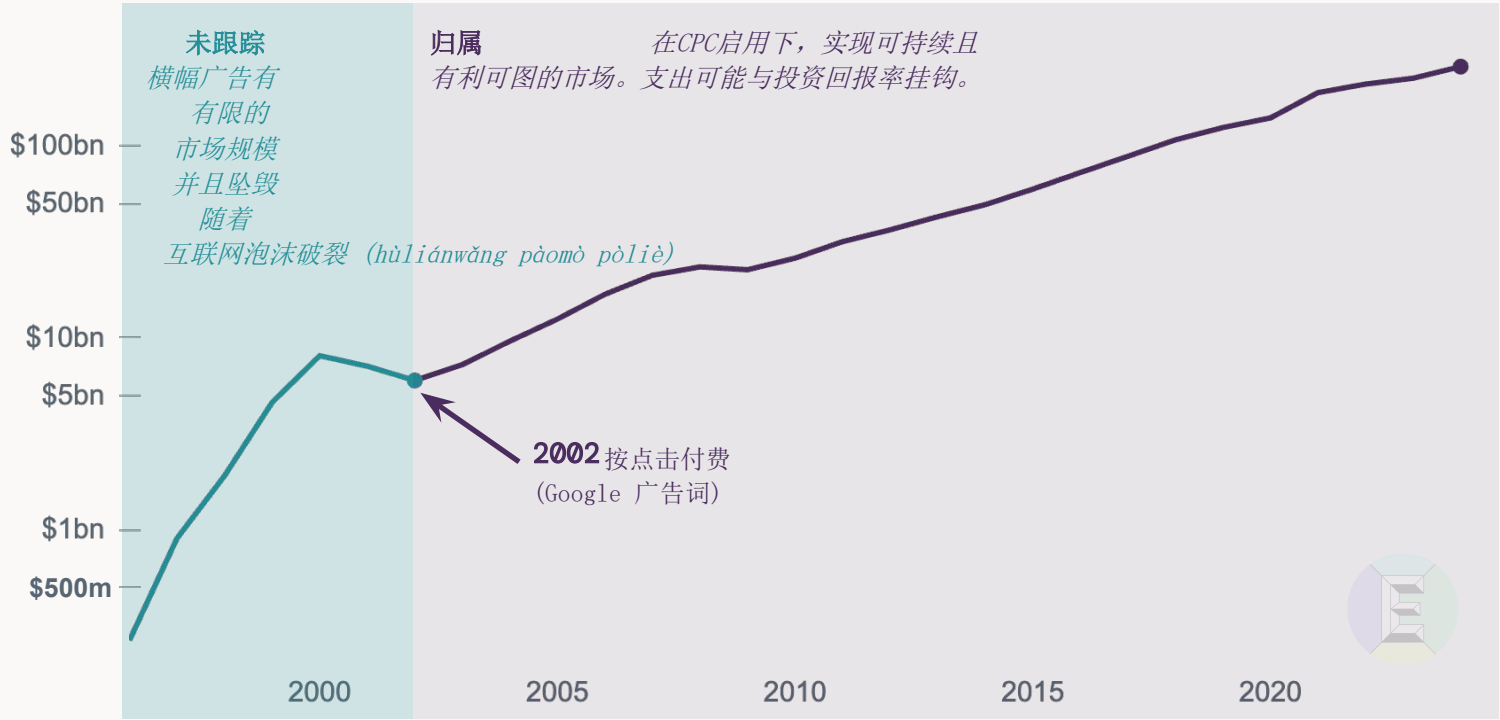
# 基于token的定价是人工智能的“按点击付费”时刻

## 人工智能定价模型的演变

1. **FREE**
   预算中未予认可
2. **订阅**
   按座位定价，不含
   追溯到特定值
3. **基于令牌的定价**
   格律用法使（并且）能够...
   需要注明项目来源

## 年度数字广告收入

$bn/year, log scale, 1996-2024

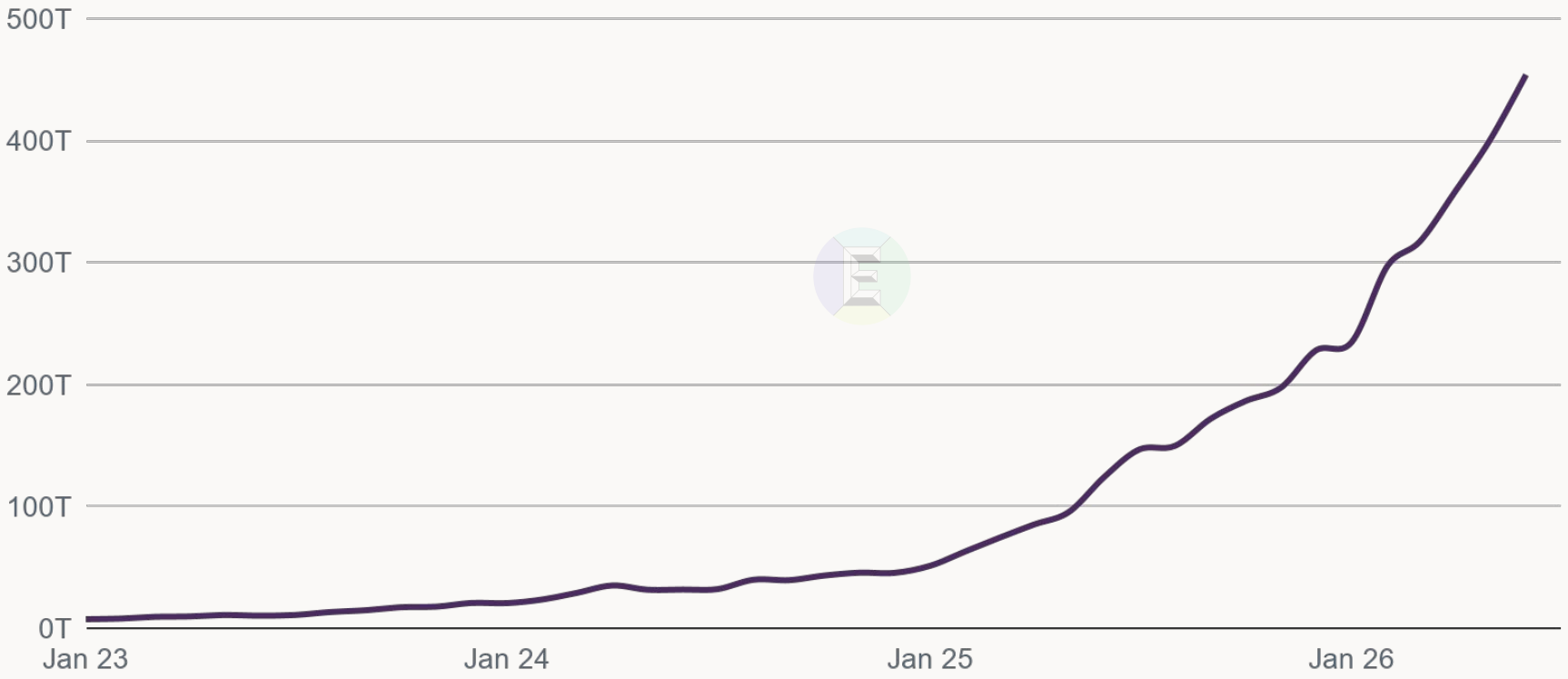



来源：Exponential View分析；IAB/PwC互联网广告收入报告

# 每一代技术都提升了每吉瓦容量的象征性产出。

**每个数据中心吉瓦容量的产出代币数**

每月每吉瓦万亿个代币



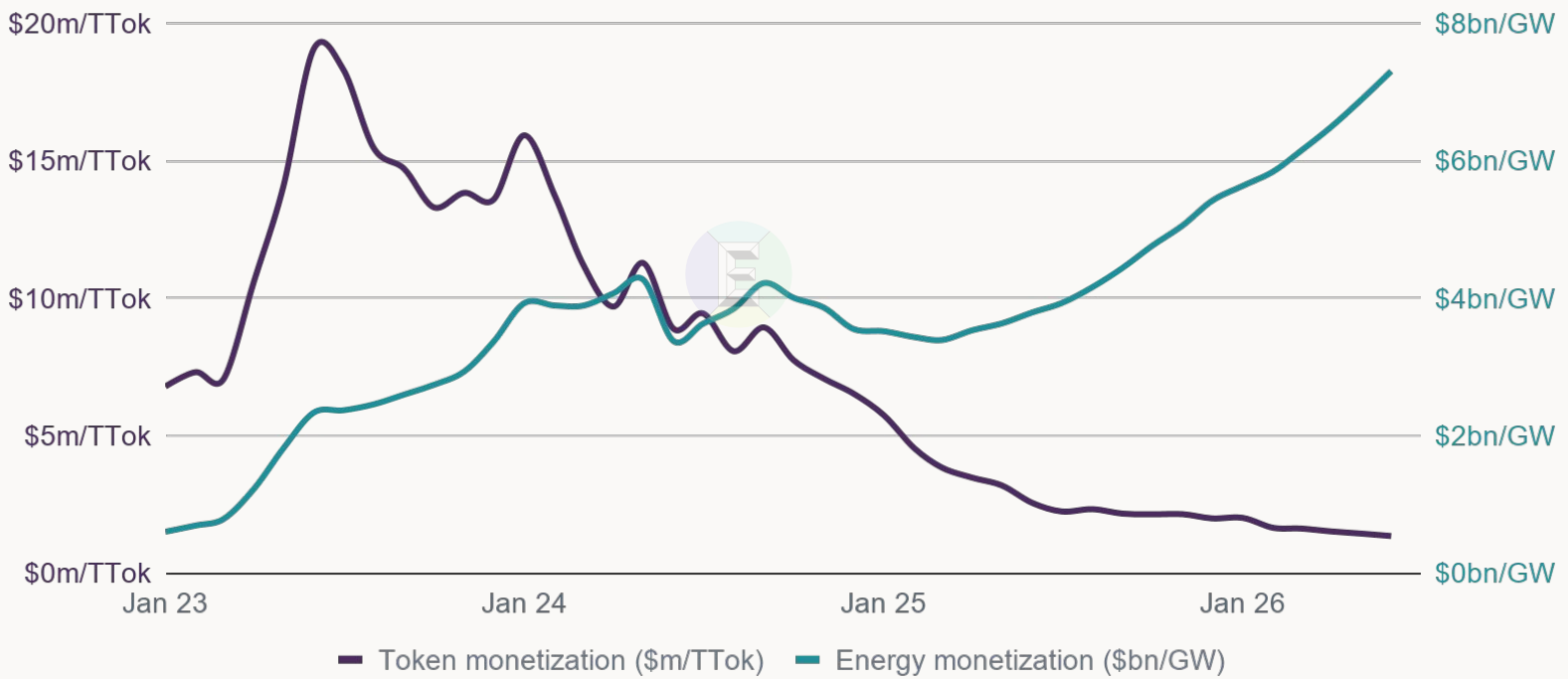

- 每吉瓦每月可购买更多代币产出。
- 即使物理限制给人工智能经济带来压力，这也使得超级周期成为可能。
- 动力来源：a）实验室通过使用更小的模型和提供更好的服务来实现更高的效率。b）硬件方面的进步，尽管进展较慢（例如：Hopper → Blackwell → Rubin）。c）工作负载组合从训练转向推理。

指数观分析。全球除中国。

# 这种效率正在提高每吉瓦容量的货币化程度，同时每枚代币的收入却在下降。

## 代币与数据中心产能产生的收入

m/TTok（左轴），bn/GW（右轴）




- 每万亿代币的收入自2023年峰值以来已下降，反映了价格的下跌。
- 效率提升推动代币价格下降，但需求增加使其得到大幅弥补。
- 全行业数据中心每吉瓦容量的收入已超过70亿美元/吉瓦。

指数观分析。全球除中国外。年收入。

# 令牌是人工智能的计费指标，但尚未成为价值单位。

电

灯

Kilowatt-hours

爱迪生的第一批客户按灯付费。
灯泡已安装。计量装置是后来才安装的。

互联网

页面浏览量

会话与点击

广告商支付注意力。
conversion:
CPM → CPC → CPA

移动互联网

兆字节

活跃用户

按用户参与度评估的应用
日活跃用户数 / 月活跃用户数，留存率

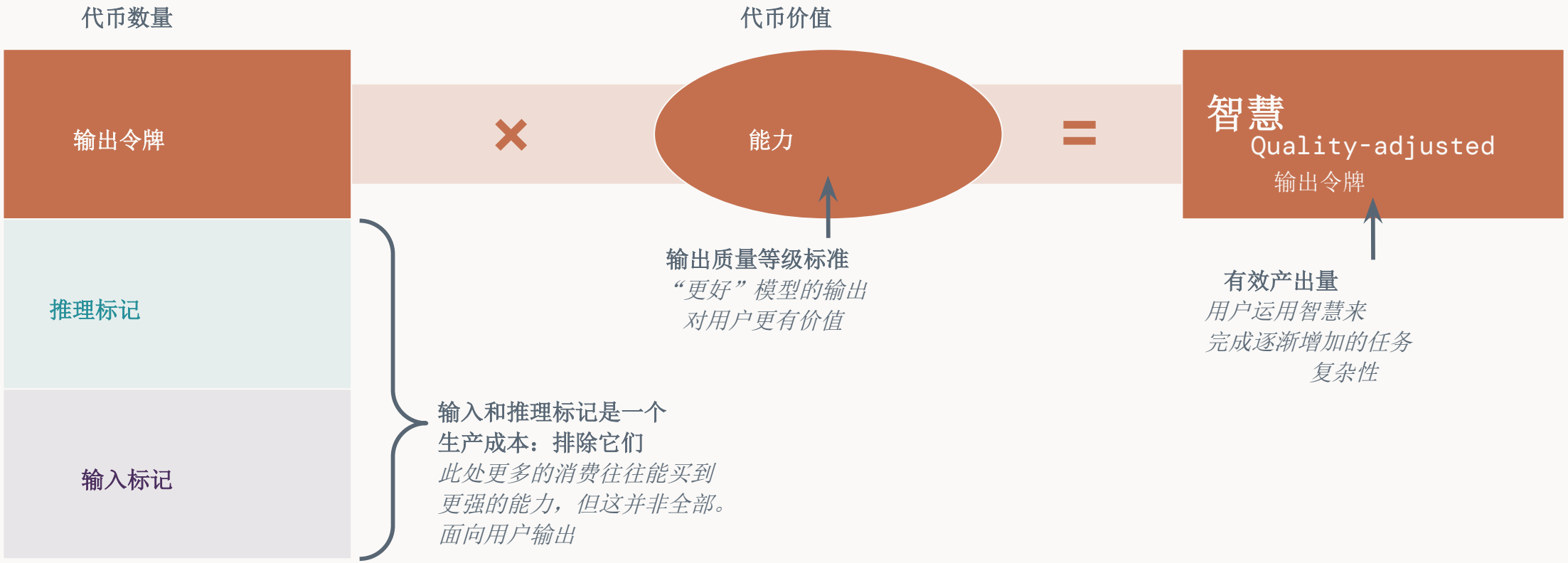
生成式人工智能

标记

“Intelligence”?

价值创造单元尚未定义

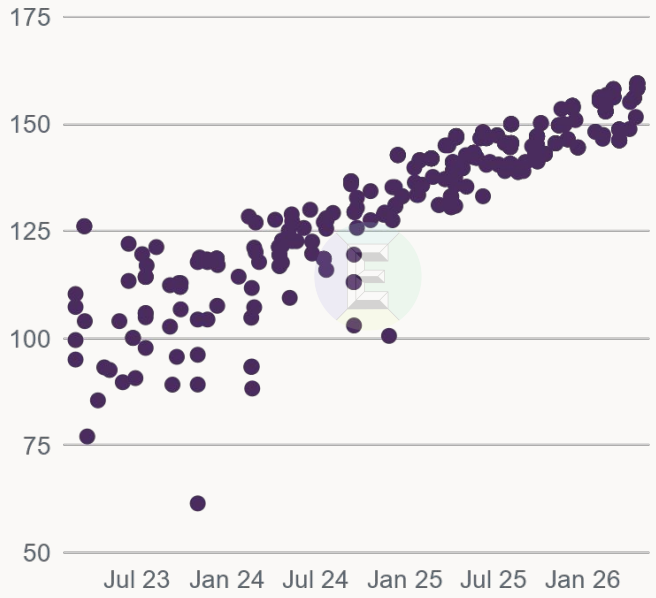
# 质量调整后的代币最接近一个可用的价值单位



# 无论你选择哪种衡量标准，趋势都在上升。

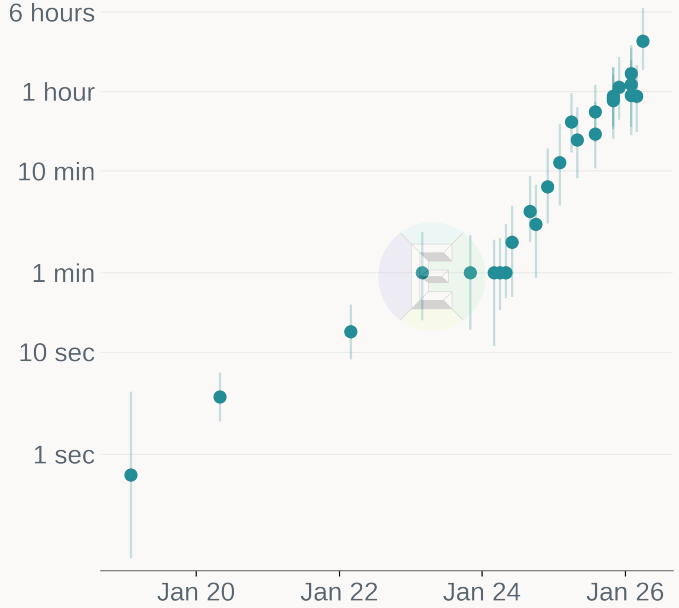
### 时代能力指数

Score: GPT-5 = 150,
Claude 3.5 Sonnet = 130



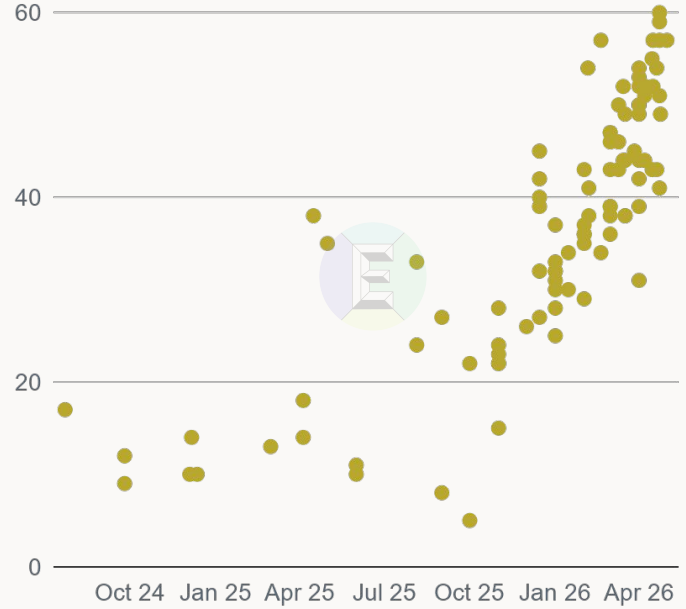

### METR任务视界

人类任务在50%模型成功率下的持续时间，对数刻度




### 人工智能分析指数

Score: 0-100

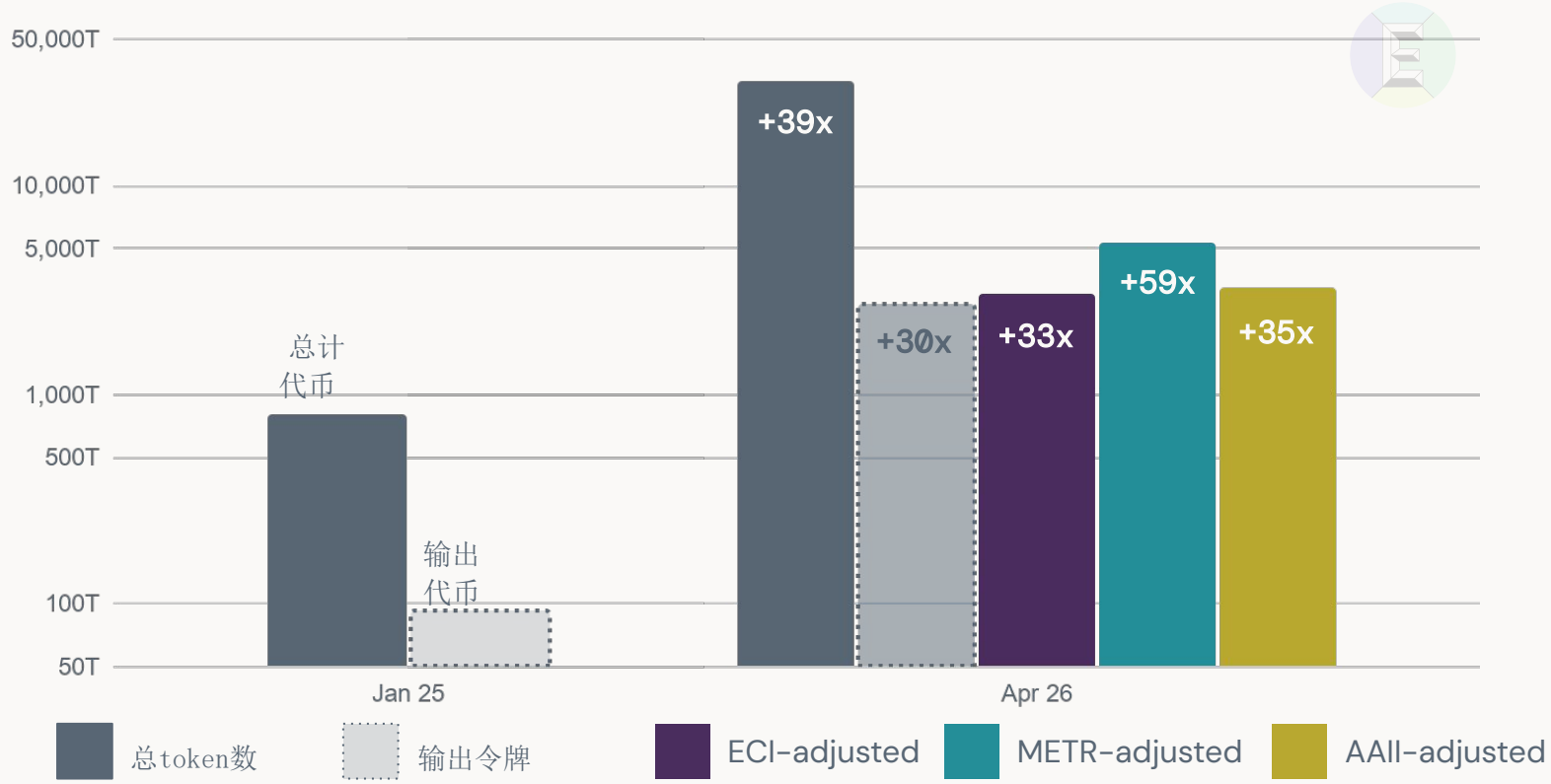



来源：指数观分析；纪元人工智能；METR；人工智能分析
注意：METR 任务主要包含软件工程、机器学习和网络安全任务。

# 质量调整后的产出与原始体量增长保持同步

**总数、输出及质量调整后的指标**

每月万亿个代币，2025年1月与2026年4月



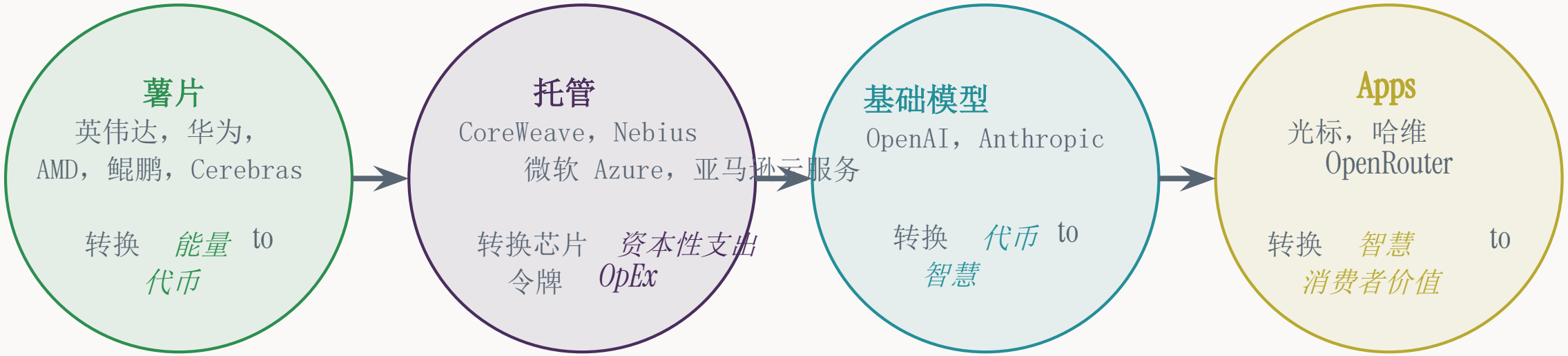

- **原始输出 token 增长速度慢于总 token（30 倍 vs 39 倍）：**投入和推理上花费的金额不断增加。
- **在分数提高方面广泛（33倍至59倍）：**不同的衡量标准和不同的评分量表意味着我们只能得出方向性的结论。
- **调整质量后的输出令牌似乎在增长：**输出量和能力均高于2025年1月。

来源：指数视图分析；OpenRouter；纪元人工智能；METR；人工智能分析。注：ECI、METR & AAII已索引至2025年1月平均分，用于比较目的。

# 5 | 堆栈：价值所在之处

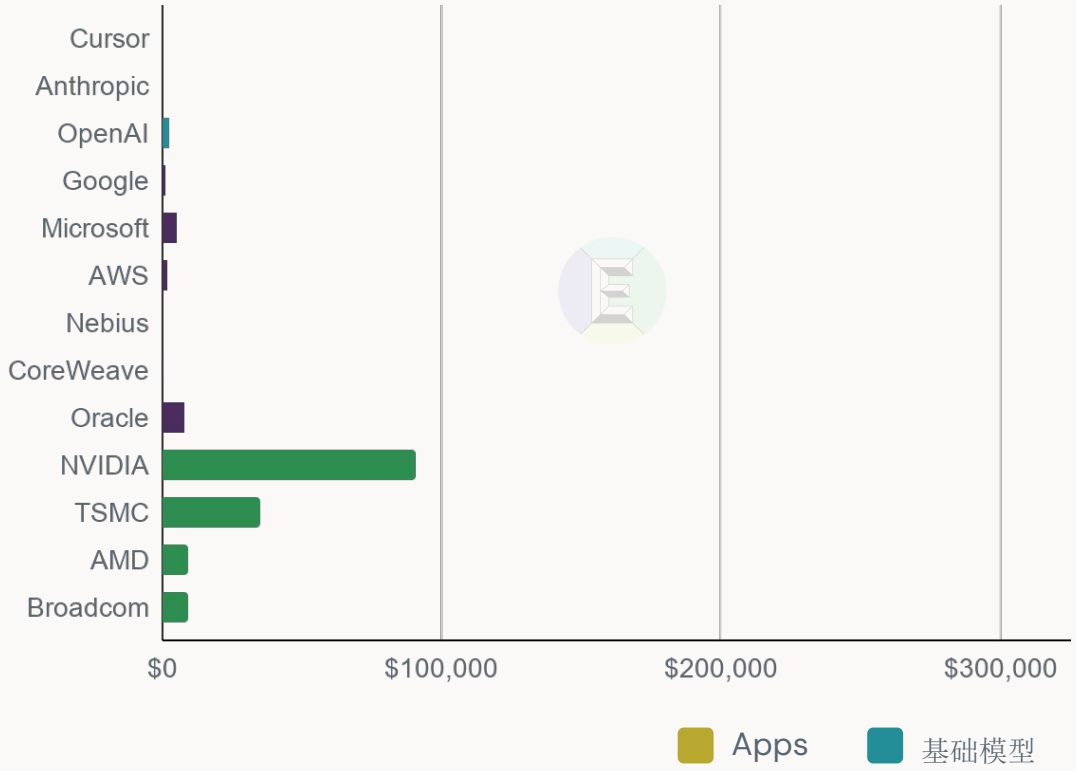
收入集中，但应用和模型正获得份额。实验室仅在掌握前沿技术时保留定价权；昨日的前沿经济价值会迅速衰减为开放权重。利润在技术栈中累积的位置将取决于技术进步，而技术进步会跨越竞争动态。

# 堆栈将资本和能量转化为认知工作。

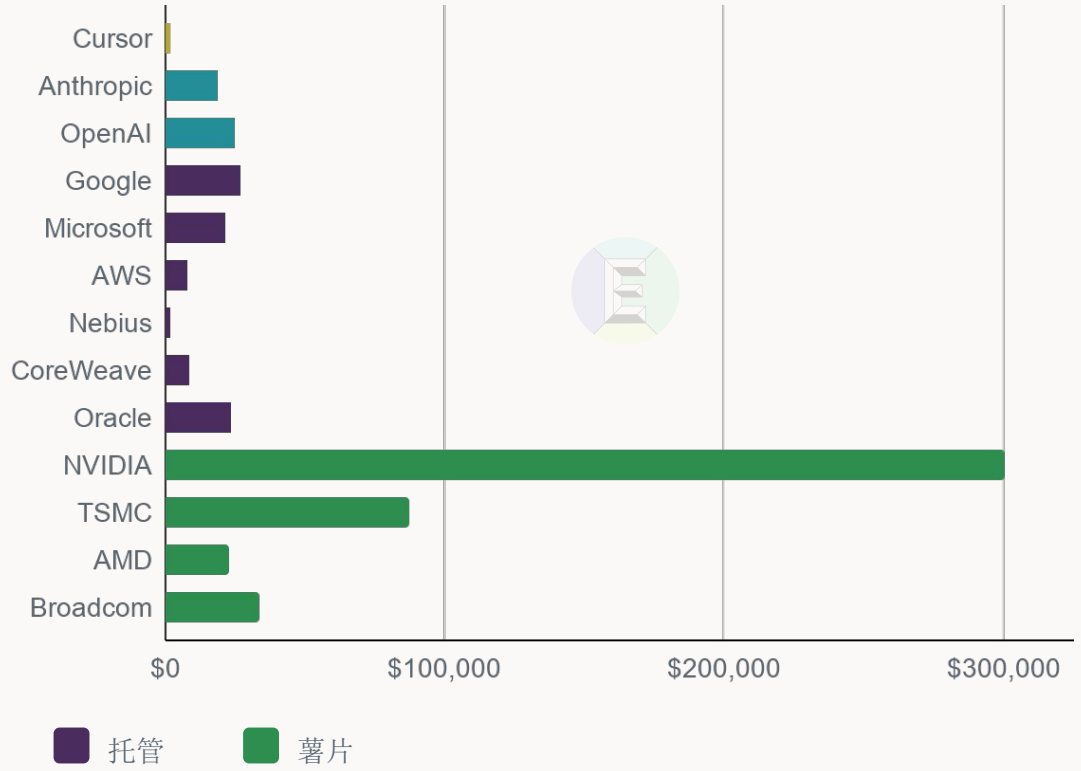


# 目前收入集中，但结构正在变化。

年化生成式人工智能收入
$million, Q1 2024




年化生成式人工智能收入
$million, Q1 2026

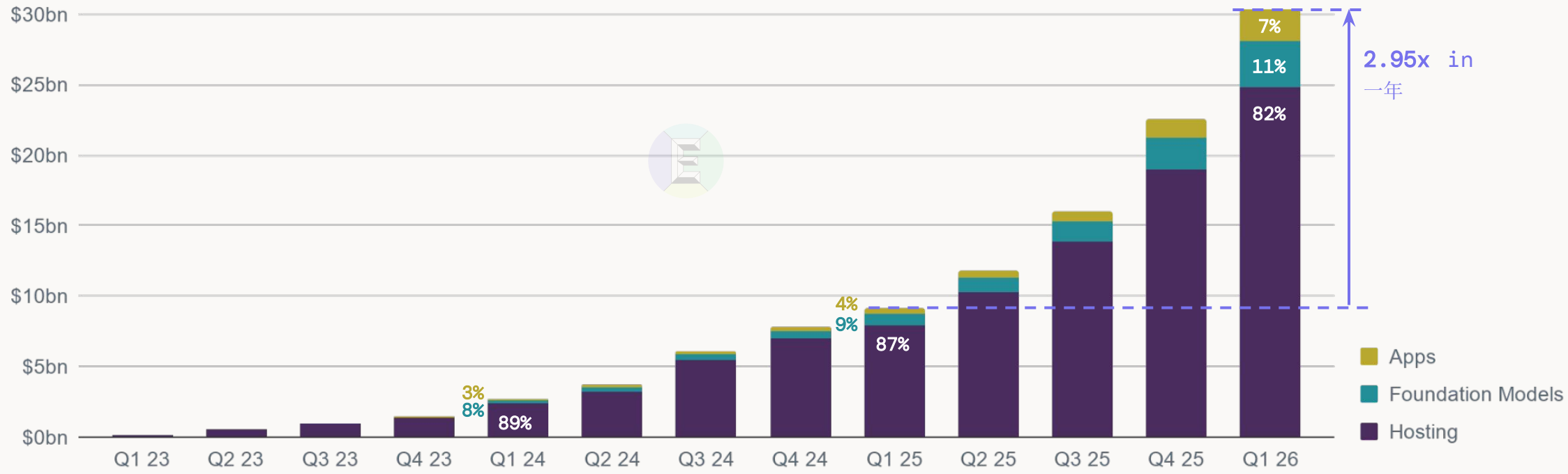



来源：指数观分析；公司提交的文件。注意：收入不适用去重调整。

# 价值正沿着价值链向上移动，朝着应用程序和模型发展。

## 按层级去重后的季度生成式人工智能收入

每季度百亿美元

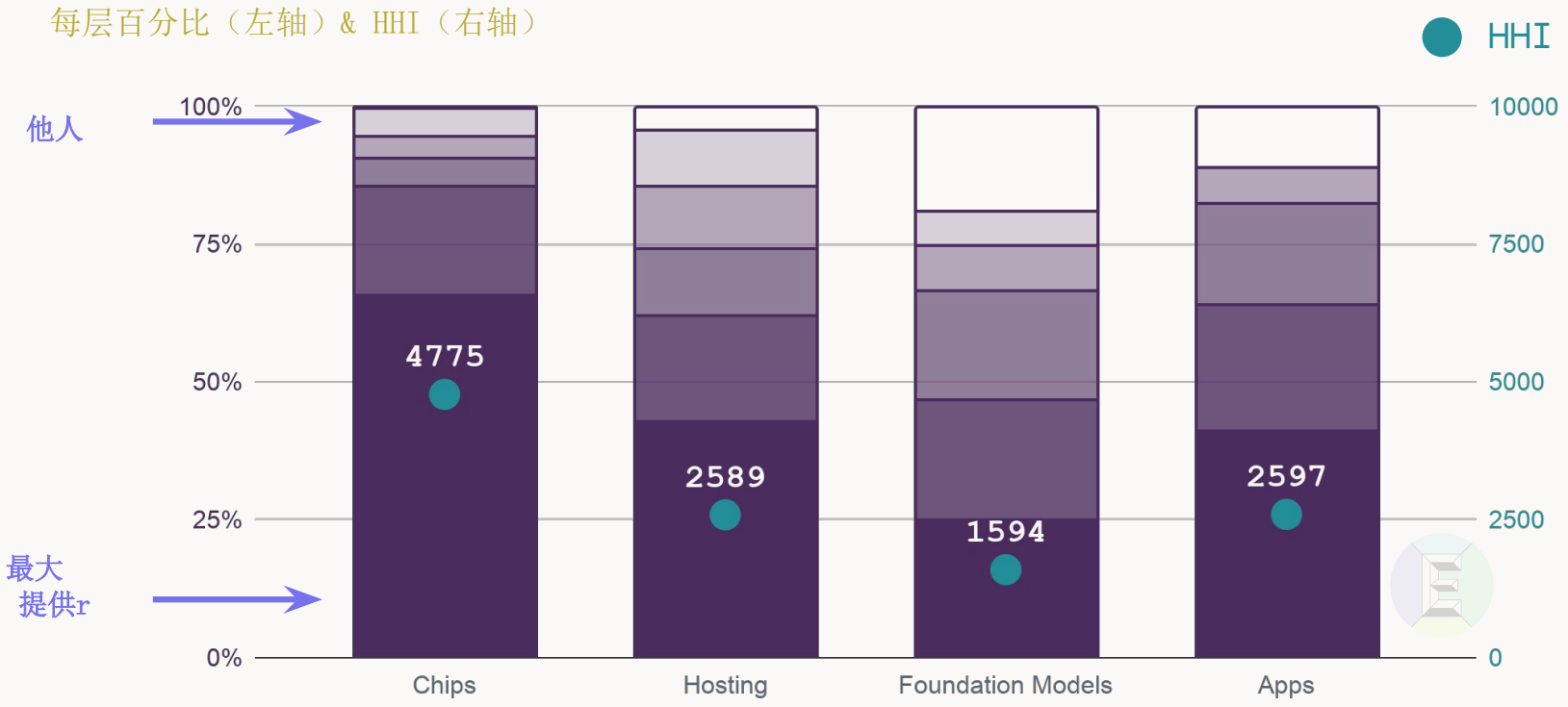



注意：全球范围（不含中国大陆）。不包括芯片（芯片被视为运营支出，由托管层资本化）。

指数观分析。由于四舍五入，数字可能不会加到100%。

# 定价权跟随竞争压力，而非堆叠位置。

## 领先企业的市场份额与赫芬达尔-赫希曼指数（HHI）




- 上游供应商为代币定价捕获所有可用的页边空白缺乏下游竞争。
- 英伟达是最大的芯片供应商，但 **亚马逊和谷歌对定制硅的垂直整合**可能降低其显著性。
- 联邦通信委员会（FM）的收入高度集中于OpenAI和Anthropic，但开放权重的模型提供了

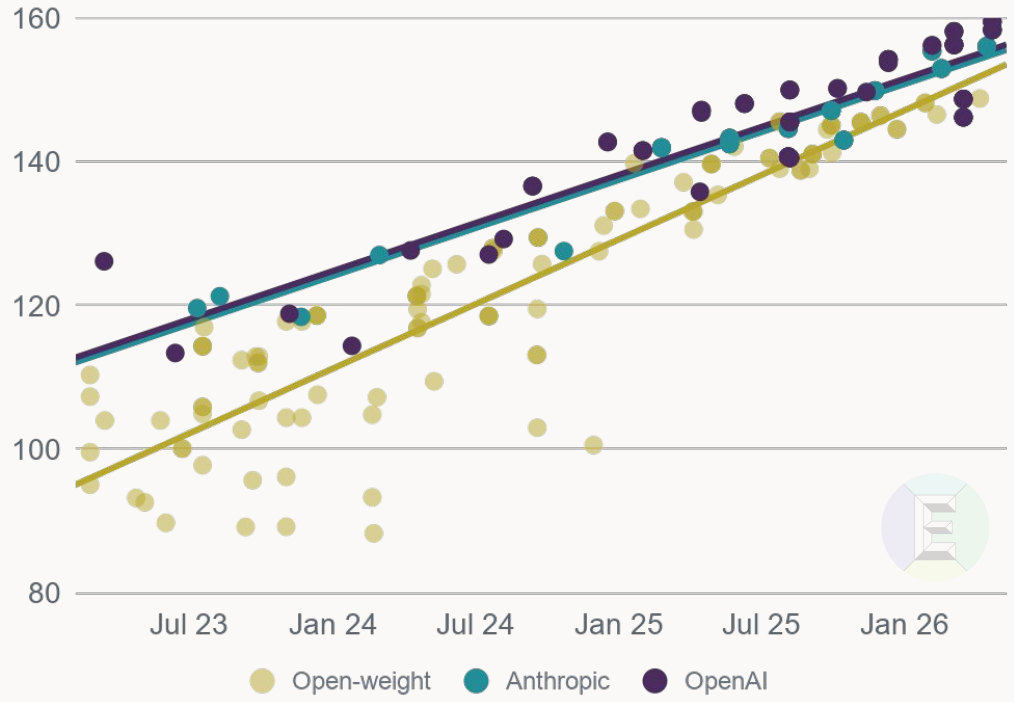
**低成本竞争优质代币。**

指数观分析。注：主机与应用程序按收入份额定义。基础模型按代币份额定义（由于开放权重竞争）。芯片按计算份额定义（H100-等效值）（由于专用超大规模芯片的垂直整合）。赫芬达尔-赫希曼指数是衡量市场集中度的指标。

# 前沿实验室目前尚能支撑高端定价。

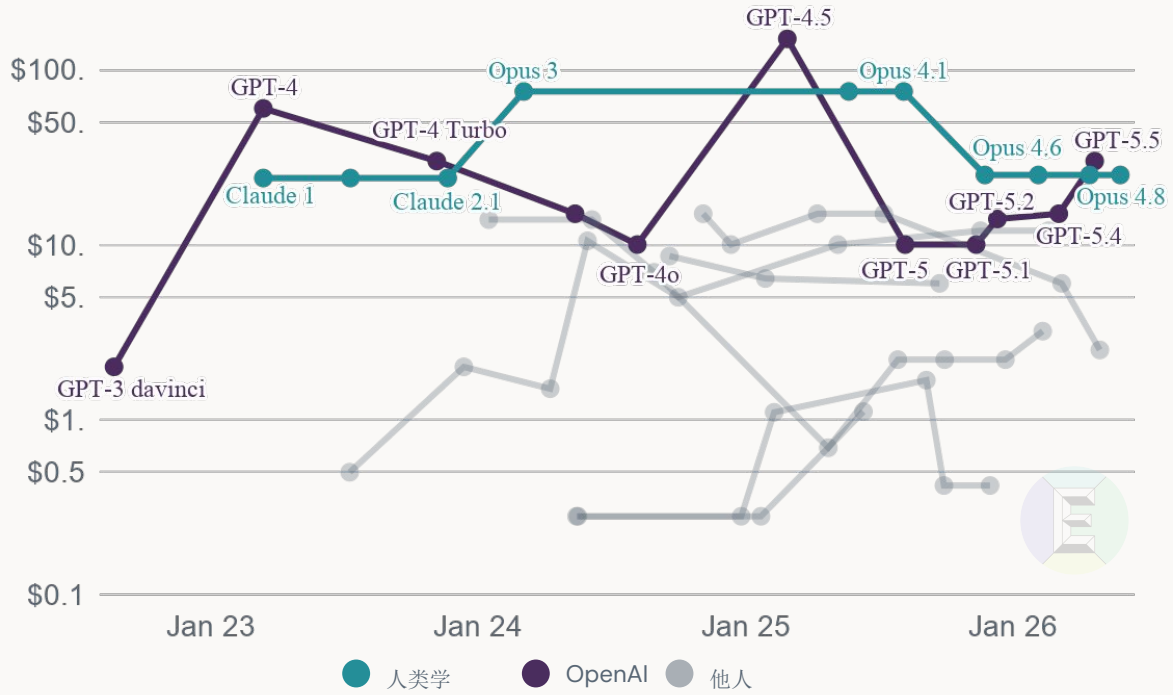
### 时代能力指数

发布时评分，OpenAI/Anthropic/开源权重



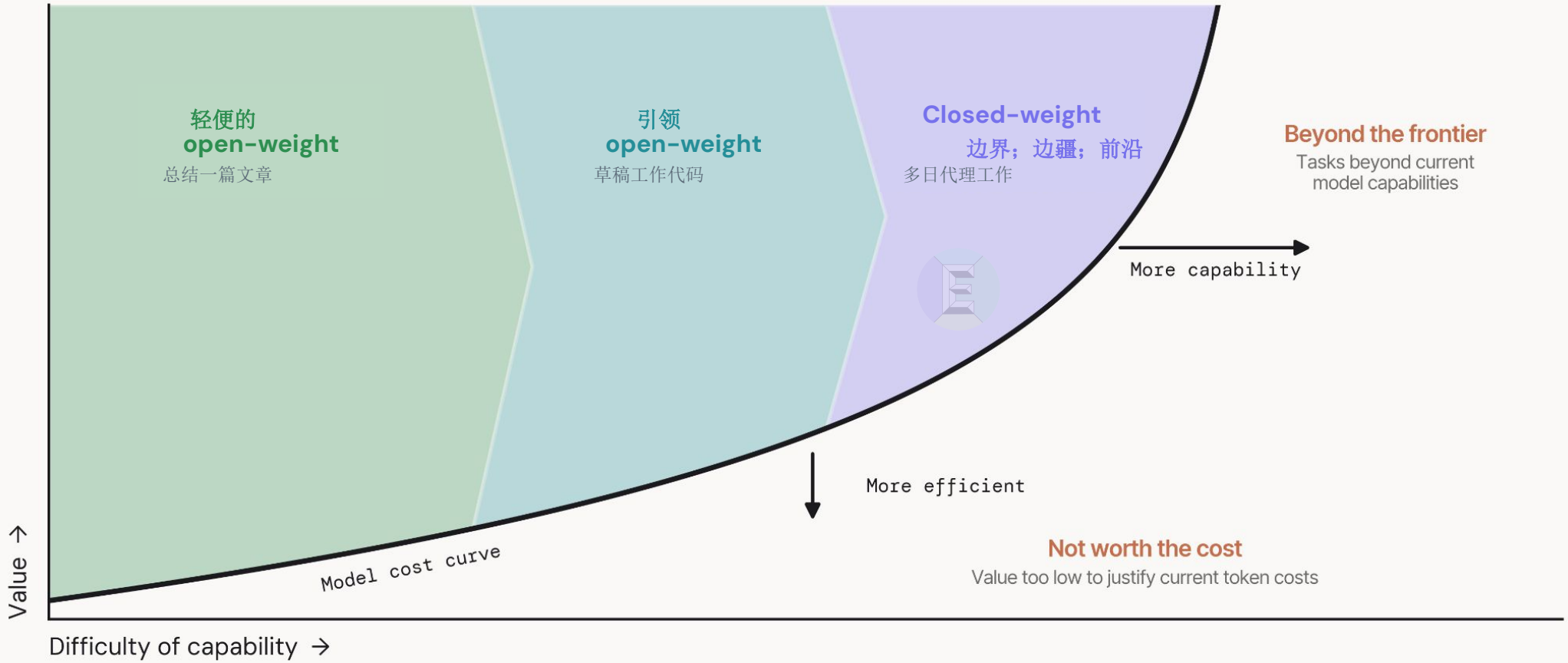

### 销售价格

百万美元的输出令牌




来源：指数观点分析；纪元人工智能。

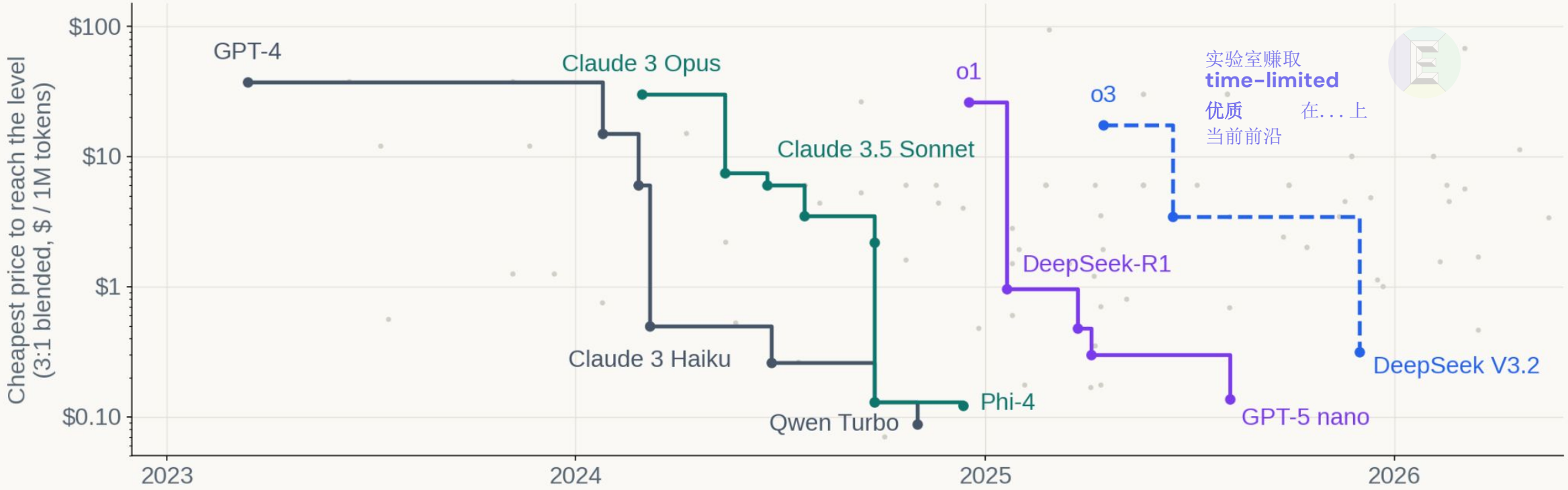
# 实验室必须超越开放式重量的商品化。保留边距



# 去年的前沿正在迅速商品化。

**每项能力前沿的价格**

混合价格美元，按性能在GPQA钻石（博士级科学）分组



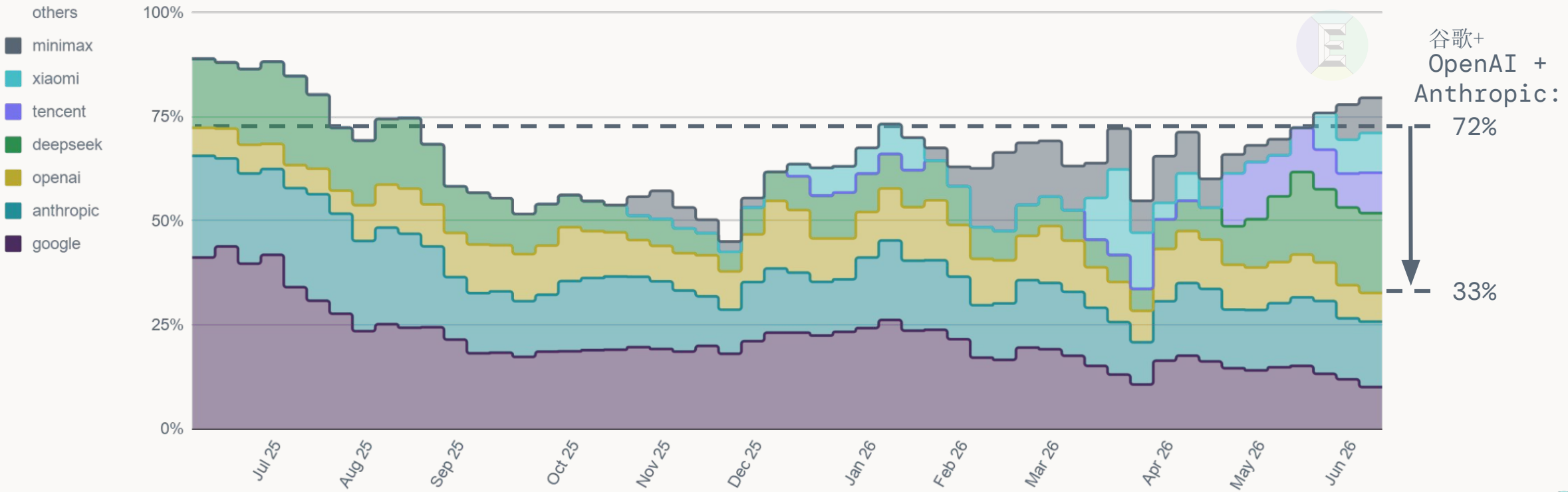

来源：指数观点分析；纪元人工智能。

# 在自行选择的 OpenRouter 用户中，代币份额正在变动。

## 开放式举重

每周 | OpenRouter | 代币份额

每款模型的作者百分比




来源：指数观分析；OpenRouter。
请注意：虽然OpenRouter并非市场横截面，但其数据显示了自我选择“模型路由”用户的行为。

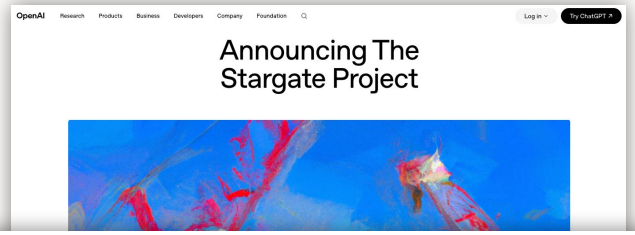
# 在价格压力下，实验室开始拓展应用程序和基础设施领域。

今天

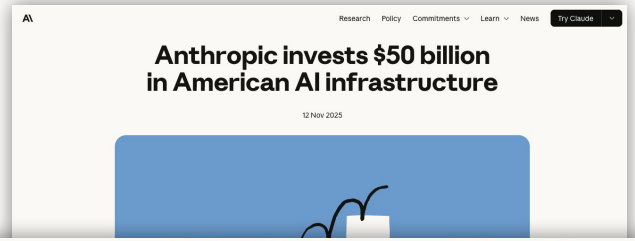

构建垂直应用：法律

LEGORA
Clio
H
法律典籍
克劳德法律


As Anthropic put it in its announcement: “An engagement might begin with the company’s engineering team sitting down with clinicians and IT staff to build tools that fit into the workflows that staff already use... Engagements like this will run across mid-sized companies across industries, each shaped by the people closest to the work.”




The Stargate Project is a new company which intends to invest $500 billion over the next four years building new AI infrastructure for OpenAI in the United States. We will begin deploying $100 billion immediately. This infrastructure will secure American leadership in AI, create hundreds of thousands of American jobs, and generate massive economic benefit for the entire world. This project will not only support the re-industrialization of the United States but also provide a strategic capability to protect the national security of America and its allies.




Today, we are announcing a $50 billion investment in American computing infrastructure, building data centers with Fluidstack in Texas and New York, with more sites to come. These facilities are custom built for Anthropic with a focus on maximizing efficiency for our workloads, enabling continued research and development at the frontier.

来源：Exponential View分析；TechCrunch；OpenAI；Anthropic。

# 压缩每一层，消费者获取剩余价值。

今天


场景1： **开放式重量模型迎头赶上**

*无法收取许可证费用的实验室*
*边界模型*

*对应用程序来说，使用代币更便宜 &*
*主承办服务*

*基础模型实验室构建应用层*
*以及他们自己的基础设施*

场景2： **通用模型**

*边界模型可以替代人工智能*
*应用层/集成层*

*前沿模型实验室承担集成工作；*
*通用“外壳”定价权受到挤压*

*应用程式以专有数据防卫。*
*特定领域的工 作流程与自有模型*

场景3： **降低计算需求**

*模型变得更小，计算*
*更高效*

*模型和应用程序的运行成本更低，从而扩大了*
*适用市场*

*部分本地托管，芯片已分发*
*边缘设备*

场景4： **全部三个合在一起**

*效率的完美风暴*
*改进*

*成本降低扩大了适用市场*

*显著的消费者剩余作为收益*
*物美价廉*

真实需求与需求价格弹性：成本降低扩大了市场规模，并增加了消费者剩余。

人工智能的需求验证了其收入，这比
以往任何平台转型都更为有效。

投资的价值在于，不断下跌的价格能否推动足够的代币交易量，从而收回资本支出成本。

# 方法论：我们如何核算收入、资本支出和代币

## 我们计入什么，以及我们排除什么

- 我们计算
**收入**
在堆栈的每一层（应用程序（订阅和广告软件）、基础模型API以及作为独立服务销售的AI云和计算服务）每一层都代表着付费客户的实际支出。一家跨越多层的公司（例如，拥有面向客户的应用程序的基础模型提供者）拥有其传统软件，以及来自AI的广告提升（主要是Alphabet和Meta）。我们还将从收入衡量指标中排除资本支出和融资。
- **资本支出**
该部分计入七家堆栈构建者（超大规模企业和新云服务商）的AI相关基础设施支出，包括现金固定资产和租赁。

- **标记**
每个经过处理（输入和输出）的令牌，在所有主要服务提供商和平台上都计入统计。

## 我们如何去重，以及为何要这样做

- **收入**
每一层都会计算，但不会跨层累加：例如，100美元的应用内支出，其中60美元支付给模型提供方，后者又花费30美元用于推理的云托管服务，根据其增值贡献，分别归因40/30/30美元，总和为100美元，避免了重复/三重计算，否则会导致错误的190美元数字。
- **资本支出** 该项资产在资产负债表上仅由实际所有者计算一次。对于由多方共同租赁或由基础模型提供方从超大规模计算服务商处租赁的资产，应归属于运营所有者，而非承租方。

**代币**
- 当推理For一家为其他公司运行基础模型的铸造厂提供时，这些代币将只归功于实际运行的模型，因此通过铸造厂提供的模型不会被重复计算。

## 来源

所有图形都是构建的
**自下而上，源自基础和专业资料**
并与最高层级估算值和替代数据进行了三角测量。
**收入和资本支出**
其依据为公司文件（SEC 10-K、10-Q 和 8-K）及高管披露信息，并通过云归属（例如 Azure 和 Bedrock）进行交叉核实，以确认私营公司的收入是否出现在公开公司的账目中。我们的系统每日扫描和爬取可用资源，以创建、维护和改进数据的广度和深度，并确保来源 **归因与置信度分级**
这个高质量数据集用于构建
**完整的财务模型**
（包括损益表）适用于大公司，而基于驱动因素的模式适用于小公司。
**资本性支出的AI占比**
每家公司均单独核算，并与硅片（芯片供应商收入）、建造成本、业务板块构成以及卖方研究报告进行核对。

# 作者


**阿兹尔·阿扎尔**
创始人

**威廉·吉尔德**
产品经理

**汉娜·佩特洛维奇博士**
资深研究员

**纳撒尼尔·沃伦**
资深研究员

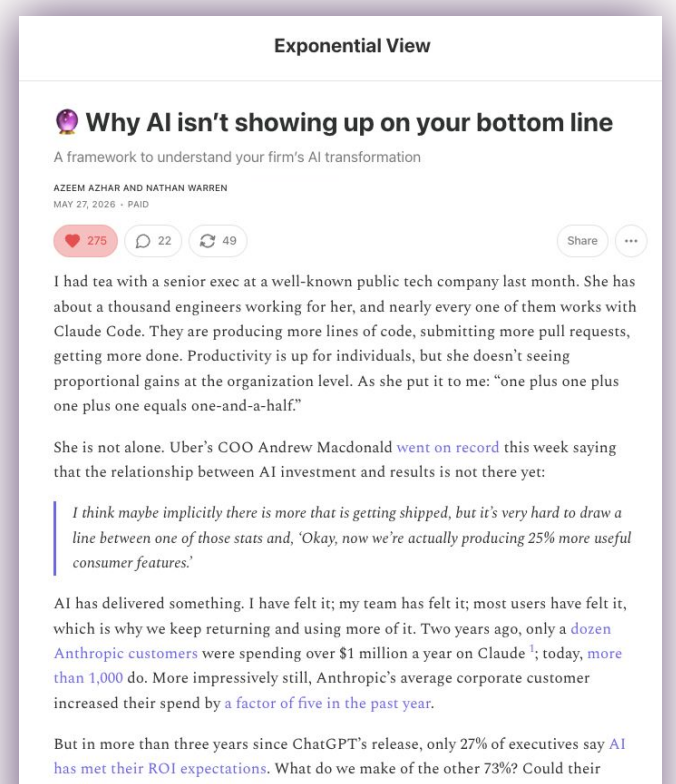
**玛丽亚·加夫里洛娃**
总经理


我们欢迎反馈和贡献。
at aieconomy@exponentialview.co

For advisory requests and institutional inquiries, please contact helen@exponentialview.co

# 请关注我们关于指数观点的分析 (www.exponentialview.co)

订阅指数观，每周在您的收件箱中接收我们的研究。



Exponential View

## 🔮 Why AI isn't showing up on your bottom line

A framework to understand your firm's AI transformation

AZEEM AZHAR AND NATHAN WARREN

MAY 27, 2026 · PAID

275 22 49 Share

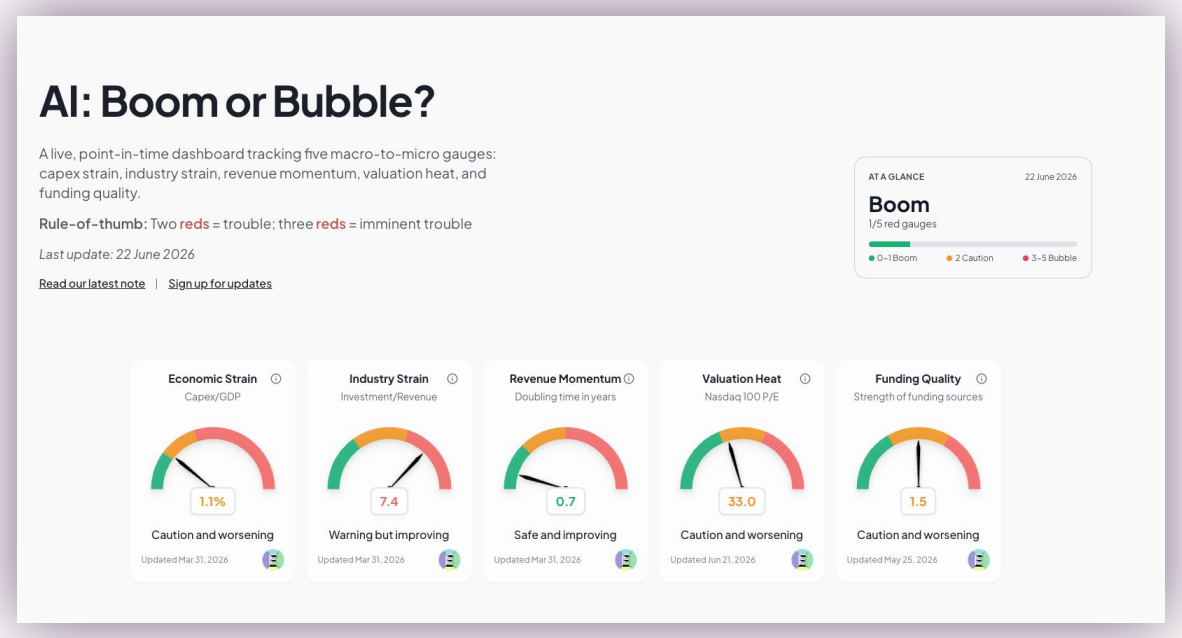
I had tea with a senior exec at a well-known public tech company last month. She has about a thousand engineers working for her, and nearly every one of them works with Claude Code. They are producing more lines of code, submitting more pull requests, getting more done. Productivity is up for individuals, but she doesn't seeing proportional gains at the organization level. As she put it to me: "one plus one plus one plus one equals one-and-a-half."

She is not alone. Uber's COO Andrew Macdonald went on record this week saying that the relationship between AI investment and results is not there yet:

> *I think maybe implicitly there is more that is getting shipped, but it's very hard to draw a line between one of those stats and, 'Okay, now we're actually producing 25% more useful consumer features.'*

AI has delivered something. I have felt it; my team has felt it; most users have felt it, which is why we keep returning and using more of it. Two years ago, only a dozen Anthropic customers were spending over $1 million a year on Claude [1]; today, more than 1,000 do. More impressively still, Anthropic's average corporate customer increased their spend by a factor of five in the past year.

But in more than three years since ChatGPT's release, only 27% of executives say AI has met their ROI expectations. What do we make of the other 73%? Could their



## AI: Boom or Bubble?

A live, point-in-time dashboard tracking five macro-to-micro gauges: capex strain, industry strain, revenue momentum, valuation heat, and funding quality.

**Rule-of-thumb:** Two reds = trouble; three reds = imminent trouble

*Last update: 22 June 2026*

Read our latest note | Sign up for updates

AT A GLANCE 22 June 2026

**Boom**

1/5 red gauges

0–1 Boom · 2 Caution · 3–5 Bubble

| Economic Strain | Industry Strain | Revenue Momentum | Valuation Heat | Funding Quality |
|---|---|---|---|---|
| Capex/GDP | Investment/Revenue | Doubling time in years | Nasdaq 100 P/E | Strength of funding sources |
| 1.1% | 7.4 | 0.7 | 33.0 | 1.5 |
| Caution and worsening | Warning but improving | Safe and improving | Caution and worsening | Caution and worsening |
| Updated Mar 31, 2026 | Updated Mar 31, 2026 | Updated Mar 31, 2026 | Updated Jun 21, 2026 | Updated May 25, 2026 |