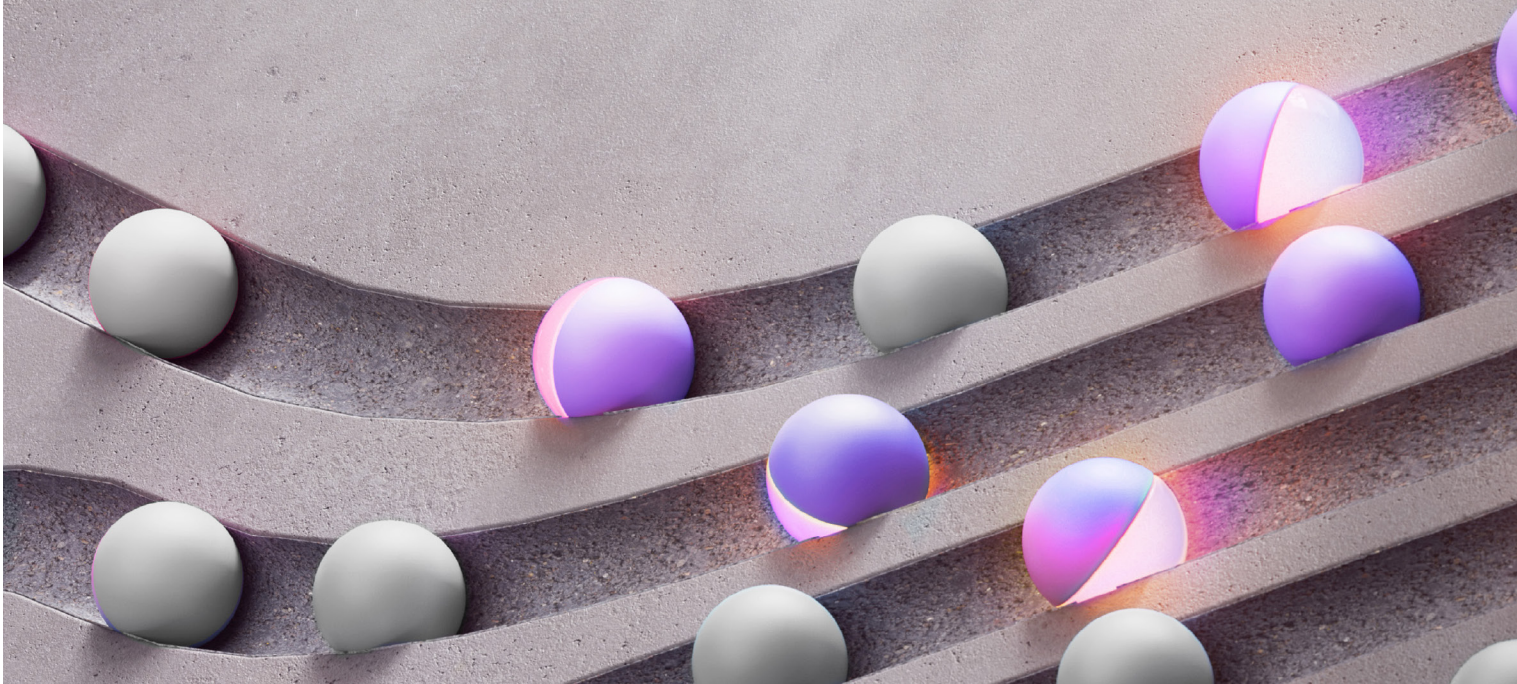




词元经济学 (Tokenomics) : CIO行动指南

从AI投入到价值产出



作者



阿乔伊·梅农
(Ajoy Menon)

埃森哲资深董事总经理，
数字核心主管



关岚
(Lan Guan)

埃森哲首席AI与数据官



詹姆斯·伯罗斯
(James Burrows)

埃森哲资深董事总经理，
数字核心，技术战略主管



苏瑞亚·慕克吉
(Surya Mukherjee)

埃森哲商业研究院研究总监

人工智能 (AI) 正在催生一项新的经营成本：词元 (Token)。随着AI应用规模不断扩大, 词元经济正逐渐演变为一项兼具风险与机遇的重要议题, 并日益成为管理层和董事会关注的焦点。

董事会要求CEO厘清AI投入的风险与回报, CEO则希望CFO建立相应的管控机制, 而CFO又将目光投向CIO, 期待其能够让AI资源消耗更可预测、成本可控、价值可量化。那些能够系统掌握并驾驭词元经济规律的企业将比那些仅仅为AI使用买单的企业更具优势。

基于全球调研、分析与咨询实践, 埃森哲发现, AI带来的词元挑战既是成本问题, 更是管理问题。研究显示, 企业若想在控制成本的同时实现AI规模化应用, 需要践行五大要务。领先企业通常会在各类AI应用上线前**建立监测机制** (observability), 确保AI使用情况和资源消耗透明可见; 通过成本分摊 (chargeback) 落实**财务责任**; 根据业务需求**选择最具成本效益的模型**, 实现投入产出最优; 要求所有新增应用在上线前**明确预期价值和衡量标准**; 同时**提升员工和开发人员的应用能力**, 使其能够合理选择模型并高效运用AI。这些企业将AI词元视为一项需要精细化运营的战略资源, 其管理标准与资本支出、云基础设施和软件资产等关键技术投入保持一致。

从本质上看, 词元已经成为AI时代的新型生产成本, 而管理这一成本正逐渐成为CIO的重要职责。CIO需要协调财务、业务、人力资源等关键职能, 将成本管理、价值衡量和资源配置纳入统一框架, 从而推动企业实现AI的规模化应用。

什么是词元经济学 (Tokenomics) ?

词元经济学基金会 (Tokenomics Foundation) 将AI词元经济学 (AI Tokenomics) 定义为一套研究如何将能源和资本转化为AI能力, 并通过高效利用AI服务, 实现可量化业务价值的管理方法¹。词元是AI处理信息的基本单位。每一次提问、回答、信息检索和智能体交互, 都会消耗词元。本报告及其中概述的各项要务, 均围绕词元的成本与价值展开。

词元治理带来的收益十分可观。我们的研究显示, 在未来24个月内, 率先建立系统化词元管理机制的企业, 有望将词元成本涨幅降低44%², 按年词元支出1000万美元估算, 企业可避免约440万美元的成本增长; 若年支出达到1亿美元, 则可节约4400万美元。

除此之外, 企业还能破解价值量化难题: 目前, 每1美元的词元支出中, 只有20美分能够量化为财务成果向董事会和CFO汇报。通过建立成本分摊和应用级监测机制, 企业可将这一比例提升至32美分, 这意味着每1000万美元支出中, 约115万美元的业务价值可首次被验证。综合来看, 在无需部署新模型或增加AI投资的前提下, 仅提升管理能力, **每1000万美元年支出即可收获约560万美元的综合收益 (包括成本节降和新增可验证价值); 年支出1亿美元的企业, 这一收益可达约5600万美元。**

许多企业仍然沿用传统软件时代管理思维来管理AI, 即在成本产生后再关注支出情况。这种模式在今天已难以为继。随着企业支出模式从固定的软件许可费用转向按量计费的AI服务, 且员工、智能体, 以及各类复杂工作流对AI的使用持续增长, 词元成本已成为经营变量。

随着词元成本上涨, 董事会越来越关注三个问题: 是什么在推高AI成本? 这些成本的增速有多快? AI创造了哪些业务价值? 许多CIO目前仍难以给出清晰答案。这很大程度上是因为他们缺乏对词元消耗来源的可视化管理, 无法说明哪些业务场景在大量消耗词元、哪些模型导致了这些消耗, 以及这些消耗带来了怎样的业务价值。此外, CIO往往以技术语言而非财务语言阐述问题, 而后者恰恰是CFO协助CEO向董事会汇报所需要的语言, 这进一步增加了成本归因、风险管控和AI规模化应用的难度。

践行前述五大要务的企业将词元视为需要主动管理的生产要素，而不是事后审查的成本项目。借助这一体系，CIO便可在控制成本、验证业务价值的同时，实现企业AI规模化落地。

研究概览

我们于2026年7月对来自17个国家及地区、年营收超10亿美元企业的750位高管开展了调研，并于2026年6月至7月间，对《财富》500强企业的技术和财务主管进行了15次深度访谈。受访者对其所在企业如何管理AI资源消耗、优化、调配、问责，以及价值衡量进行了评估，为理解AI部署背后的经济逻辑和规模化治理机制提供了重要洞察。本次研究过程运用了生成式AI，我们的研究专家恪守埃森哲负责任AI企业规范，尽可能地以传统研究方法对生成式AI输出的结果进行了审查和验证。

1	业务挑战	07
2	解决之道	22
3	行动建议	27
4	结语	35
5	埃森哲助您布局 词元经济	36



业务挑战

CIO面临的新命题

根据我们的调研,过去一年,受访企业在AI词元上的总支出约为25亿美元。如果不加干预,这一数字将在24个月内攀升至36亿美元³。企业间的支出规模差异显著。例如:某全球科技公司每年支出逾2000万美元,某欧洲保险公司的AI推理预算达到1.5亿至2亿美元,某政府承包商则将其全球AI支出上限设定为2.5亿美元。鉴于企业所处发展阶段和组织规模不同,词元支出水平存在较大差异,因此本报告以每年1000万美元词元支出为基准。对于年支出为3000万美元的企业,可将相关测算结果乘以3;若支出达3亿美元,则乘以30。在各类AI成本驱动因素中,词元支出目前位列第三,仅次于基础设施以及软件开发与维护成本。

就当前规模来看,词元支出本身或许尚不足以引起董事会的重视,但其增势不容忽视。随着用量攀升,成本不确定性也会随之上升。词元支出很可能在短时间内从一项运营成本,演变为需要CFO重点管控的财务风险。这种压力会逐级传递,CFO需要向CEO阐明潜在风险,而CEO也越来越频繁地需要向董事会汇报。

归根结底，管理层的顾虑集中在两点：一是词元用量及相关支出难以按季度精准预测；二是AI投入难以清晰证明其所创造的业务价值。由此形成了一种两难局面：当成本缺乏可预测性，财务部门便不愿追加预算；而当价值缺乏可见性，管理层就找不到理由持续投入。因此，CIO不得不肩负起一项新的财务使命：在AI成本问题升级为董事会议题前，让AI的经济效益可预测、可衡量、可管控。

真正需要警惕的是增长趋势。调研显示，在24个月内，受访企业的词元消耗量预计将平均增长78%，而单位词元价格仅下降19%。

这种担忧绝非杞人忧天。

优步 (Uber) 曾仅用四个月就耗尽了全年AI词元预算⁴。一位来自全球金融数据与媒体公司，负责云、平台和安全架构的CTO坦言：

“这是实打实的难题……每天、每分钟都在增长……追踪（词元用量和支出）极具挑战，尤其当AI应用分散在多个业务部门和团队。”

美国一家大型金融机构的技术负责人透露，其AI词元支出在短短半年内增长了5~10倍。即便按最低增幅估算，原本1000万美元的年支出，也将在一年内攀升至5000万~1亿美元；词元价格的下降，此刻不过是杯水车薪。同样，一家欧洲大型银行的年度词元支出在一年内从几乎为零增长至500万欧元（约580万美元），在人工干预前，仅一项工作负载的支出就曾达到每周20万欧元（约23.1万美元）。

有三分之一的企业在年底前就用完了全年词元预算。**词元支出不会自行稳定下来，沿用过去那种“先使用、后结算”的管理模式将难以为继** (图1)。

图1 三种增长情景揭示词元支出为何无法自行稳定

词元支出不会自行稳定

当前, 每1000万美元词元支出……				当前优化率: 23% 盈亏平衡点: 31%
情景	年度支出规模及增长逻辑			管理启示
1. 管控式增长 需将优化率提升8个百分点, 达到31%这一盈亏平衡点	支出维持1000万美元不变 24个月内词元消耗量上涨78% 单位价格下降19% 优化率达到31%			增长带来的成本压力是可以被部分抵消的, 但前提是在用量激增前, 企业已提升模型调用、缓存复用、提示词管理等能力。
2. 放任式增长 未采取优化举措	24个月内支出达到1440万美元 24个月内词元消耗量上涨78% 单位价格下降19% 未采取优化举措			尽管词元单价有所下降, 但用量涨幅超过降价幅度, 年度词元支出仍将增加约440万美元。
3. 爆发式增长 六个月增长5-10倍	年支出达到5000万-1亿美元 基于美国某大型金融机构的快速落地案例			传统年度预算机制失效。需要在工作负载层面持续管控成本、主动干预。

数据来源: 埃森哲词元经济学调研 (样本量=750)。

注: 成本走势取决于管理方式。以每年1000万美元词元支出为基准, 管控式增长可将支出维持在当前水平; 放任式增长将使支出增至1440万美元; 而爆发式增长则可能将年度支出推高至5000万美元以上, 而此时词元价格下降所带来的成本抵消效应尚未发挥作用。情景1、2为调研假设。情景3基于访谈压力测试情景, 非预测结果。

词元消耗主要来自三类业务场景: 员工和开发人员使用的工具 (约占支出的30%); 内部流程自动化和企业软件 (约占29%); 以及服务于客户的应用、分析和智能体工作流 (约占39%)。其中, 智能体工作流目前占比最小, 仅为11%, 但其每个步骤都会触发推理调用, 词元消耗极易成倍飙升, 因此最有可能成为下一轮预算外成本上涨的主要驱动因素。



词元问题的一体两面: 成本管控与价值证明

企业面临的两个核心挑战, 一是成本难以预测, 二是价值难以验证。看似是两个独立问题, 实则源于同一个管理盲区: 企业看不清哪些AI应用在消耗资源, 也难以准确评估这些投入创造了哪些业务价值。

1. 支出难以预测

企业真正担忧的并非词元支出的绝对金额, 而是越来越难以预测、管控和解释这笔支出。与此同时, 词元用量的增速远超价格降幅, 且这一差距仍在持续拉大。在智能体工具普及之前构建的预算模型已无法应对这种变化。

问题已经在领先实践企业中显现。一家亚洲大型银行的资深工程负责人向我们分享了一项经历: 该行为开发团队部署了一款AI编码助手, 最初预期生产效率提升带来的收益将远超词元成本。然而三个月后, 实际支出显著高于预期。数千名工程师不约而同地选用了最强大的模型, 而很多工作低成本模型足以胜任。没有人做错什么, 现有预算机制根本无法预见这种情况的发生。

2. 价值难以验证

另一方面, AI创造的价值真实存在。企业正在提升生产效率、加快决策流程、改善客户体验。问题在于, 这些回报目前尚不能以财务语言与词元支出明确关联。因此, CFO拿不到决策依据, CEO无法验证投资价值, 董事会也难以评估投入产出, 这条价值链条尚未真正打通。

研究显示,企业每5美元的词元支出中,仅有1美元能够量化为财务成果,例如营收增长、成本节降,或资本效率提升。企业普遍认为其余4美元同样创造了价值,只是无法有效验证。价值确实存在,缺少的是一套能够将其转化为财务指标的内置追踪机制。

即便是在AI应用较为成熟的企业中,价值归因依然是一项挑战。一家《财富》500强金融服务机构的AI战略总监道出了许多企业的共同困境:AI为整个企业创造的价值有目共睹,但这些价值却难以映射到所消耗的词元上。生产力显著提升、业务成果持续改善、应用规模不断扩大,但词元消耗与这些成效之间的投入产出关系,在很大程度上仍难以量化。

要解决这些问题,必须从根源着手:看清每一项AI应用的资源消耗、明确责任主体,并落实财务责任。

AI价格下降,为何支出反而增长

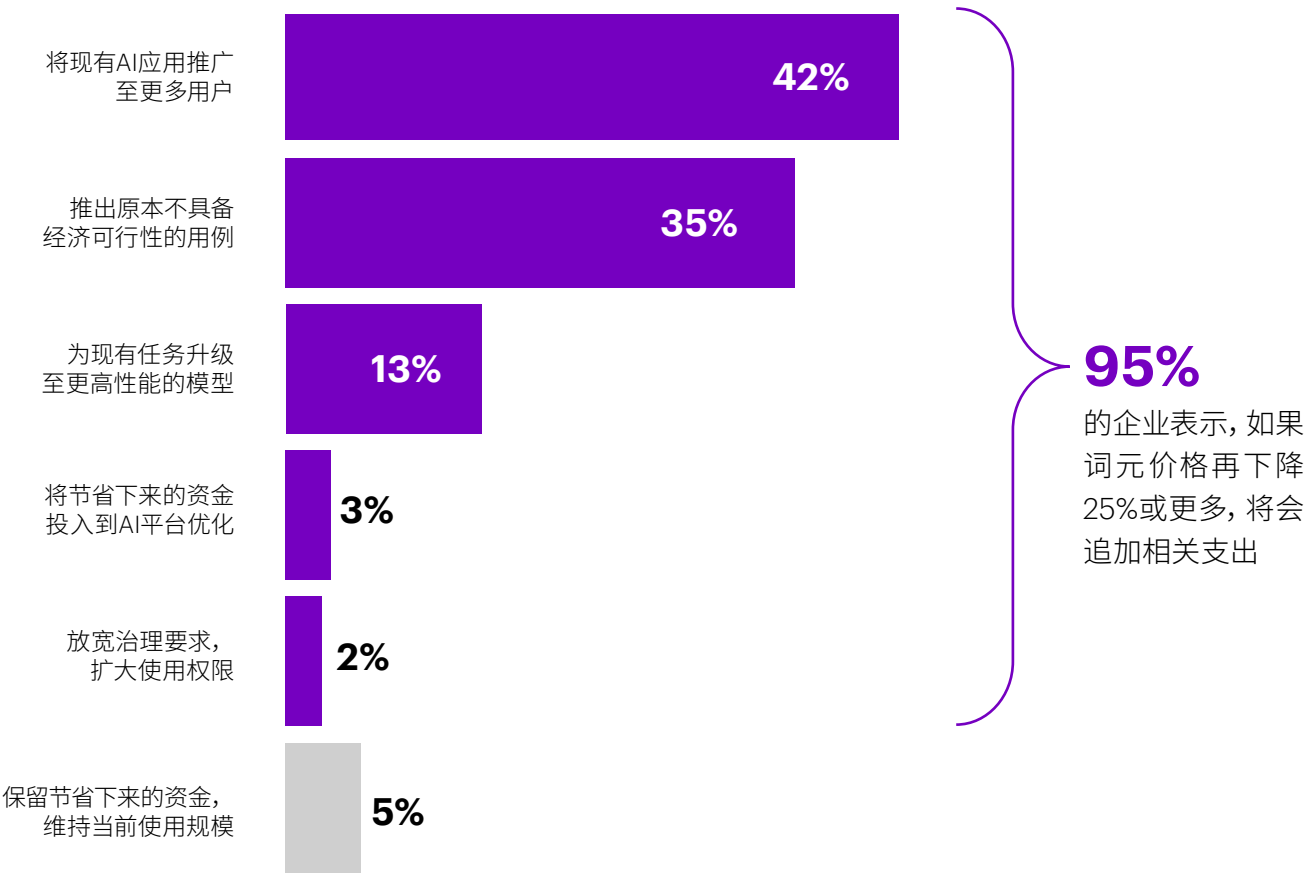
悖论: AI越便宜,企业反而花得越多

单位词元的价格正一路走低。18个月前,每百万词元定价约1美元的模型,如今价格已大幅缩水。按理说,企业的AI支出会随之下降,但现实恰恰相反。

这正是词元经济学中的杰文斯悖论 (Jevons paradox) : 正如更高效的蒸汽机非但没有减少、反而推高了英国的煤炭消耗, 词元降价也促使企业追加AI部署 (图2)。当被问及“如果词元价格再下降25%, 企业将采取何种举措”时, 仅有5%的受访高管选择保留节省下来的资金; 更多企业倾向于加码投入: 触达更多用户、拓展应用场景、开发更高性能的模式。

图2 企业对词元降价的反应

问题: 如果词元价格再下降25%或更多, 您的企业最有可能采取什么举措?



数据来源: 埃森哲词元经济学调研 (样本量=750)。



一位来自全球大型科技基础设施公司, 负责AI、网络安全和数据业务的首席技术顾问 (Field CTO) 精准描述了这一现象:

“即使模型变得更便宜, 团队也只会用得更多。这就好比……你当年第一次拿到1GB硬盘的时候? 当时并没有人会想, ‘天哪, 我们用不了……这么多硬盘空间。’ 不, 使用量反而呈指数级增长, AI的应用逻辑也是如此。”

曾经因成本过高而难以开展的工作, 如今突然变得经济可行, 新的应用场景持续涌现。去年因价格高昂而被搁置的智能体应用, 可以按照今年的价格顺利部署了。用量的爆发式增长完全抵消了价格下降带来的成本红利, 最终导致AI账单不降反升。

总部位于旧金山的法律AI公司Harvey, 为来自1300多家律师事务所和企业法务团队的逾10万名律师提供服务。数据显示, 其月度词元消耗量在2026年1月约为1万亿个, 到同年5月已暴涨至12万亿~13万亿个⁵。

这充分说明, 当AI开始深入核心业务流程并进入规模化生产阶段后, 词元消耗将极速攀升。即便单位成本大幅下降, 也会被激增的用量迅速抵消。

六大盲区导致预算失控

词元的真实成本隐藏在财务团队难以察觉的六大盲区之中（图3）。

图3 六大盲区导致AI预算超支



任务成本难以预估

绝大多数企业成本在投入前即可预估，例如购买软件许可证、服务器、新增人力成本，但AI词元消耗并非如此。一条简单的指令就可能触发一连串智能体调用，消耗数百万个词元，且事前根本无法预估——你无法提前知晓一项任务的成本是1美元还是100美元。



默认使用最强模型

员工习惯性选择最强模型，一方面是他们不了解模型之间的成本差异，另一方面也是因为企业没有提供充分的理由或激励措施，引导他们选择成本更低的模型。

埃森哲对O*NET数据库内9368项工作任务进行了分类研究，发现真正需要前沿模型能力的任务不足10%，其余90%都在占用单位词元成本高出实际所需10~20倍的模型来处理。正如某全球大型科技公司的企业AI战略副总裁所观察到的那样：

“大约10%的人会不断要求使用最新、最强大的模型，而组织中的大多数员工其实只是把AI当作搜索工具使用。”

智能体倍增效应

智能体工作流会大幅增加词元消耗。一次请求可能会触发多轮模型调用、不断扩展对话场景并反复处理同一信息。随着工作流愈发复杂，成本上涨速度甚至会超过用量增速，传统预算机制因此逐渐失灵。

某企业在大规模部署AI编码助手后发现编码成本增长远超预期。深入分析后，它们意识到问题并非出在模型价格，而在于架构设计。随后，该企业通过两项举措遏制了成本持续上涨：将开发人员从高级模型切换到足以胜任大多数编码任务的低成本模型，同时开启提示词缓存，避免每次交互都重复读取庞大的代码库。

仅模型切换这一项优化举措就减少了约90%的词元用量，缓存启用则进一步节省了成本。AI成本更多地取决于模型选择、上下文管理和架构设计，而非模型本身的价格。因此，用户培训和架构层面的指引对于成本管控至关重要。

预算机制过时

大多数企业的AI预算制定于智能体普及之前。结果，有三分之一的企业在年底前就用完了全年词元预算。某全球快餐连锁企业的首席技术与信息官 (CTIO) 回忆道：

“在项目初期，我们并不清楚会花费多少预算……有些团队明显超支了，实际支出是预算的两到三倍。”

价格体系分散

大型企业普遍同时运行着来自不同服务商的模型，而各家在定价、上下文窗口机制、缓存策略和输出词元费率上差异显著，企业很难以全局视角管控AI成本。此外，新模型屡屡在毫无预告的情况下发布，使得企业基于现有模型做出的成本预测很快失效。价格体系的分化速度与模型迭代一样快。

一家欧洲银行的高级技术主管告诉我们，服务商越来越倾向于按照应用场景创造的业务价值、而非词元用量来给AI定价。

呼叫中心解决方案的成本约为通用AI平台的10倍，专业法律工具的定价甚至更高，这是因为服务商会参照AI所替代的专业人士的薪资水平来定价。随着定价逻辑从“按用量计费”逐步转向“按价值收费”，即使处理的是类似任务，费用也可能截然不同。对单个服务商做成本预测或许更简单了，但企业整体预算管理将变得更加复杂，总支出也可能进一步攀升。

此外，服务商的定价机制还可能导致无法预见的成本跃升。专注于B2B客户支持与智能体AI的Pylon公司发现，随着授权用户突破150位并采用企业版定价模式，其年度词元支出预计将从约40万美元飙升至140万美元。这意味着财务团队需要提前将服务商的定价层级和门槛纳入预算模型，而不是等到续约时才发现成本大幅上涨⁶。

成本可见但责任缺失

企业的词元消耗量通常隐藏在服务商的后台看板中，财务和开发人员无权访问，管理员也不会实时查看。成本展示 (showback) 仅展示用量、不绑定财务责任；成本分摊 (chargeback) 会按照实际用量向使用团队计费，从而促使其主动减少消耗。目前，超过半数的企业采用成本展示机制，只有7%做到了成本分摊。成本展示有助于培养成本意识，与关键绩效指标 (KPI) 等结合时方能发挥作用；成本分摊则可以建立更直接的问责机制，不过仍需配套其他激励措施，才能真正推动行为改变。

80%的支出无法量化为业务成果

财务部门仅能量化不到五分之一的支出回报

当管理层尝试用可量化的财务成果——无论是带动营收增长、成本节降，或是资本效率提升——追溯词元支出时，他们发现，能达到这一标准的支出不足五分之一。换言之，每花费1000万美元，仅约200万美元能够与明确的业务成果建立量化关联。这使得CFO难以有效管理投资、CEO无法充分验证投资价值，而这正是董事会日益关切的问题（图4）。

图4 仅有少部分AI回报能向董事会清晰呈现

CFO仅能将不到五分之一的支出与业务成果挂钩

CFO对每一笔支出都了如指掌，
但仅能对不到五分之一的回报加以量化验证

年度词元支出

作为一项成本
完全可见

每**1000万美元**中……→

200万美元

价值回报可被
量化验证的支出
约占20%

水位线——CFO可以捍卫什么

800万美元

无法量化证明
价值的支出
约占80%

数据来源：埃森哲词元经济学调研（样本量=750）。

AI价值归因的现实困境

从成本端来看，AI投入相对容易拆解。埃森哲调研显示，词元占AI总支出的比例通常不超过25%，且能够清晰识别和追踪到具体AI应用及团队。

但AI创造的价值却贯穿整个工作流，其带来的生产力提升往往是多重因素共同作用的结果，包括所使用的模型、检索的数据、支撑运行的基础设施、设计工作流的工程师，以及终端用户。无论多么细致地衡量，都很难精准计算出价值回报中有多少来自词元本身。

一位来自全球大型科技基础设施公司，负责AI、网络安全和数据业务的首席技术顾问一语道破：

“真正的经济学问题，不是消耗了多少（词元），而是这些词元究竟创造了什么价值？”

48%的企业采用IT与财务部门联合责任制，但没有明确的责任人统筹AI投入与产出。技术团队虽能查看各部门或个人的词元用量以及对应成本，却无法判断这些消耗是否真正创造了与之相称的业务价值。这一判断应由业务负责人作出，因为他们最清楚哪些工作流程、客户交互和业务决策由AI驱动。但现实是，大多数企业尚未建立由业务负责人主导价值评估的机制。

目前，仅有35%的企业能够在其规模最大的应用场景中，将成本与业务成果建立量化关联；59%的企业仍以粗略估算为主。一家《财富》500强金融服务机构的AI战略总监表示：

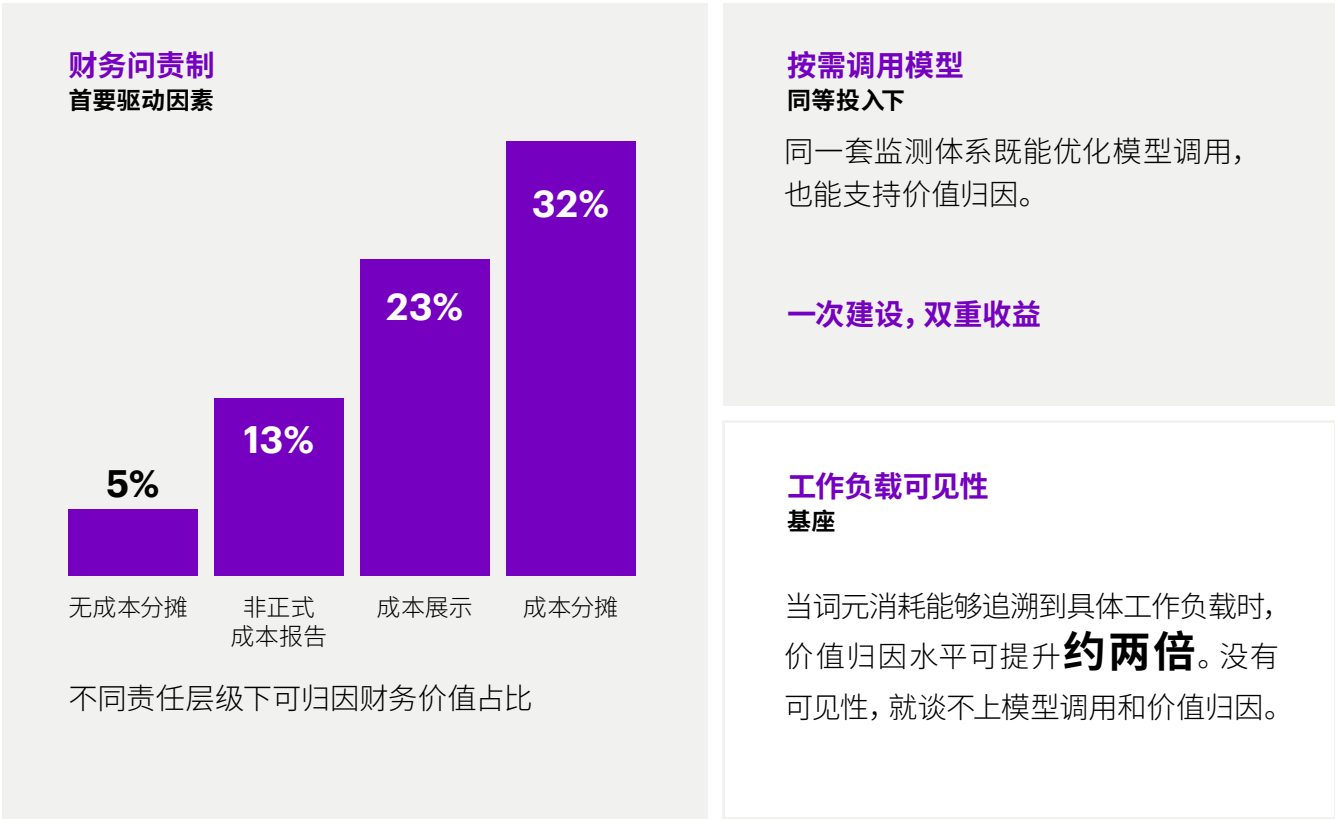
“我认为，我们当前绝大多数AI项目确实在创造价值，但这些价值并不是通过词元消耗来衡量的。”

关于哪些因素能够推动财务价值归因，数据给出了清晰且可落地的结论（图5）。

图5 问责制、按需调用与可见性让AI价值得以证明

三大因素驱动财务价值归因

财务问责制是首要驱动因素；
责任越明确，价值归因水平越高



数据来源：埃森哲词元经济学调研（样本量=750）。



财务问责制

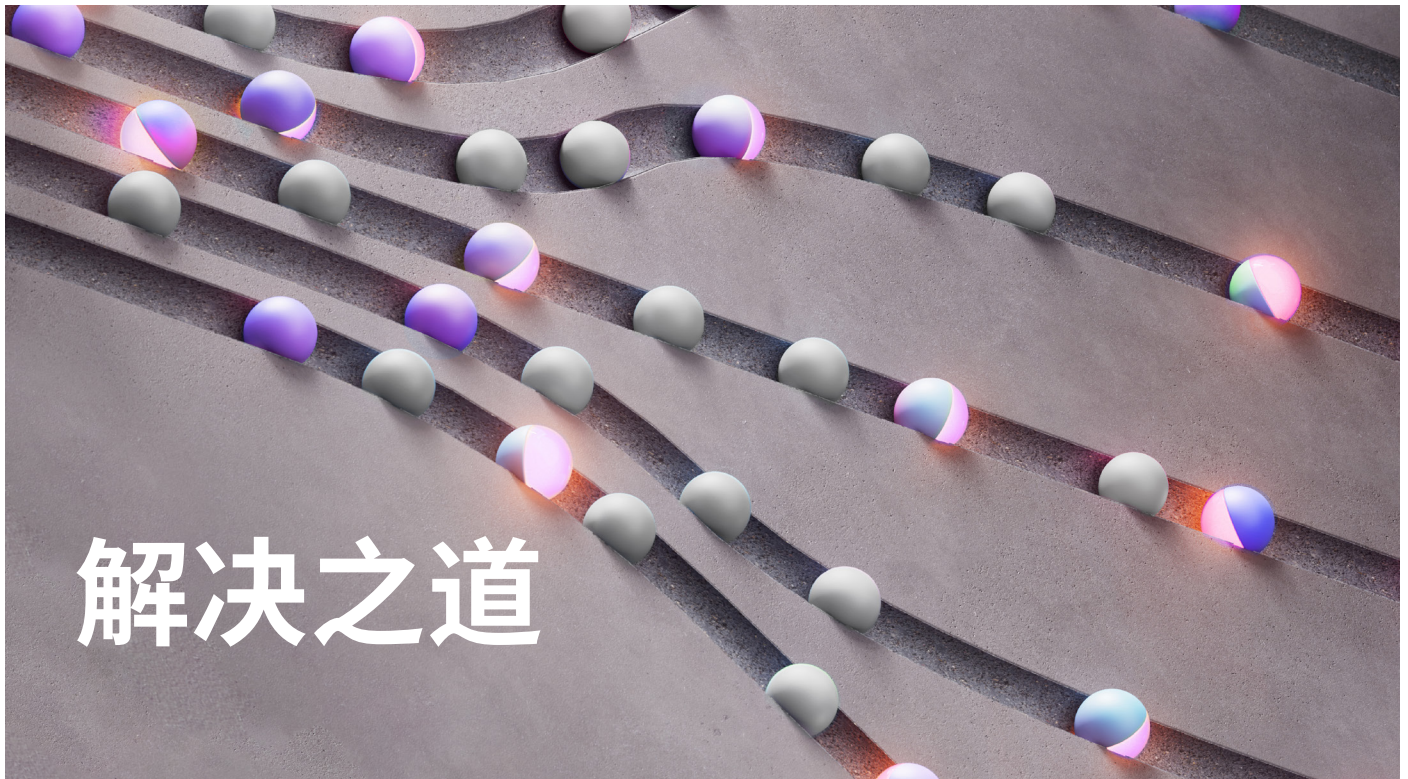
财务问责制是AI价值量化的关键驱动因素之一。不论企业所处行业、规模或发展成熟度如何，从无成本分摊到全面成本分摊，财务问责制每提升一个层级，可归因财务支出约提升6~7个百分点。在无成本分摊时，可归因财务价值仅约占5%；采用非正式成本报告后可提升至13%；成本展示下达23%；实现全面成本分摊则可进一步提升至32%。

按需调用模型

支撑模型选择的前提，是企业能够看清AI资源究竟消耗在什么地方。只有掌握每项AI应用运行在哪个模型上、花费了多少成本、创造了哪些业务价值，才能在成本与效果之间找到最佳平衡。模型调用与价值归因并非两项独立的工作，而是同一套管理体系的两个方面，需要同步规划、同步推进。

工作负载可见性

能够将词元消耗追溯至具体用户、团队或业务场景的企业，其整体价值归因能力是其他企业的近两倍。因为只有先看清资源消耗发生在哪里，企业才能进一步优化模型使用，并评估这些投入是否真正创造了业务价值。



解决之道

只需一套管理体系，释放超过50%的价值空间

研究发现，有四项关键管理实践能够同时提升AI投入产出效率：明确业务责任、建立成本分摊机制、强化效果评估，以及按需调用模型。这些举措分别作用于成本端和价值端，但依托同一套治理能力：让每项AI应用都有明确责任人、监测机制和问责制；一套机制建设，既能管住成本，也能看清价值。

明确每项AI应用的负责人，有助于提升成本可预测性和价值可衡量性；缺少明确的责任主体，成本便无从追溯，价值也难以归因。成本分摊机制促使团队主动精细化管理成本、验证价值回报。财务问责制不仅能够强化成本意识，也是企业证明投资回报的驱动力。精细化衡量可以降低预测偏差，识别出真正创造业务价值的AI应用场景，为下一轮投资决策提供可靠依据。根据业务需求匹配合适的模型，不仅能够有效控制成本，也有助于看清不同AI应用究竟创造了多少价值。

一位来自全球金融数据与媒体公司，负责云、平台和安全架构的CTO分享了他们的实践经验：

“我们刚刚完成了一项重要工作，将责任归属细化到组织、团队层级，在大多数场景下，我们甚至可以追溯到个人。这样做的目的很简单，就是从项目一开始就把成本管理机制建立起来，让每一笔支出都有人负责。”

建立治理机制能带来什么回报

目前，大多数企业的词元消耗优化率为23%。我们的模型显示，在未来24个月内，词元用量将增长近80%，企业需要把优化率提升至约31%，才能将支出维持在当前水平，这中间存在8个百分点的提升空间。如果不采取任何措施，1000万美元的年度支出在24个月内将增长至1440万美元，在36个月内将进一步攀升至1730万美元。

这套治理体系同样将带来价值提升。现阶段，每1美元的词元支出中，仅有20美分能够转化为财务回报。通过实施成本分摊和精细化衡量，这一比例可提升至约32美分。对于每1000万美元支出而言，相当于有约115万美元的价值能够首次得到量化验证。在CFO看来，这是CIO必须解决的问题：不是单纯地削减词元开支，而是要建立完善的衡量体系和管理机制，让每一笔投入产出都能在财务账目上清晰呈现。

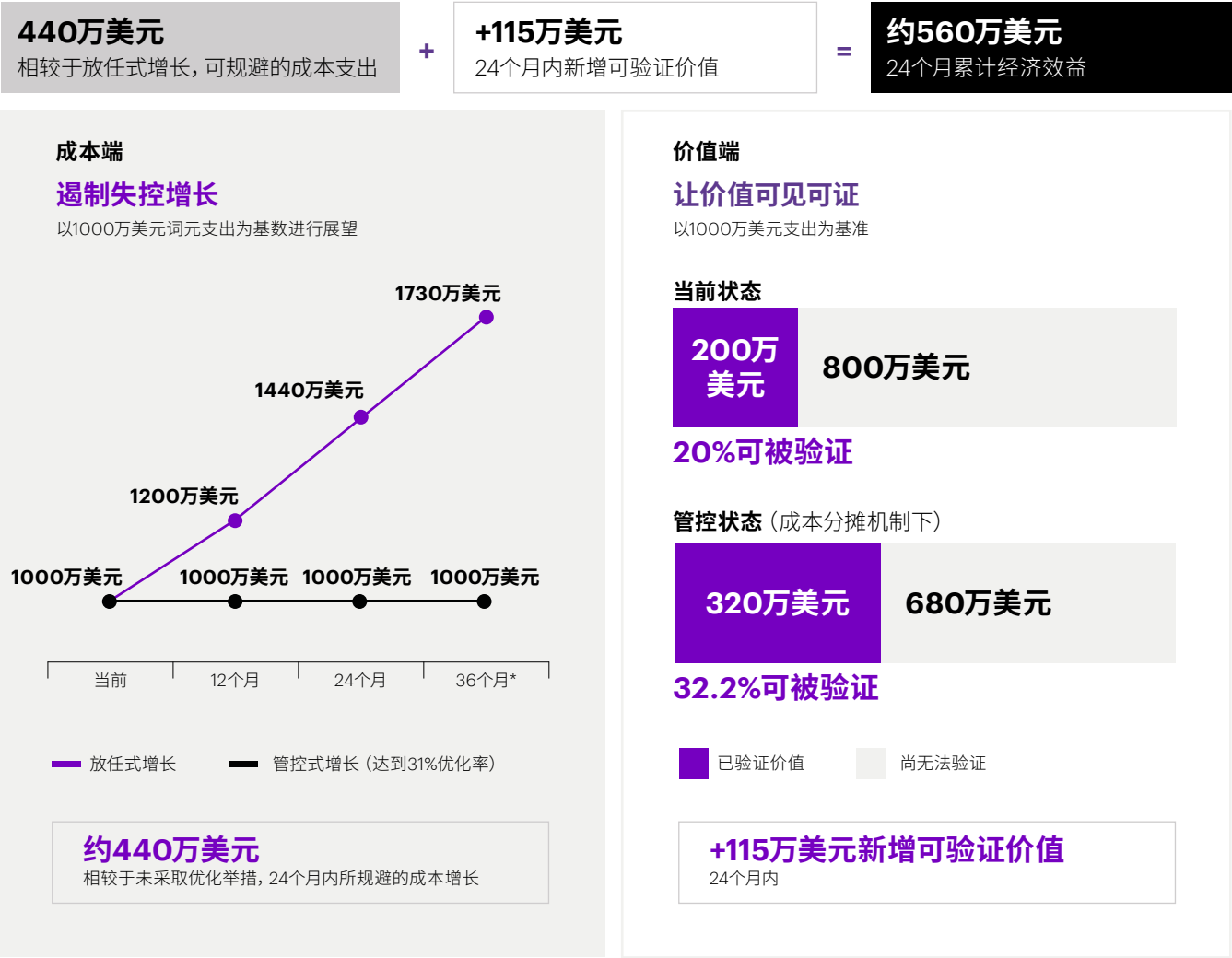
一套治理体系实现双重收益：每1000万美元的年度支出，可规避约440万美元的成本增长，同时创造约115万美元的可验证价值。两项收益来自同一套管理体系，明确每项AI应用的责任主体，建立相应的成本管理机制，并持续跟踪投入与产出之间的关系。

已建立起这套能力体系的企业，关注的不再只是成本支出，更是价值回报（图6）。

图6 一套治理体系、双重收益、复合效应

同一套词元经济学原理，既能抑制成本增长，又能规模化验证价值创造

每1000万美元词元支出……



灵活调配开放权重模型和专有模型以优化成本

前文主要讨论了大多数企业目前所采用的模式，即依托服务商API或云平台按词元付费，这种模式覆盖了70%的企业。

然而，决定AI成本的并不只是词元消耗量，部署架构的选择对成本的影响远超大部分CIO的预期。企业可以选择在CPU或自有GPU环境中部署开放权重模型，也可以在公有云上运行开放权重模型；或是通过云平台、直接调用API的方式使用专有模型。不同部署方案在成本结构、响应速度、控制能力、模型性能以及基础设施运维压力等方面各有优劣，企业需要结合业务场景综合权衡（图7）。

图7 不同业务场景对应不同的成本结构

模型部署	成本结构	适用场景	利弊权衡
基于CPU的开放权重模型	推理成本极低；但基础设施、人力和运维成本较高	大量、低复杂度的任务，如分类、摘要总结和信息提取	响应速度较慢；较高的机器学习运维（MLOps）成本
基于自有GPU的开放权重模型	规模化场景下推理成本较低；但基础设施、人力和运维成本较高	大规模、高频次的标准化工作负载	前期投入高；需自行承担全部基础设施建设与运维压力
公有云开放权重模型	推理成本较低；但需额外承担云基础设施和管理成本	希望依托开放权重模型获得成本优势，同时规避基础设施运维压力的企业	增长最快的模式；控制力弱于本地部署
通过云平台使用专有模型	按词元计费；基础设施和运维成本已包含在内	大多数生产级工作负载	目前企业主流模式；可获得前沿模型能力；享受企业级服务协议（SLA）保障
直接调用模型服务商的专有模型	按词元计费；基础设施和运维成本已包含在内	直接调用API	在大多数应用场景下，其功能与云平台等效

数据来源：埃森哲分析。



对于约90%无需前沿推理能力的任务，开放权重模型在成本方面极具优势。调研数据也印证了这一趋势：与主要依赖专有模型的企业相比，那些能灵活调配开放权重模型和专有模型的企业，在模型选择方面更加成熟。平均来看，这类企业有55%的AI请求可匹配到最合适的模型，而主要依赖专有模型的企业这一比例仅为47%；与此同时，61%的企业实现了AI生产级规模化应用，专有模型组的比例仅为32%。

除成本效益外，开放权重模型还具备三项核心优势：能够更好地保障数据主权、抵御前沿模型出口管制风险，以及微调后在非英语和特定业务场景上表现更佳。

但开放权重模型并非没有代价。根据埃森哲调研，词元占AI总支出的比例通常不超过25%；采用自有开放权重模型往往需要企业承担额外的基础设施、人力和运维成本。此外，企业对商业级服务保障的需求和自身能力缺口也长期制约着开放权重模型的广泛应用。尽管随着新一代开放权重模型和基础设施不断成熟，这些门槛正在逐步降低，但对于许多企业而言，相关管理成本依然不可忽视。

因此，开放权重模型虽然是优化成本的重要选择，却未必适合作为企业AI体系的唯一答案。对于缺乏专门团队和运营能力的企业而言，开放权重模型带来的词元成本优势很可能会被上述其他成本所抵消。

更合理的做法是制定精细化选型策略：根据工作负载按需匹配模型，在成本、性能、数据治理和运营复杂度之间寻求最佳平衡。

行动建议

CIO的五项要务

基于全球调研和客户实践，埃森哲提炼出支撑企业AI规模化落地的五项关键要务：看清每项AI应用的成本与价值、财务问责制、智能调用、价值验证与AI人才。这五项要务构建起一套统一的运营治理体系，各能力相互协同，并非需要按固定顺序推进的线性任务。工作负载可见性是其余四项要务的基础，因此应优先建设；其余要务可视企业自身AI成熟度同步或分阶段推进。

■ 要务一： 在规模化扩展之前，看清每一项AI应用的成本和价值

企业应部署AI网关或监测机制，对每一次AI交互进行记录和追踪，包括工作负载名称、责任团队、所用模型、词元成本及输出结果。当AI进入规模化应用阶段后，企业需要通过自动化治理规则和策略控制机制开展管理，而不能再依赖人工逐项审核。因为无论是成本管理还是价值衡量，都建立在一个前提之上：先看得见。

LangSmith、Arize和Helicone等工具能够以较低的成本，将AI成本追踪到具体应用、团队和业务场景⁷。

大量企业实践证明，过晚建设这类能力可能会付出高昂代价。优步曾鼓励工程师“尽可能地”使用AI，甚至设立了内部词元用量排行榜。结果到2026年4月，公司仅用四个月就耗尽了全年AI预算，最终不得不为每位员工设置了1500美元的使用上限⁸。

Shutterstock的案例进一步印证了尽早构建监测衡量机制的重要性。该公司在审查AI成本结构时发现，输出词元占到总词元消耗的75%；与此同时，ChatGPT许可证数量在一个月内从50个激增至750个。此外，它们还发现有25万美元已获批的AI预算存在闲置风险。面对上述问题，Shutterstock没有简单地限制AI的使用，而是将AI和云成本划归至财务运营部门，在进一步扩展前将相关数据整合到财务统一流程。最终，Shutterstock得以及时掌握AI使用情况和对供应商的支出承诺，发现并化解了潜在资源浪费⁹。

某全球金融数据与媒体公司的技术主管在接受我们的访谈时，阐述了尽早构建可见性所能带来的收益：将成本分摊机制前置到架构设计阶段，而非事后补建，企业便可将成本与价值归因到组织、团队，大多数场景下甚至能落实到个人。

■ 要务二： 谁使用，谁承担成本

将词元成本分摊至实际使用团队，可先从技术开发部门以及消耗量最高的三至四个职能部门开始实施。成本展示虽能提升团队的成本意识，但由于不直接绑定财务责任，仍需配套相应激励措施才能真正推动行为改变。调研表明，成本分摊是驱动成本优化和强化价值管理最有力的手段之一，然而目前只有不到十二分之一的企业采用了这一机制。

这个机制十分直观：当团队需要为自己的词元消耗买单时，团队主管便会主动审视支出是否合理，技术人员在选择模型时也会把成本纳入考量。实施了成本分摊的企业通常在治理能力和监测机制方面也更为成熟。成本分摊机制并非孤立发挥作用，而是需要以工作负载级可见性和明确的责任机制为基础。对于AI成熟度较低的企业，仅靠成本分摊往往难以取得预期成效。某全球大型科技基础设施公司的首席技术顾问分享了一种“标准集中制定，责任分散落实”的运营模式。在这一模式下，每个正在运行的工作负载都应指定一个业务负责人，负责解释预期成果、成效衡量指标，以及规模化应用后能够带来的经济效益。

■ 要务三： 为每项任务匹配最合适的模型

企业需要识别出真正需要前沿模型能力的工作负载与业务环节，这类场景通常仅占约10%，主要涉及复杂推理、规划、多源信息整合，以及高风险内容生成等任务；而其余大部分场景，中等或开放权重模型足以胜任。更重要的是，这套策略应作为网关规则强制执行，而非仅作为参考指南由开发人员自主选择。

模型间的价格差异显著，前沿模型的使用成本可到达低层级模型的15倍以上，而后者往往也能胜任同等任务。由加州大学伯克利分校主导开发、发布于ICLR 2025的开源框架RouteLLM，在标准测试任务中实现了高达85%的成本节降，同时保留了旗舰模型95%的性能¹⁰。

某全球大型科技公司的企业AI战略副总裁直言不讳地指出：大多数企业AI应用仍局限于信息搜索和内容检索，只有一小部分高阶用户需要高级模型来完成复杂任务。如果将这两种需求交由同一种前沿模型处理，原本可以避免的成本便会不断累积。

某全球快餐连锁企业的CTIO从运营角度描述了类似问题：用户最初即使是执行一些简单任务，也会默认选择成本最高的前沿模型，因为他们尚不清楚哪些场景低成本模型足以胜任。而一个具备动态模型调用机制的中央网关，能够有效防止此类默认倾向引发支出失控。

埃森哲的一个内部平台每周处理约8.7万亿个词元，通过自主部署开放权重模型进行推理并结合智能模型调用，我们将其成本控制在全部使用前沿API的约六分之一¹¹。行业发展也正呈现这一趋势。IDC《2026年AI与自动化未来展望》预测，到2028年，70%的AI领先企业将采用先进的多工具架构，在不同模型之间动态自主调用¹²。Vercel公司于2026年公布的生产环境数据显示，这一模式已经开始在企业中大规模落地，生产团队平均运行35个模型，并将其组建成一套模型调用决策体系。曾经的前沿实践正迅速演变为企业运营基础能力¹³。

如何按需调用模型

如前所述，前沿模型不应成为默认选项。埃森哲对O*NET数据库内9368项工作任务进行了分类研究，通过五项测试检验这些任务是否真正需要前沿AI模型。在剔除体力劳动以及需要人工审核和确认的任务后，仅有10%的任务满足前沿模型使用条件。其余90%的任务，要么使用中端模型即可顺利完成，要么几乎不需要复杂推理，模型选择对最终输出结果的影响微乎其微。

这90%涵盖网络设计评估、根据既定模板起草合同、安全事件排查，以及参照已知标准分析数据等，并非简单的任务。

完成这些任务同样需要专业知识的支撑，问题在于它们并非不可简化，而这正是前沿模型与其他模型的区别所在。

当一项任务没有既定解决方案，当某个前期错误在悄无声息地破坏后续流程，当任务一旦拆分便失去意义，当需要融合两个不同专业领域的知识，或是当价值高度依赖整体连贯性时，这项任务就是不可简化的。如果以上五种情形均不成立，无论工作看起来多么艰巨，一个称职的中端模型就足够了。

模型调用带来的成本差异显而易见，前沿模型的单位词元成本通常是同等能力中端模型的10~20倍。目前，有54%的AI请求被分配给了能力超出任务实际需求的模型。近四分之一的高管认为，模型调用是提升效率的最有效手段，他们的观点不无道理。实践表明，当请求被匹配给能够以最低成本实现预期结果的模型时，可节省40%~60%的成本。

■ 要务四： 在工作负载上线前，明确价值论证与管理标准

在投入资金开发新的AI应用前，企业应首先厘清三件事：现有流程的财务成本和时间成本分别是多少；AI将改善哪些具体财务成果；以及如何以财务价值衡量这些成果。然后，设置明确的准入门槛，确保每一个AI应用在上线前都具备以下基础条件：已部署监测工具从工作负载层面衡量资源消耗；优化方案已朝着收支平衡稳步推进；预期价值可以转化为可量化的财务成果（无论是营收增长、成本节降，还是资本效率提升）。

无需设计得过于复杂，关键在于机制前置，而不是等到项目上线后再设法论证其合理性。

Match Group的实践清晰展示了企业如何从“鼓励试验”转向“价值把关”。该公司过去给予AI试点项目近乎无上限的预算支持,但据CFO史蒂夫·贝利 (Steve Bailey) 介绍,自2026年起,公司大幅提高了投资门槛。任何达到一定规模的AI投入,都必须提交明确的商业论证,证明项目能够通过降本增效、提升营收或优化用户体验等方式创造价值。试点阶段可以保留一定灵活性,规模化推广则需要经得起财务评估的价值回报¹⁴。

荷兰银行 (ABN AMRO) 已将价值论证变成了一项制度规定。该行要求所有AI用例在部署前至少对应五项预设价值成果中的一项。截至2026年5月,该行已有40个用例落地上线,每个场景都预先映射到一项价值成果,且均有明确的数据支撑。例如,其ICS语音机器人承接每月约6万通电话中的80%,每周可节省约500小时人力;其Delphi风险评估工具为金融犯罪风控业务每月节省约800小时。这些指标在应用部署前即已明确设定,并在上线后被持续监测¹⁵。

Klarna将AI应用于客户支持业务时也采取了类似的做法。公司在架构设计阶段就内置了成效衡量机制,用以追踪问题处理时长、重复联系率和自助解决率等关键指标。公开信息显示,该项目2024年上线之初,就节省了约4000万美元的招聘成本,到2025年第三季度,这一数字已提升至约6000万美元¹⁶。之所以能够如此清晰地衡量、验证应用成果,是因为可见性能力在项目启动阶段便已被纳入整体设计。

提前设立相关标准的企业,正逐渐与尚未采取行动的同行拉开差距。在优步,智能体目前提交的代码约占10%,但其COO最近坦言,公司仍难以阐明词元投入如何转化为了产品价值¹⁷。支出已经发生,但相应的价值评估体系尚未搭建完备。为弥补这一缺口,优步正逐步完善管控机制,包括为每位员工设置每月1500美元的使用限额、设立内部监控看板,以及对超额支出进行额外审批。优步的经历印证了这一规律:领先企业往往会在开放AI使用权限的同时配套搭建价值衡量与监测体系,而不是等到发现预算超支后再被动补救。

某全球大型科技基础设施公司的首席技术顾问谈到,那些贸然推动AI规模化应用的企业几乎都会面临相同的问题:平台团队不得不事后编制各种报告,解释成本为何超出预期,并试图证明项目究竟创造了多少价值。如果一个项目无法在立项阶段以财务语言论证其预期价值,就不应被视为正式项目,而应作为试点进行管理和投入:设定明确的部署范围和时间节点,要么在规定时间内论证价值,要么就此停止。

本次调研发现,多数无法验证财务回报的企业,并非未能创造价值,而是未能衡量价值。要务四要求企业在支出发生前搭建相应的价值衡量机制,而不是等到事后再设法论证其合理性。

价值的定义取决于具体职能和行业。在客户运营领域,价值体现在自助解决率以及单次闭环交互成本上;在软件开发领域,价值体现在开发周期和每位开发者的产出上;在后台运营领域,价值体现在每笔交易带来的财务价值上;在风险合规领域,价值体现在分析师的决策量和误报率上。价值的定义因场景而异,但管理准则始终如一:在AI应用上线前,论证预期成果、设定基准并构建衡量机制。

■ 要务五: 提升全员AI能力

大多数情况下,模型选择不当并非使用者判断失误,而是由于缺乏明确指导。开发人员和员工倾向直接选择能力最强的模型,一方面是因为他们看不到成本差异,另一方面通常也缺少简明的标准来指导他们如何按需选择模型。培训有所帮助,但作用有限:65%的企业已开展各类用户培训,但模型过度调用问题依然存在。这说明仅靠提高认知并不足以改变使用行为,关键是在做出选择那一刻,让规则得到强制执行。

某全球大型科技公司的企业AI战略副总裁直言不讳地指出：

“你的丰田凯美瑞完全够用，暂时没必要升级到法拉利。”

这位高管还进一步说明了这套指导规则应如何落地：将词元投入产出率纳入开发者KPI；按用户角色和用例分配预算；监测AI用量的同时兼顾使用效率。推行上述举措，需要确保员工充分理解企业AI应用规范并为他们提供相应资源和支持。

设定目标和规则，并非是要让员工去承担他们难以从全局视角感知的成本。AI的总成本并不总是一目了然，例如，较低成本模型可能需要更长的处理时间、更多轮的提示词交互，或者输出结果需要投入更多人工审核。因此，企业的核心目标是做好行为引导，搭建一套简单明确的框架，帮助员工按需匹配模型，并开放成本查询权限给相关人员。在此基础上，通过网关或终端预设默认模型，让最适配的模型成为最便捷的选项，把人工判断留给真正需要专业权衡的场景。

领先企业会明确使用预期、提供必要的信息支持，并依靠治理机制承担约束职责。当规则清晰易懂，员工将自觉遵守执行。CIO的职责就是让这一切简单易行。

这不是单靠培训就能解决的问题。词元经济学涉及技术、财务、人力和业务等多个领域。CIO需要与CFO协同，将价值量化和成本问责纳入AI投资决策；与CHRO协同，将模型选择和成本意识融入员工日常工作规范；同时推动业务负责人在立项之初就为每个AI工作负载明确权责、完成价值论证，并对最终业务成果负责。

这五项要务共同塑造了企业的技术运营模式，确保其有效运作是全体管理层的责任。CIO的职责在于统筹协调。

结语

管理机制决定成败

AI时代的领导力，归根结底取决于管控能力和治理水平。真正领先的企业，会在每一个词元转化为需要解释的成本前就实施有效管理，并将其与可验证的价值成果建立明确关联。

许多企业先开放了AI权限，治理往往要等到账单生成后才被提上议程。这种先部署、后治理的路径并非无法纠正，但在预算压力下补建治理体系，无论是难度还是成本，都远高于在部署之初便同步建立相应机制。不过，选择正确路径的企业也并非个例：有技术高管在项目启动之初就建立了成本和价值追踪机制；有技术负责人在AI工具规模化推广之前便追问：“这个词元实际做了什么？”荷兰银行要求每个AI应用在投入生产环境前都必须对应到一个明确的业务成果和衡量标准。

调研结果同样印证了这一规律——处于早期试点阶段的企业，平均优化了约15%的词元支出，并将约30%的请求匹配到了适当模型，而实现AI规模化部署的企业，这两个比例分别提升至约28%和58%。这些成果源于可见性、模型调用、运营模式，以及严谨的治理体系——像制造商运营生产线一样管理AI基础设施：控制单位成本、减少浪费并持续衡量。

率先建立起这套管理体系的企业将在未来数年持续扩大领先优势。AI能力将不断提升，模型价格会不断变化，消耗规模也将持续上涨。但这一切不会改变根本逻辑：价值源于词元创造的成果，而非购买词元本身。

本报告试图弥合的正是AI投入与价值产出之间的鸿沟。留给企业行动的窗口期就在下一代模型出现、下一次智能体扩展，或不可预测的AI账单成为实质性管理问题之前。一旦那一刻到来，问题将沿着管理链条迅速传递：从董事会到CEO，再从CEO到CFO、CIO。届时，已提前行动的CIO将能给出清晰有力的答案。词元经济学正逐渐成为企业运营AI的标准管理实践。及早采用这一管理方法，CIO不仅能够帮助CFO清晰评估AI投资的经济效益，让CEO看到可量化的价值回报，更能推动企业基于实证而非期望，持续规模化释放AI价值。

埃森哲助您布局 词元经济

企业的技术基础不再只是支撑业务运转的后台能力,而是创造、积累并持续放大竞争优势的重要引擎。构建、投资并持续完善这一能力体系,正是CIO的核心使命。

要履行这一职责,技术领导需要统筹三大关键要素:坚实的技术基础、有效的运营模式,以及清晰的投资逻辑。与此同时,还需与业务部门紧密协作,制定以业务价值最大化为导向的发展路线图。目前,能够真正将这些要素整合起来的企业仍属少数,这也正是大部分AI潜力至今未能转化为切实业务成果的原因。

埃森哲着眼于业务重塑,而不仅仅是技术架构升级。我们帮助客户围绕现代企业架构的三大核心维度,构建面向未来的数字核心:引擎 (Engine) 支撑企业运行;知识 (Knowledge) 建立对自身的可信认知;大脑 (Brain) 通过AI模型、协同机制和智能体解读认知并将其转化为行动和价值。安全与治理能力则为整个体系提供运营保障。

埃森哲内部平台每周处理约8.7万亿个词元,通过自主部署开放权重模型进行推理并结合智能模型调用,我们将其成本控制在全部使用前沿API的约六分之一。基于自身实践以及全球客户经验,埃森哲形成了一套成熟的词元经济治理方法论,总结为三个步骤:看见问题、解决问题、持续优化。这一方法由埃森哲AI词元导航器 (Accenture AI Token Navigator) 支持,平台内置七项专业工具,覆盖AI资产全景诊断、实时模型性能评测、预算即代码 (budgets-as-code) 与策略即代码 (policy-as-code) 运行治理、业务价值成本实时监测,以及配套的工程体系与人才培养机制,从而将成本治理能力嵌入企业日常运营,实现长期、可持续的价值提升。

关于本研究

我们的调研综合了三类数据来源：面向17个国家及地区750位高管开展的调研、与《财富》500强企业技术和财务主管的15次深度访谈，以及基于6000个高管模拟样本的推演，以对调研发现的关键模式进行压力测试和验证。所有受访者均来自年营收超10亿美元的企业并直接参与AI投资决策。

主要调研：于2026年7月开展，调研对象为来自17个国家及地区、年营收超10亿美元企业的750位高管。模拟分析：基于Aaru平台构建，包含6000个虚拟决策主体，覆盖六个企业成熟度层级，相关数据均明确标注为模拟数据。定性访谈：于2026年6月至7月间开展，对《财富》500强企业的技术和财务主管进行了15次深度访谈，访谈内容均已匿名化处理并加以汇总。统计方法：采用皮尔逊相关分析、多元回归分析和方差分析。除非另有说明，所有回归和相关性分析结果均达到统计学意义 ($p < 0.05$)。

鸣谢

研究团队：

阿马尔·塞巴斯蒂安 (Amal Sebastian)、阿迪蒂·阿比吉特 (Aditi Abhijeet)、比斯瓦迪普·戈什·哈兹拉 (Biswadeep Ghosh Hazra)、阿洛克·兰詹 (Alok Ranjan)、阿贾伊·加格 (Ajay Garg)、克里希纳·坎特·苏特拉耶 (Krishna Kanth Sutraye)、格里·法尔科娃 (Gerry Farkova)、娜塔莉娅·西先科 (Nataliya Sysenko)、伊格纳西奥·马莫内 (Ignacio Mamone)、凯文·加拉格尔 (Kevin Gallagher)、加尔吉·查克拉巴蒂 (Gargi Chakrabarty)、普拉迪普·罗伊 (Pradeep Roy)

业务专家与领导：

托尼·莱拉里斯 (Tony Leraris)、史蒂夫·考特尼 (Steve Courtney)、马特·拉戈津斯基 (Matt Lagodzinski)、迈克·艾森斯坦 (Mike Eisenstein)

市场营销部：

布里安娜·陈 (Brianna Chan)、阿特·根卡雷利 (Art Gencarelli)、劳伦斯·麦金 (Laurence Mackin)、埃德·马尼 (Ed Maney)、珍妮·斯坦库斯 (Janine Stankus)

大中华区业务联系人

曹琦峰

埃森哲大中华区AI与数据事业部总裁

qifeng.cao@accenture.cn



关键概念与指标说明

可归因财务价值 (复合归因)——企业能够以量化财务成果 (例如, 营收增长、成本节降, 或资本效率提升) 来阐释的词元总支出。这是一个复合指标, 因为财务归因需要经过两个阶段, 这两个阶段分别对应调研中的两个问题:

- **价值关联:** 能够关联到可衡量业务成果的词元用量占比。
- **财务表达:** 已关联的词元用量中, 能够进一步量化为财务价值的占比。

由于第二阶段以第一阶段为前提, 因此复合指标的计算方式为:

复合归因 = 价值关联 × 财务表达

例如, 一个企业将其60%的词元消耗与某个价值成果相关联, 又将其中60%进行了财务表达, 则可实现 $0.60 \times 0.60 = 36\%$ 的复合归因。

优化率——企业通过各类效率优化举措 (模型调用、缓存优化、提示词机制、上下文管理等) 所减少的词元消耗, 以相对于未优化状态减少的消耗占比表示。

盈亏平衡优化率——在24个月内, 在用量增长和单价下跌的情况下, 保持词元总支出不变所需达到的优化水平。该指标并非调研数据, 而是推导得出:

盈亏平衡 = $1 - 1 / [(1 + \text{用量增长}) \times (1 + \text{价格变化})]$

调研预测24个月内用量增长 (+78%)、价格变化 (-19%), 代入计算后可得: 盈亏平衡优化率约为31%。

成本展示与成本分摊——成本展示会向使用团队披露其词元消耗情况, 但不产生实际财务责任。成本分摊则按照实际用量向使用团队计费, 从而直接从财务侧激励团队主动管理词元消耗。本报告将无成本分摊 > 非正式成本报告 > 成本展示 > 成本分摊的演进过程, 视为财务问责能力逐步提升的不同阶段。

智能调用——被分配至能够满足任务需求的最低成本模型的AI请求占比。如果请求被分配给比任务所需更高成本的模型, 则称为“错误调用”。

应用级监测能力——将词元消耗追溯到具体用户、团队、应用或工作负载的能力, 是模型调用和财务归因赖以实现的监测基础。

数据解读

以下是对报告核心数据的简要说明。

术语	简单来说
复合归因	能够向董事会证明已转化为财务成果的词元投入占比。目前每1美元支出，可转化约20美分。
优化率	通过优化举措（调用、缓存、提示词）减少的词元消耗。调研所示平均优化率约为23%。
盈亏平衡优化率	随着用量增长和价格下降，保持词元总支出不变所需达到的优化率。24个月内需达到约31%。
成本展示与成本分摊	成本展示让团队看到成本，成本分摊让团队支付成本。只有成本分摊才会改变团队行为。
智能调用	被分配至能够以最低成本模型实现预期成果的请求占比。比例越高越好。



参考资料

- [1] 词元经济学基金会, [什么是AI词元经济学?](#), 2026年9月1日。
- [2] 埃森哲分析。若未采取优化举措, 词元成本预计将在24个月内增长约44%, 主要原因是用量将增长78%, 尽管价格下跌19%能抵消部分成本。实现31%的优化率则可使支出维持在当前水平。少数已全面实施成本分摊的企业(样本量~53) 称其优化率达到了近48%; 鉴于样本量较小, 该数据仅做参考, 不视作基准水平。
- [3] 基于741位已披露词元支出的受访者数据, 按其年均词元支出的加权平均值337万美元测算得出。24个月预测沿用了同一套基准数据, 并采用了调研得出的两个假设: 用量增长(+78%)、价格下降(-19%)。
- [4] 娜塔莉·隆 (Natalie Lung), [优步限制Claude Code等AI工具的使用来管理成本](#), 彭博社, 2026年6月2日。
- [5] 本·希姆库斯 (Ben Shimkus), [Harvey CEO称公司5月份的AI词元用量达到12万亿美元](#), Business Insider, 2026年6月19日。
- [6] 亨利·尚多内 (Henry Chandonnet), [AI词元成本迫使企业重新思考招聘、预算制定和使用管理机制](#), Business Insider, 2026年6月17日。
- [7] 克里斯·舒普特林 (Chris Shuptrine), [2026年追踪供应商AI词元用量的七大工具](#), Torii, 2026年6月。
- [8] [“我们创造了一个怪物”: 由于成本挤压预算, 各企业纷纷限制AI使用](#), 《金融时报》, 2026年6月19日。
- [9] 麦肯齐·霍兰 (Makenzie Holland), [Shutterstock和Prudential Financial分享AI成本策略](#), CIO Dive, 2026年6月11日。
- [10] [AI成本核算: 2026年模型支出的合理优化](#), Digital Applied, 2026年6月23日。
- [11] [AI已计入企业损益表, 而多数企业只读懂了一半](#), 埃森哲, 2026年7月29日。
- [12] [全球AI与自动化: 2026年预测, 文档编号US53858125](#), IDC FutureScape, 2025年10月。
- [13] [AI成本核算: 2026年模型支出的合理优化](#), Digital Applied, 2026年6月23日。
- [14] [Match Group CFO为2026年AI支出设定了“更高标准”](#), CFO Dive, 2025年12月23日。
- [15] 荷兰银行创新与技术部, 投资者关系, [荷兰银行与AI: 投资者会议](#), 2026年5月20日。
- [16] [Klarna AI客户支持效率](#), Twig, 2026年6月10日。
- [17] 杰克·安杰洛 (Jake Angelo), [优步在四个月内就花光了2026年全部AI预算。其COO开始质疑这笔投资是否物有所值](#), 《财富》, 2026年5月26日。

关于埃森哲

埃森哲注册于爱尔兰，帮助全球各行业的领先企业构建数字核心，运用AI，实现高效、全面的价值提升。我们致力于成为客户信赖的重塑转型卓越伙伴，引领广泛、安全的AI应用，坚持以客户为本、以AI赋能，建立全球最佳职场。公司汇集全球81.4万名专业人士，依托自有资产与平台、深厚的流程与行业专长，以及领先的生态合作关系，为客户提供端到端解决方案，并在规模化运营中实现可衡量的业务成果。

我们通过重塑服务，提供网络安全、数字核心、财务、行业与企业、Song、供应链与工程、人才等领域广泛的专业服务，具备AI与数据、行业与流程，以及技术服务方面的前沿能力。我们为全球约九千家客户提供服务，2026财年全球营收约为740亿美元。

埃森哲在中国市场开展业务近40年，运营和办公地点分布在北京、上海、大连、成都、广州、深圳、杭州、香港和台北等城市。

了解更多，敬请访问埃森哲中文主页[accenture.cn](https://www.accenture.cn)。

关于埃森哲商业研究院

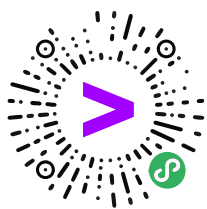
埃森哲商业研究院针对全球企业组织面临的重大问题，洞悉发展趋势，提出前沿思想。我们将数据科学驱动的分析等创新研究方法，与对行业和技术深刻理解相结合；研究团队由分布于全球20个国家的300名研究员组成，每年发布数以百计的报告、文章和观点。我们在研究工作中运用生成式人工智能，并遵循埃森哲负责任人工智能标准，由研究专家在条件允许的情况下，以传统研究方法对生成式人工智能的产出进行审阅与验证。我们与世界领先组织共同开展的研究富有启发性，助力客户拥抱变革、创造价值，充分释放技术与人类智慧的力量。

敬请访问埃森哲商业研究院主页www.accenture.com/research。

欢迎与我们互动



埃森哲中国官方微信



埃森哲中国官方小程序

免责声明：本报告中的材料反映了编制时的可用信息，但全球形势正在快速演进，情况很可能会发生变化。报告内容仅作为通用参考信息，并未考虑阅读者的各种具体情况，也不用于替代埃森哲专业顾问的咨询意见。在可适用法律允许的最大范围内，对于本文中信息的准确性和完整性，以及任何基于这些信息所采取的行动或发生的疏漏，埃森哲均不承担责任。埃森哲未在文中提供任何法律、法规、审计或税务建议。读者有责任从自己的法律顾问或其他有资质的专业人士处获得此类建议。文中引用了归第三方所有的商标。所有这些第三方商标分属其各自的所有权人。相关内容没有任何明示、暗示或表示得到了该商标持有人的赞助、认可或批准。