

FUNDA REPORT

Deep|LLM: 企业 AI 应用调研 (Vol.5) - 降本与应用扩张并行, 业务回报影响后续投入

Enterprise AI budgets are still rising as APIs and app rollouts offset seat saturation and model cost cuts.

摘要

多数受访企业仍计划增加AI支出, 但增速随工具覆盖和应用推广阶段而分化。模型切换、缓存与采购议价降低部分成本, 应用扩张仍带动新增投入: 某二手车交易平台持续优化成本, 同时预计未来6个月支出增长约30%。部分企业在订阅席位趋稳后转向增加API与现有用户用量, 后续预算更取决于应用落地及降本、增收效果; 前沿模型采用仍集中于探索、规划和复杂任务。

目录

1. 本期要点:
2. 一、AI 支出增长与预算预期
3. 二、订阅与 API 支出
4. 三、ROI 与投资决策
5. 四、预算与使用管理
6. 五、前沿模型
7. 六、开放权重与闭源模型
8. 七、应用场景

本周(样本为9月15日-20日)我们继续访谈了的10家企业, 覆盖支付、电信、薪酬与人力资源服务、电商、二手车交易、广告媒介、数字营销、制药、多元化控股及工业软件等行业。

多数样本企业仍计划增加AI投入, 但增长动力和增速出现分化。员工订阅席位基本配齐后, 新增支出更多来自API项目和使用量; 仍在扩大工具覆盖、推广智能体的企业, 则继续增加订阅与应用投入。与此同时, 模型选型、缓存和采购谈判正在降低部分费用, 应用回报也影响后续预算。

本期要点:

1. **AI支出继续增长, 部分企业订阅席位采购趋稳后增速放缓, 新增应用仍带动投入。**某多元化控股企业的订阅人数和人均费用预计稳定, 后续增长主要来自API; 某薪酬与人力资源服务企业预计至年底再增约25%, 增速放缓。某二手车交易平台过去6个月增长约40%-60%, 到年底预计再增约30%。某制药企业和某电信运营商则预计增长加快, 分别由业务智能体推广和扩大员工工具覆盖带动。
2. **员工订阅席位配齐后, 部分企业新增支出转向API项目; 扩大员工覆盖仍需增加订阅投入。**某支付技术服务企业预计未来约6个月新增约10万美元月度支出, 全部用于API项目; 某多元化控股企业的新增支出也主要来自API。某电信运营商仍需为更多员工采购工具, 预计订阅贡献约八成增量。API占比提高还可能来自采购方式变化: 某薪酬与人力资源服务企业计划把部分订阅席位与积分采购转为API调用, 迁移后月度支出预计大致稳定。

3. **成本优化正在降低模型费用，但应用推广和用量增长仍带动多数企业增加投入。**某数字营销技术与服务企业换用更便宜的模型后，支出仍随新客户和流量每月增长约2%-6%。某二手车交易平台通过缓存、去重等措施降低单个模型的成本，同时继续向更多地区推广应用。某广告媒介代理企业收紧 API 用例审批后，API费用下降10%-20%，订阅费用增长约10%，总额大致持平。
4. **已显现的 AI 回报体现在业务处理量增加和外购软件替代，部分企业据此追加预算。**某电信运营商用自建应用替代营销自动化平台，原平台年许可费约 20 万-25万美元。某二手车交易平台约35%-40%的 AI 项目回报已转正；某工业软件企业通过展示项目收益争取到专项预算。员工通用工具的收益仍较难追踪，新增 API 项目则更多需要证明降本或增收效果。
5. **数据控制、定制需求和任务成本推动部分企业扩大开放权重模型的使用。**某工业软件企业计划将开放权重模型承担的工作负载由约40%提高至明年的约60%，优先用于私有数据和内部工具。某二手车交易平台让内部模型与开放权重小模型处理大部分常规查询，再将复杂问题交给闭源大模型。某薪酬与人力资源服务企业则计划通过自建网关接入更小的模型，以降低现有工具的使用成本。
6. **Fable / Astra主要用于探索、规划和复杂研发，常规生产任务继续使用现有主力模型。**某数字营销技术与服务企业在订阅工具中使用前沿模型探索和规划，生产 API 仍以中端模型为主；某工业软件企业将 Fable 用于代码库分析和复杂推理。支付、电商运营及多元化控股企业当前均未使用Fable / Astra。数据质量、系统整合和应用开发仍影响新用例落地。

一、 AI 支出增长与预算预期

多数样本企业仍将增加 AI 投入，增长动力随应用推广阶段而变化。支付企业完成全员订阅席位采购后，新增投入转向 API 项目；多元化控股、薪酬服务和二手车企业预计增速放缓，但支出仍将增加。制药企业继续推广智能体，电信运营商扩大员工工具覆盖，两家企业预计增长加快。广告媒介企业过去半年收紧 API 审批后，API费用下降、订阅费用上升，总额大致持平；未来半年两项预计均保持稳定。

样本企业	历史支出变化	未来约 6 个月支出预期	2027年支出预期
某数字营销技术与服务企业	过去 6 个月约增15%-30%，主要来自客户和流量增长；当前约 4 万美元/月	至年底约增20%；未来 6 个月约增25%（FUNDA估计），增长速度大致延续	约增30%-45%
某广告媒介代理企业	整体大致持平：API费用下降约10%-20%，订阅费用增加约10%；近期集团 AI 支出约 900 万美元/月	订阅与 API 费用均预计持平；重大新模型出现时可能增加投入	预计低速增长，与 IT 预算增速相近；FUNDA估计约增10%
某多元化控股企业	过去 6 个月约增30%；营销图像与内容工具从低基数增长约500%；近期约 80 万美元/月	约增3%（FUNDA估计），较此前放缓；API增长，订阅及其他费用大致稳定	约增3%（FUNDA估计）
某工业软件企业	广义 AI 支出由约 17 万增至 20 万美元/月，增长约18%，主要来自 API 与内部智能体应用	约增20%（FUNDA估计），增速大致延续；智能体工作流程继续向更多部门推广	约增40%（FUNDA估计）
某制药企业	历史自然增长约20%/年，折合半年约9.5%（FUNDA测算）；当前约3,000万-4,000万美元/月，为专家粗估	约增14%（FUNDA估计），增速加快；现有智能体扩大部署，新用例增加投入	约增30%（FUNDA估计）；新增投入取决于用例推广及业务部门审批

某支付技术服务企业	由零增至约 20 万美元/月：今年 3 月开始开通员工工具，6 月底已覆盖全员	约增50%，月度支出由约 20 万增至 30 万美元；新增约 10 万美元全部来自 API 项目，预计 2027 年一季度开始投入	2027年情景下，月度支出较当前约增100%（FUNDA估计）；取决于智能体支付等项目的推进与采用
某电信运营商	过去 6 个月约增100%；由约5,000增至 1 万新元/月，当前约合8,000美元/月	约增200%，月度目标由 1 万增至 3 万新元，预计约 6 个月内逐步达到；增速加快	2027财年月度目标为 3 万新元，约为当前的 3 倍
某薪酬与人力资源服务企业（欧洲业务）	今年 1 月至今约增100%；当前约 3.5 万-5万美元/月	至年底约增25%，增速放缓；订阅席位基本稳定，增量来自使用量	预计大致持平
某二手车交易平台（印度业务）	过去 6 个月约增40%-60%；当前约 20 万-30万美元/月，不含云与平台费用	约增30%，较此前放缓；AI应用由12-14个城市推广至约 20 个城市	约增40%（FUNDA估计）
某电商运营服务企业	今年 5 月开始推广AI，支出已上升，具体增幅未量化；当前内部支出约 2 万-3万美元/月	年底约 3 万-4万美元/月，按区间中值测算增约40%；增量主要来自API	尚无具体增幅；后续投入和工具选择将视回报调整

来源：FUNDA企业调研；FUNDA测算。

注：历史变化和近期预测按各公司所述期间列示；2027年数字含 FUNDA 估计、预算方向及月度支出目标，具体见表内说明。工业软件企业的历史序列为包含人员、基础设施等费用的广义 AI 支出；其外购模型与工具费用另见第二节。电信运营商的增长率按新元金额计算。制药企业的总额为专家粗略估计。

1. 员工工具集中开通和成熟应用推广，带动了本轮样本的支出增长。

过去一段时间，10家样本中 9 家的 AI 支出上升，广告媒介企业大致持平。支付企业向全员集中开通 AI 助手，薪酬服务企业和电信运营商的支出约翻倍；二手车企业则在成熟应用向更多城市推广的过程中增长约 40%-60%。随着员工订阅席位陆续配齐，后续增量将更多取决于现有工具的使用深度，以及业务应用能否持续扩大。

a. 员工工具推广和使用扩大，带动部分企业支出增加。

- 某支付技术服务企业今年 3 月开始向员工开通 AI 助手，至 6 月底已覆盖约2,000名员工。月度支出由年初的零增至约 20 万美元，目前全部为订阅费用。
- 某薪酬与人力资源服务企业（欧洲业务）今年 1 月至今的 AI 支出约翻倍，当前约 3.5 万-5万美元/月，以员工工具订阅和额外积分为主，也包含少量客户产品的 API 费用。
- 某电信运营商过去 6 个月的月度 AI 支出由约5,000新元增至 1 万新元。增长来自内部 AI 应用替代外购软件，以及编程工具使用增加；当前订阅与 API 费用各占约一半。
- 某电商运营服务企业今年 5 月开始推广AI，当前内部支出约 2 万-3万美元/月，其中 API 及按量使用费用约占85%-90%。

b. 成熟应用扩大部署范围，仍在带动费用增加。

- 某二手车交易平台（印度业务）过去 6 个月 AI 支出增长约40%-60%。公司计划将相关应用由12-14 个城市推广至约 20 个城市，后续支出预计再增约30%，增幅低于此前，但订阅和 API 使用仍将同步扩大。

c. 收紧 API 使用审批后，广告媒介企业的总支出大致持平。

- i. 某广告媒介代理企业今年加强 API 用例审批，并清理闲置用户和密钥，原定推出的高阶平台计划也被取消。过去 6 个月 API 费用下降约10%-20%，订阅费用随新增用户增长约10%，总支出大致持平。

d. 已有 AI 应用的企业，费用随业务量和使用范围扩大而增加。

- i. 某多元化控股企业过去 6 个月支出约增30%，部门之间差异较大。营销部门的图像与内容工具从较低基数增长约500%，其他领域增幅较小。
- ii. 某工业软件企业的广义 AI 支出由约 17 万增至 20 万美元/月，约增18%。增长主要来自 API、内部自主智能体系统，以及研发、财务、运营、法务和人力资源等部门的应用集成。
- iii. 某数字营销技术与服务企业过去 6 个月支出约增15%-30%，主要随新客户和业务流量增加。公司此前已通过模型评估和切换降低费用，目前继续随业务量增长。

2. 企业后续投入节奏分化，部分企业增幅回落，制药与电信样本预计加快。

多元化控股、薪酬服务和二手车企业预计后续增速放缓。制药企业通过已验证的智能体扩大使用，电信运营商继续增加员工工具订阅席位，两家企业预计增长加快。支付企业已完成全员订阅席位采购，下一阶段的新增投入将转向 API 项目；电商运营企业则计划扩大工作流应用，继续增加调用支出。

a. 部分企业支出仍在增加，但增幅低于此前。

- i. 某二手车交易平台（印度业务）后续支出预计再增约30%，低于过去 6 个月的约40%-60%。增长仍来自既有 AI 应用向更多城市推广。
- ii. 某多元化控股企业的员工人数和人均订阅支出预计基本稳定，后续增量主要来自 API。API 目前约占支出的10%，即使 API 费用增长约30%，对总支出的拉动也约为 3 个百分点，低于此前约30%的总体增幅（FUNDA估计）。

b. 制药企业和电信运营商仍在扩大应用，支出增速预计加快。

- i. 某制药企业已将采购、事件登记和销售支持等智能体从试点推向规模化应用。内部 AI 助手已覆盖约90%的员工，后续增长主要来自每名员工执行更多任务，以及新获批用例增加投入。渠道预计支出增长加快；FUNDA估计未来 6 个月约增14%，高于历史自然增长折合的半年约9.5%。

'There is an acceleration, but this acceleration happens because of the value you are bringing to the table.'

- 1. 某电信运营商计划将月度 AI 支出由约 1 万新元增至 3 万新元，预计约 6 个月内逐步达到，增幅约200%，高于过去 6 个月的约100%。新增投入包括面向财务、市场和销售的员工工具，以及更多内部生产应用。

2. 订阅席位基本配齐后，薪酬服务企业的后续增速放缓。

- a. 某薪酬与人力资源服务企业（欧洲业务）今年 1 月至今支出约翻倍，需要员工工具的主要使用人群已基本覆盖。至年底支出预计再增约25%，新增费用主要来自现有员工的积分与 Token 用量，而非继续增加订阅席位。

3. 员工工具普及后，新增支出更多取决于使用深度和业务应用推广。

支付企业、多元化控股企业和数字营销企业已向全员提供 AI 工具，薪酬服务企业也已覆盖主要使用人群。订阅席位基本稳定，并不意味着 AI 投入停止增长：支付企业计划增加 API 项目，数字营销企业预计人均订阅支出上升，制药企业则继续扩大现有用户的智能体使用。新增订阅需求仍集中在工具覆盖率较低、使用权限升级或应用推广尚未完成的企业。

a. 二手车企业仍需随地域推广增加订阅和 API 使用。

- i. 某二手车交易平台（印度业务）约75%-80%的员工已有企业版 AI 工具权限。公司继续向更多城市推广 AI 应用，订阅仍占支出约70%-75%，API约占25%-30%，近期比例预计大致稳定。

b. 工业软件企业和电信运营商仍在扩大 AI 使用人群。

- i. 某工业软件企业预计明年使用 AI 的员工约翻倍。法务、财务和人力资源等内部工具陆续上线，将带动更多员工使用AI，并增加相关 API 调用。
- ii. 某电信运营商目前不足10%的员工拥有付费 AI 订阅席位，目标是提高至30%-40%。新增采购主要面向财务、市场和销售等知识工作岗位。公司预计明年订阅占 AI 支出的比例由约50%升至70%。

c. 工具覆盖率较高的企业，API和智能体应用继续提供增长空间。

- i. 某多元化控股企业已为约1,000名员工购买 AI 工具权限，暂无扩编计划，人均订阅支出也预计稳定。更多现有工具和面向客户的产品接入外部模型，后续支出增量预计全部来自API（FUNDA估计）。
- ii. 某制药企业的内部 AI 助手已有约 7.2 万月度用户，约占员工总数的90%。后续投入主要用于让现有用户执行更多任务，以及推广由产品线或业务线支持的新用例。

4. 2027年投入仍以增加为主，新项目进展和回报将影响扩张幅度。

展望 2027 年，8家样本仍预计增加 AI 支出，薪酬服务企业预计大致持平，电商运营企业尚未给出具体增长预期。工业软件和二手车企业仍在推广既有应用，制药企业继续扩大智能体使用；支付企业能否进一步增加 API 投入，则取决于智能体支付等新业务的推进。部分企业还需要用节省的其他开支支持 AI 预算，因此应用效果会影响后续投入节奏。

a. 地域推广和业务回报，将影响后续项目投入。

- i. 某二手车交易平台（印度业务）继续推进跨城市推广和工作流自动化，2027年支出预计增长约40%（FUNDA估计）。公司已有15-20个项目投入生产，另有 50 多个项目在推进，后续费用将随实际部署扩大。
- ii. 某电商运营服务企业预计近期内部 AI 支出继续增加，但 2027 年的具体增幅尚不明确。公司将结合回报调整使用方式，如果现有模型未能带来足够价值，也在考虑引入混合模型方案。
- iii. 某数字营销技术与服务企业预计 2027 年支出增长约30%-45%。现有客户与流量仍在增加，人均订阅支出也有上升空间，增长并非只依赖新增员工订阅席位。

二、订阅与 API 支出

部分企业在订阅席位采购趋稳后，新增支出主要来自 API 项目；电信运营商继续扩充员工订阅席位，订阅贡献大部分增量。支付和多元化控股企业的后续增量预计全部来自API，电商运营和薪酬服务企业也主要增加按量使用。电信运营商仍在推广员工工具，订阅预计贡献约80%的新增支出。薪酬服务企业计划将现有订阅和积分采购转为 API 调用，在增加模型选择的同时控制整体费用。

样本企业	API vs. 订阅的占比	AI支出未来增长与驱动因素（API vs. 订阅）
某数字营销技术与服务企业	①API约占60%-70%，主要用于生产应用； ②订阅约占30%-40%，按岗位分档； ③约700-800名员工均有 AI 工具权限	①未来 6 个月总支出预计增约25%（FUNDA估计）； ②人均订阅支出预计增20%-30%，订阅席位数量基本稳定； ③生产 API 随客户和流量增长，两项增量贡献尚未分别量化
某广告媒介代理企业	①订阅约70%，API约30%； ②Copilot基本全员可用，Gemini与 Claude 也广泛开放； ③合同包含基础订阅席位及超额用量收费，活跃使用率约60%-70%	①未来 6 个月订阅与 API 费用均预计持平； ②2027年预计随 IT 预算低速增长； ③此前 API 费用下降约10%-20%，订阅费用随新增用户增长约10%
某多元化控股企业	①近期订阅约60%、API 约10%、垂直工具约30%； ②约1,000名员工均有 AI 工具权限，活跃使用率约60%； ③API包括面向客户产品所消耗的 Token 费用	①员工人数与人均订阅支出预计稳定； ②API预计增约30%，按当前10%的支出占比，拉动总支出约 3 个百分点； ③后续增量预计全部来自API（FUNDA估计）
某工业软件企业	①外购模型与工具约 4 万-5万美元/月，其中 API 约 3 万美元、订阅约 1 万-2万美元； ②另有自托管与租赁机房费用约 5 万-6万美元/月； ③AI应用继续向更多员工开放	①内部智能体与跨部门应用带动 API 调用增加，预计明年 AI 使用人数约翻倍； ②外部许可成本通过自托管开放权重模型优化； ③后续增量中，API约占60%、订阅约占40%（FUNDA估计）
某制药企业	①订阅约40%，广义 API 相关支出约60%，为专家粗估； ②内部 AI 助手月度用户约 7.2 万人，覆盖约90%的员工； ③付费订阅席位数量未披露	①现有用户使用更多智能体和任务执行功能； ②已验证用例扩大部署，新获批准用例增加业务线投入； ③订阅与 API 对增量的贡献未分别量化
某支付技术服务企业	①当前约 20 万美元/月，全部为订阅； ②约2,000名员工均已开通 Copilot 付费订阅席位； ③订阅席位内智能体由订阅费覆盖，目前无单独 API 支出	①未来约 6 个月月度总支出预计由 20 万增至 30 万美元； ②订阅大致稳定，新增约 10 万美元全部来自API； ③届时 API 约占总额的三分之一，贡献全部增量（FUNDA测算）
某电信运营商	①当前订阅与 API 各约50%； ②订阅主要为 IT 编程工具，API用于内部应用的生产推理； ③付费订阅席位覆盖不足10%，计划提高至30%-40%	①月度目标由 1 万增至 3 万新元； ②订阅占比预计由50%升至70%，API由50%降至30%； ③据此测算，订阅和 API 分别贡献约80%和20%的增量，API绝对支出仍增约80%（FUNDA测算）
某薪酬与人力资源服务企业（欧洲业务）	①订阅和额外积分占大部分支出，API用于少量客户产品； ②约 200 个 ChatGPT 订阅席位，约 60 名开发者另用编程工具； ③订阅及积分费用约 3.6 万美元/月	①订阅席位基本稳定，至年底总支出预计增约25%，主要来自使用量； ②计划建设大模型网关，转向以 API 为主的采购； ③网关方案预计约 4 万美元/月，2027年总支出预计大致持平

某二手车交易平台（印度业务）	①订阅约占70%-75%，API约占25%-30%；②约75%-80%的员工有企业版 AI 工具权限；③当前费用不含云与平台成本	①AI应用继续向更多城市推广，总支出预计再增约30%；②近期订阅与 API 比例大致稳定；③若比例保持不变，两项将分别贡献约70%-75%和25%-30%的增量（FUNDA情景测算）
某电商运营服务企业	①约 150 个 Claude 企业版订阅席位，合计订阅费用折合约3,000美元/月；②订阅约占10%-15%，API及按量费用约占85%-90%；③当前内部支出约 2 万-3万美元/月	①订阅席位数量预计基本稳定，更多员工通过内部产品使用AI；②年底月度支出预计约 3 万-4万美元，较当前中值增约40%；③增量主要来自API，具体贡献比例未量化

来源：FUNDA企业调研；FUNDA测算。

注：订阅费用可能包含平台许可和额外使用额度，并非全部为固定订阅席位费。制药企业的 API 相关支出为较广范围估计。覆盖率与实际使用率分别指获得权限和实际使用的情况，不能相互替代。增量贡献为某项新增支出占总新增支出的比例，不等于该项未来支出占比；表内 FUNDA 估计与情景测算已单独标明。

1. 订阅席位基本稳定后，后续增长更多来自 API 调用和现有用户的使用量。

多元化控股企业和支付企业已覆盖全员，预计后续新增支出全部来自API；电商运营企业的订阅席位数量也基本稳定，主要通过内部产品扩大 AI 使用。薪酬服务企业则先增加现有工具的积分与 Token 用量，再计划通过大模型网关转向 API 采购。员工权限已开通后，应用能否形成持续调用量，比继续增加账号更能影响这几家企业的投入。

a. 支付和电商运营企业订阅席位趋稳，新增费用主要来自 API。

- i. 某支付技术服务企业当前约 20 万美元/月的 AI 费用全部为员工订阅，公司预计订阅席位数量和档位保持稳定。筹备中的 API 项目用于智能支付、交易保护、欺诈和拒付管理，预计 2027 年一季度开始增加费用。未来约 6 个月，API月支出预计增至约 10 万美元，将总额推至约 30 万美元。
- ii. 某电商运营服务企业约 150 个企业版订阅席位预计基本稳定，更多员工将通过内部 AI 产品使用模型。当前内部支出约 2 万-3万美元/月，年底预计增至 3 万-4万美元，新增费用主要来自 API 调用。

b. 薪酬服务企业的近期增量来自使用量，采购方式计划随后转向 API。

- i. 某薪酬与人力资源服务企业（欧洲业务）约 200 个 ChatGPT 订阅席位已覆盖主要使用人群，约 60 名开发者另用编程工具。当前订阅及额外积分费用约 3.6 万美元/月；至年底总支出预计再增约25%，主要来自现有员工使用量增加。公司计划通过大模型网关接入不同模型，减少对现有工具及额外积分采购的依赖。

c. 员工覆盖和地域推广尚未完成的企业，订阅仍是重要增量。

- i. 某电信运营商计划将付费 AI 订阅席位覆盖从不足10%提高至30%-40%，新增采购主要面向财务、市场和销售。公司月度支出目标由 1 万增至 3 万新元，订阅占比预计由50%提高至70%。按这一目标测算，订阅费用增长约320%，API费用增长约80%，分别贡献约80%和20%的新增支出（FUNDA测算）。API 占比虽然下降，绝对费用仍在增加。
- ii. 某二手车交易平台（印度业务）约75%-80%的员工已有企业版 AI 工具权限，但跨城市推广仍需要增加订阅和 API 用量。订阅目前占支出约70%-75%，API约占25%-30%，近期比例预计大致稳定。若这一比例保持不变，订阅也将贡献约70%-75%的新增支出（FUNDA情景测算）。

d. 订阅和 API 仍可同步增长，各自的增量取决于使用方式。

- i. 某数字营销技术与服务企业已向全员提供 AI 工具，人均订阅支出仍预计在未来 6 个月增长20%-30%。生产 API 主要服务客户沟通等应用，随客户和流量增加而增长。两项都可能增加费用，但其各自对总增量的贡献尚未分别量化。
- ii. 某工业软件企业预计明年使用 AI 的员工约翻倍，新增内部工具和跨部门智能体将带动 API 调用。公司同时通过自托管开放权重模型控制许可费用。FUNDA 估计后续增量约60%来自 API、40%来自订阅，应用扩张与单位成本优化将共同影响最终费用。

2. 采购方式调整也会提高 API 占比，未必带来同等幅度的总支出增长。

薪酬服务企业计划把部分工具订阅和额外积分采购迁入大模型网关，预计以 API 为主的方案与当前总费用相近。工业软件企业则计划增加开放权重模型自托管，降低外部许可费用，但仍需承担算力和运维支出。API 占比上升可能同时包含新增应用和原有采购的迁移，不能全部视为新增需求。

a. API审批收紧后，广告媒介企业的费用扩张受到约束。

- i. 某广告媒介代理企业的 API 用户数量较少，部分应用费用却较高。公司要求新增用例取得高级管理人员支持，并清理闲置用户和密钥。过去 6 个月，API费用下降约10%-20%，订阅费用增加约10%，总额大致持平；未来 6 个月，两项费用也预计保持稳定。

b. 大模型网关可改变采购结构，薪酬服务企业预计总费用大致稳定。

- i. 某薪酬与人力资源服务企业（欧洲业务）计划通过大模型网关，为不同任务选择适合的模型，并逐步转向 API 采购。拟议方案约 4 万美元/月，与当前约 3.5 万-5万美元/月的总费用大致相近，2027年支出预计持平。采购方式变化主要提高模型选择和费用管理的灵活性，尚不代表总支出已经下降。

'...we will be in control of it, and we will be able to actually deliver better models for some users, and lower-priced models for other users, but they wouldn't notice.'

1. 扩大自托管有望减少外部许可采购，但仍需承担算力和运维费用。

- a. 某工业软件企业当前外购模型与工具费用约 4 万-5 万美元/月，其中 API 约 3 万美元；另有自托管与租赁机房费用约 5 万-6 万美元/月。公司计划扩大开放权重模型的内部部署，以减少外部许可采购。随着内部应用继续推广，API 和基础设施仍有新增需求，节省许可费并不意味着整体 AI 投入下降。

2. 新项目落地和业务回报，将影响 API 支出能否持续扩大。

支付企业后续投入取决于智能体支付等业务的采用情况；电信运营商要求新增 API 费用能够带来降本或增收；制药企业则在用例验证价值后追加投入。对这些企业而言，新应用能否进入持续运行的业务流程，比单纯增加试用更能决定后续调用规模。

a. 智能体支付的采用情况，将影响支付企业后续 API 投入。

- i. 某支付技术服务企业计划围绕智能体支付开发交易保护、欺诈和拒付管理等产品。如果相关服务得到市场采用，API 投入可能进一步扩大；若采用进度较慢，公司将维持较小规模的试验，不会立即大幅追加投入。

b. 新增应用仍须通过业务论证，预算随价值验证逐步释放。

- i. 某电信运营商要求每笔新增 API 支出说明能够削减的成本或带来的收入，内部应用替代既有软件，是其扩大 AI 预算的重要依据。
- ii. 某制药企业已将采购、事件登记和销售支持等智能体推向规模化应用，现有用户承担更多任务将增加调用量。新获批用例还可获得产品线或业务线的独立投入，因此增长既来自已有应用扩大使用，也来自新增项目。订阅与 API 各自承担的增量尚未分别量化。

三、ROI 与投资决策

AI 预算与项目回报相关：已验证价值的项目继续获得资金，新增应用需要说明降本、增收或效率改善的空间。部分企业通过替换外购软件、削减其他 IT 开支为

AI 腾出预算，或由受益的业务部门承担应用成本。收益既体现为软件许可费等现金支出的减少，也体现为同样的团队服务更多客户、处理更多业务。员工通用工具的收益较难单独衡量，生产应用和新增 API 项目则更容易对应具体的业务结果。

样本企业	AI 预算制定主要依据	ROI 具体实现情况
某数字营销技术与服务企业	①由产品工程团队（主要是 AI 部门）决定，受访者倾向认为不在 IT 预算内；②新产品先做业务调研（受访者称属常规产品管理），避免将人力与 Token 消耗在失败产品上，AI 支出多在线上线后发生；③投入偏向面向客户的产品，不用于内部流程精简	①用看板关联 AI 成本与输出质量，API 侧按每次调用追踪；②受访者称回报良好、超过 2 倍，未披露具体倍数及完整测算口径；③回报主要体现在客户留存和 AI 生成对话的质量。
某广告媒介代理企业	①API 与数据科学费用计入集团层面的 AI 预算，与项目技术费分开，目前不向客户收费；②各代理机构按用例申请 API 预算，须附业务论证、获非直属上级的高级管理人员背书并注明服务的客户；③超额费用仍会支付，由财务事后问询	①按节省工时和员工可承担的业务量评估，尚难精确折算为现金回报；②同一团队可以服务更多客户；③受访者举例，某产品可节省 30%-40% 的工作时间，用于其他有价值的工作。

某多元化控股企业	①AI与IT合并编预算，AI约占20%，预计3-5年后升至30%-40%；②公司集中管控，部门可削减其他开支转投AI，追加额度按业务绩效审批；③效果突出的工具可申请IT之外的专项预算	①上线前估算回报，受访者称无法精确计算；②收入端看新增客户数和新增用例，不看价格；③成本端看节省工时，尤其是减少加班（加班费率为正常水平的1.5倍）；④已兑现回报未量化
某工业软件企业	①受访者向CTO、CEO及投资者展示成效后，争取到单列于IT之外的专项预算，约\$250万/年；②项目开始回收投入后，管理层允许按需上调；③按研发、安全、监管合规分项预测下一年度支出，先按上限编制，再通过议价与降本压缩支出	①ROI按（总财务收益减总拥有成本）÷总拥有成本计算，收益按节省工时折算，单个方案须达30%；②上年约35%，今年截至9月约35%、全年预计至少40%-45%，明年预计50%-60%；③受访者称已为公司节省数百万（未说明币种、范围与期间）
某制药企业	①最终决策由更高层级的管理人员作出，用例由数字化部门与业务部门共同决定；②AI预算独立于IT，用例单独编预算、均须附业务论证，中央预算覆盖安全、合规、治理、培训与实验；③预算有上限，追加须论证，PoC或MVP验证价值后可获得更多预算	①按用例衡量降本、效率和增量销售，另看研发与生产指标（如候选药物产出时间、批次放行）；②采购、事件工单、销售拜访准备等智能体已从试点转入规模化部署；③回报金额或比例未量化
某支付技术服务企业	①订阅由集团统一决定、向全员开通，区域团队目前不参与决定；②API与自建项目由区域团队选定未来约6个月的方向，报集团审批（尚无API支出）；③含在IT预算内，约占5%，不单列，目前低于预算；④新增投入按回报与适用性共同评估	①以同样的客户经理团队服务更多商户，作为AI的回报目标；②订阅按固定订阅席位收费，尚未正式计算或追踪ROI；③新增支出同时评估回报与适用性，包括节省金额、欺诈与拒付拦截、停机时间及流程所需人手。
某电信运营商	①CEO决定，纳入IT预算、不单独核定，今年低于IT运营支出的5%；②数年前已结束转型，自上而下要求IT支出持平或小幅下降；③增量预算须论证可削减的其他运营支出或可带来的增收	①以现金节省衡量，对比自建成本与所替代SaaS的许可费；②原营销自动化平台许可费约20万-25万美元/年，自建应用运行费约3,000美元/月，另有约5,000美元的一次性建设支出，不含人力成本；③员工生产力收益尚未单独量化。
某薪酬与人力资源服务企业（欧洲业务）	①今年依据粗略估计编制预算，目前已不足；②目前由欧洲业务承担、集中计入IT，约占IT总成本10%-20%；③预计下一财年起下放各业务单元；另主张面向客户的产品在核算损益时应计入AI成本（未给时间）；④计划按用例登记的预估模型与用量编预算	①尚未定义ROI，也未建立效率基线；②计划结合大模型网关和用例登记，追踪模型使用与相关成本，再衡量收益；③已实现的回报尚未量化。
某二手车交易平台（印度业务）	①服务全公司的共享项目由中央IT团队出资；②只服务单一业务、无法复用的项目由业务部门出资，IT预算本身也由业务拨付；③AI支出约占IT预算10%-15%；④预算按既定优先项目核定，下一财年将依据成效重新分配	①生产应用按收入、客户体验和效率评估；②约35%-40%的项目回报为正，约40%-45%接近盈亏平衡，其余仍在改善；③相关AI业务活动量约为此前的1.5倍，不能等同于收入增长；④员工订阅的回报仍较早期，预计约6个月后更清楚。
某电商运营服务企业	①约占IT预算的高个位数百分比；②决策层包括CEO；③用例由中央IT与业务部门提出， workflow改造由技术、产品、数据与业务共同推进	①尚无全公司层面的ROI框架，主要观察员工可承担的业务量；受访者举例，同样3人的团队可由支持1家店铺增至支持3家；②目前未裁员，侧重人员转岗，未来可能减少初级岗位招聘；③后续收益取决于店铺运营、采购等工作流改造的进展。

来源：FUNDA企业调研；FUNDA测算。

注：表中回报指标包括受访者估算的 ROI

、项目盈亏状态和运营效率，彼此不宜直接比较；工时折算收益不等同于现金节省。电信运营商的金额为美元，自建应用的一次性建设支出不含人力成本；其余未单独注明币种金额沿用受访者的“\$”表述。标为“举例”的情形不代表已实现结果。

1. 部分企业通过削减原有开支增加 AI 投入，业务部门也参与承担应用费用。

AI 预算的来源正在随应用落地而分化。电信运营商要求 IT 总支出持平或下降，新增

AI 投入需要由其他运营支出的节省支持；多元化控股企业允许部门削减其他开支，增加 AI 用量。制药企业和二手车交易平台由业务部门为具体项目出资，薪酬服务企业也计划在下一财年把相关费用划归业务单元。与此同时，工业软件企业通过展示项目成效，获得了 IT 预算之外的 AI 专项资金。

a. 节省原有 IT 开支，可为新增 AI 应用提供资金。

- i. **某多元化控股企业**不把 AI 与 IT 分开编预算，受访者的解释是两者相互关联，引入 AI 工具后部分 IT 支出可能下降；AI 目前约占合并预算的 20%，预计 3-5 年后升至 30%-40%。部门可削减自身其他开支，以增加 Token 用量，追加额度按业务绩效审批。
- ii. **某电信运营商**处于常规经营期，AI 不单独核定预算。今年相当一部分 AI 支出不在原计划内（未给比例），通过削减其他成本，在已分配的 IT 预算内安排；下一财年计划把部分 IT 预算转用于 AI，业务论证须说明哪些现有企业订阅可由自建应用替代、能削减多少运营支出。

b. 部分企业正将 AI 应用费用交由使用 and 受益的业务部门承担。

- i. 某制药企业的中央 AI 预算主要用于安全、合规这类无法直接衡量回报、但属于业务运营前提的支出；越来越多的智能体用例附有业务论证，由产品线或业务线出资，而不是动用中央预算。受访者认为，这类新项目的支出较难提前估计。

'on a more of use case basis, that's much harder to... to anticipate, because it's coming from the use case... from the product budget, or from the business line'

1. 某二手车交易平台（印度业务）按项目能否复用确定出资方，两类项目的出资边界清楚，IT 预算本身也由业务拨付。

2. 某薪酬与人力资源服务企业（欧洲业务）目前把 AI 费用集中计入 IT 部门。受访者预计下一财年起 AI 费用全部划归各业务单元承担，由各部门按本部门的订阅席位和开发者用量编入预算；计划中的大模型网关可把成本分摊到个人、团队或部门。

3. 项目审批开始将 AI 调用与具体客户、业务用途对应。

- a. 某广告媒介代理企业在集团 AI 预算下按用例审批 API 支出。申请须说明服务的客户，并获得非直属上级的高级管理人员支持；超额使用会触发财务询问，反复超限且缺乏合理用途的密钥可能被停用。

4. 项目成效支持追加预算，新应用则围绕预期降本和增收争取投入。

工业软件企业通过展示项目价值争取追加资金，电信运营商以替代软件许可费支持扩大投入，二手车交易平台按项目的收入、客户体验和效率表现评估成效。制药企业继续推广已验证的智能体，支付企业准备启动 API 项

目，电商运营企业计划扩大工作流应用。企业在支持已有应用的同时，也为预期带来增收或效率改善的新项目配置资金。

a. 已验证的项目成效，为追加投入和下一年度预算提供依据。

- i. 某工业软件企业通过展示成效获得约\$250万的年度 AI 专项预算，项目开始回收投入后可申请增加资金。公司将节省的工时按任务量和小时费率折算为收益，要求单个方案 ROI 达到30%；受访者预计下一年度 ROI 可达50%-60%。
- ii. 某电信运营商已用自建应用替换外购营销自动化平台，并以更多同类替换作为扩大 AI 预算的依据。2027 财年的月度 AI 支出目标由约 1 万新元提高至 3 万新元，预计在未来约 6 个月达到这一水平。
- iii. 某二手车交易平台（印度业务）由中央团队支持共享项目，业务部门承担专属项目。生产应用按收入、客户体验和效率评估，目前约35%-40%的项目回报为正，约40%-45%接近盈亏平衡；后续预算配置将继续关注项目成效。

b. 回报尚不清晰时，企业更重视用途审批、成本治理和工作流改造。

- i. **某广告媒介代理企业**今年起把 API 改为按用例审批，因为“也该看回报了”，受访者认为 2025 年是支出高点。预计未来 6 个月支出持平，2027年随 IT 预算低速增长。

'let's create the workflows that actually add value, rather than keep prototyping till the end of time'

1. 某薪酬与人力资源服务企业（欧洲业务）下一年度 AI 预算预计持平。公司尚无 ROI 框架，计划通过大模型网关、用例登记和费用追踪建立管理基础，同时减少高价模型的不必要调用。
2. 某电商运营服务企业预计，内部 AI 月度支出由约\$2万-\$3万增至年底约\$3万-\$4万，增量主要来自 API 应用。公司尚未建立整体 ROI 框架，后续更关注 AI 能否扩大员工可承担的业务量，以及店铺运营、采购等工作流能否落地。
3. **员工订阅的收益较难单独衡量，新增 API 和生产应用的投入论证更具体。**
员工订阅通常按订阅席位收费，节省工时后带来的业务收益较难直接归集。API 与生产应用则更容易将调用成本对应到具体项目：数字营销企业追踪每次调用的成本与质量，广告媒介企业按用例审批 API 使用，电信运营商要求新增支出说明降本或增收效果，支付企业也将回报与适用性列为新项目的审批条件。

a. 员工工具主要改善个人效率，API项目更容易对应具体业务收益。

- i. 某支付技术服务企业目前的 AI 支出全部用于订阅，订阅席位固定收费，管理层没有单独追踪 Token - 消耗或正式计算 ROI。公司希望在客户经理人数不变的情况下服务更多商户。拟新增的 API 项目则同时考察回报和适用性，包括流程的重要性、输出准确性以及预期节省。
- ii. **某二手车交易平台（印度业务）**按项目衡量 AI 回报，各项目回报高低不一（见第 4 条）；员工订阅的回报，受访者认为“现在说为时尚早”，预计约 6 个月后才看得清。
- iii. **某数字营销技术与服务企业在 API 侧**按每次调用追踪成本与表现；订阅侧的个人使用无法追踪，也无法比较不同厂商订阅工具的使用效果。

- iv. **某电信运营商**认为若不监测员工订阅工具带来的生产力收益，就难以论证其回报；额外的 API 支出则须以降本或增收论证，未来 API 用量要看需求，很难预估。

b. 缺少效率基线的企业，仍需先建立收益衡量机制。

- i. 某薪酬与人力资源服务企业（欧洲业务）尚未明确 ROI 定义，也没有效率基线。公司计划结合模型网关和用例登记，追踪应用、使用量与成本，为后续评估提供基础。

'It hasn't been defined what is the return of investment. It has only been stated that we would like to be more efficient. And if you don't have a baseline, you cannot state if you are more efficient.'

1. AI 回报更多体现为员工承接更多业务，现金节省则取决于是否减少实际支出。

广告媒介企业把节省的工时用于服务更多客户，支付企业希望在客户经理人数不变的情况下扩大商户覆盖，电商运营企业则关注团队可管理的店铺数量。这类收益首先体现为业务承接能力，提高利润仍需新增收入或实际费用节省配合。电信运营商通过自建应用替代外购软件，减少了许可费支出；工业软件企业则把节省工时的价值纳入项目回报。

a. 客户覆盖、留存和服务能力，是部分企业衡量 AI 价值的主要指标。

- i. **某广告媒介代理企业**的数据科学团队没有减员，集团赢得更多客户后，同一批人服务更多付费客户。受访者举例，某产品为一组员工省下30%-40%的时间，员工将节省的时间用于服务其他付费客户，回报通过利润率体现，效果显现较慢。
- ii. **某数字营销技术与服务企业**把 AI 支出主要投向面向客户的产品，不用于内部流程精简，回报体现为客户留存和 AI 生成对话的质量。
- iii. 某二手车交易平台（印度业务）按收入、客户体验和效率衡量生产项目的回报，相关 AI 业务活动量约为此前的 1.5 倍。约35%-40%的项目回报为正，约40%-45%接近盈亏平衡，其余仍在改善。

b. 员工人数不变时扩大业务处理能力，或减少加班，也是效率改善的体现。

- i. 某支付技术服务企业希望用 AI 帮助客户经理服务更多商户，减少业务增长对新增人手的依赖。公司尚未建立正式的 ROI 计算与追踪体系，后续 API 项目将进一步评估业务适用性、效率改善及预期节省。
- ii. **某电商运营服务企业**的客户数量有限，效率提升难以通过增加服务店铺数来兑现，受访者认为目标是尽量减少相关工作所需人员，并将员工转至价值更高的工作；目前未裁员，预计今后少招初级员工。
- iii. **某多元化控股企业**成本端以节省工时衡量回报，可通过减少人员或缩短工时体现，其中特别强调减少加班，因为加班费率是正常水平的 1.5 倍；受访者也把“部门减人、多用 Token、业绩更好”视为应当批准追加预算的情形。

c. 替代外购软件可直接减少许可费；业务增长时，节省工时也可用于处理更多工作。

- i. 某工业软件企业的 ROI 公式把节省工时乘以任务量和小时费率计作收益，即“按劳动价值计算”；单个方案 ROI 须达到30%才算成功，受访者认为这一门槛对 AI 项目已经很高。
- ii. 某电信运营商用约 3 个月自建应用，替换了一套年许可费约 20 万-25万美元的营销自动化平台。自建应用运行费约3,000美元/月，一次性建设支出约5,000美元、不含人力成本。公司计划继续替换其他外购订阅，以实际费用节省支持下一年度 AI 投入。

四、预算与使用管理

企业正通过用户额度、项目审批和模型选型控制 AI 成本，同时继续支持已验证价值的应用。制药企业对开发工具设限，却允许成熟的生产智能体扩大部署；广告媒介企业收紧 API 审批后，相关支出有所回落。成本优化与总支出增长可以同时发生：更便宜的模型、缓存和采购议价降低了部分费用，客户流量、区域推广和新 workflow 又带来新增需求。模型选择也更重视任务匹配，小模型承担常规工作，复杂任务才使用高成本模型。

样本企业	当前管理方式	额度 / 超额使用	成本控制 / 模型路由
某数字营销技术与服务企业	Budgeting: 未提及公司设定的 Token 配额，通过按用例评估和选择 API 模型控制成本	①企业版订阅按岗位配置；②订阅端的个人使用效果较难单独衡量；③API侧按每次调用追踪 Token 费用与输出质量。	①通过第三方平台统一接入API，并追踪每次调用的成本与质量；②每4-6个月及新模型发布时评估，生产以中端模型为主；③约一年前更换模型后费用明显下降，此后随客户流量增长回升。
某广告媒介代理企业	Budgeting: API使用由去年的基本放开改为按用例审批，须有高级管理人员背书并注明服务客户	①Copilot额度较高，受访者未见有人达到上限；②超额自动计费、财务事后问询，反复超限且无合理用例者会被停用；③曾因忘记关闭模型，一夜产生\$数千费用	①无技术路由，也无统一选型规范，先看客户所用平台，再看团队偏好；②做过一次闲置用户与权限清理（时间未说明）；③以智能体 workflow 替代员工反复调用 Copilot，减少 Token 用量
某多元化控股企业	Budgeting: 核心 AI 与 IT 预算集中管控，团队与事业部可削减其他开支以增加 Token 用量（受访者未明确归类）	①订阅内含 Token 额度，超出后按 Token 计费；②编程用量大的团队用额度更高的高级套餐；③超额金额与频率未量化	①为对客智能硬件产品自建按意图分流的机制：对话请求交由云端模型处理，控制指令交设备端小模型；②内部智能体的路由向供应商采购；③按回答风格分别选用两家闭源厂商的模型，不用最新前沿模型
某工业软件企业	Budgeting: 约\$250万/年 AI 专项预算，编制时先按上限估算，再通过议价与成本优化压缩；管理层对追加投入不设硬上限	①员工与团队的 Token 额度未谈及；②云端 API 按预测用量向云厂商预购配额，换取折扣和可预期的成本	①一项 API 采购经议价由约\$25万降至约\$5万，服务质量不变（期间及计费范围未单列）；②研发复杂任务使用Fable，内部工具逐步采用 Llama、Mistral 等开放权重模型，DeepSeek在自托管测试中；③推进缓存、自托管和中间件管理，下一年度重点加强 AI 费用管理（FinOps）。

某制药企业	Budgeting与 Maxing 并存：通用开发与编码工具有使用限制；已上线、价值可衡量的生产智能体，在安全与任务边界明确后不限制部署	开发与编码工具设限，跟踪 Token 消耗并衡量价值；额度数值与超额情况未量化	①按任务性能、消耗和复杂度评估模型，优先选用能够胜任的低成本模型；②标准化任务使用较小模型，复杂推理采用前沿模型；③持续优化生产智能体的 Token 和算力消耗。
某支付技术服务企业	Budgeting：订阅席位费纳入 IT 预算，目前支出低于预算；受访者认为无明显预算约束，深度使用受限于监管与合规风险	①订阅席位固定收费，订阅席位内 Token 不另计费，受访者称管理层不在意消耗量；②预算为员工入离职预留约5%订阅席位余量；③API尚未启用	①据受访者理解，办公套件 AI 助手的自动模式按请求选用最低成本的可用模型；②凭借与供应商的合作关系，议价下调订阅席位价格；③管理层对 API 持审慎态度，仍在评估
某电信运营商	Budgeting：通常按预算执行，新增 API 支出须说明预期的降本或增收效果	①订阅内用户自选模型与思考档位，额度过早耗尽后自行调整；②顶级前沿模型有限开放；③个人上限的设定方未说明；④超额付费安排未谈及	①无动态路由，设计阶段按应用固定模型，每个应用使用一个密钥；②云厂商模型平台上的 AI 网关按应用监测 Token 费用，对敏感提示脱敏；③编码与复杂任务用 Opus，文本转 SQL 经评测用 Sonnet 或 Haiku，摘要也用较低档位模型
某薪酬与人力资源服务企业（欧洲业务）	Budgeting：已对开发者设用量上限；受访者认为整体治理尚未到位；今年预算基于粗略估算编制，目前已不足	①订阅内含额度有限，重度使用需另购积分；②开发者达到用量上限后须说明业务理由并申请追加；③通过限额和拟建模型网关控制高价模型调用。	①厂商默认调用前沿模型，受访者认为支出因此偏高；②计划自建大模型网关，按用户分配模型，领取密钥前须登记用例
某二手车交易平台（印度业务）	预算管控：按用户层级设置 Token 池，每日补充；不同岗位的使用上限不同。	①工程、数据科学与 AI 团队基本不设用量上限；②运营人员设额度上限；③具体额度和追加安排未量化。	①小模型处理常规任务，前沿模型处理复杂或含义不明确的任务；②采用缓存、上下文记忆、去重、批处理和量化等优化；③受访者估计，特定工作负载的低成本方案费用可低至较高成本方案的约十分之一。
某电商运营服务企业	Budgeting：初期不设限、鼓励全员使用，因额度消耗过快，改为按人设置支出上限	①每人每月支出上限为\$1,000，包含 API 用量；②超出上限须申请批准。	①无统一路由（FUNDA归类）；②API支出约70%用于 Opus，可节省比例未估算；③面向无订阅席位员工的内部 AI 平台在计划中

来源：FUNDA企业调研；FUNDA测算。

注：预算管理方式反映当前执行安排；允许申请追加额度，不等于没有预算约束。模型选型、统一 API 接入和按请求动态路由并非同一机制，表中分别说明。具体成本比较仅适用于所列项目或工作负载，不代表全公司可实现同等降幅；未量化的项目不据此视为不存在。

1. 额度和审批管理趋于严格，已验证价值的应用仍可获得追加投入。

限额更多用于约束缺乏明确用途的消耗，而非停止所有增长。电商运营企业允许员工在说明理由后申请超额使用，薪酬服务企业对开发者追加积分设审批，制药企业对生产智能体继续扩大部署。广告媒介企业则在收紧 API 审批后减少了相关支出。预算控制的效果因此取决于被压缩的是低效使用，还是能够带来业务收益的新增应用。

a. 员工用量设置上限后，有业务理由的需求仍可申请追加。

- i. 某电商运营服务企业为每人设置每月\$1,000的支出上限，包含 API 用量；超出上限须申请批准。公司仍计划扩大内部 AI 平台和工作流应用，预计内部 AI 月度支出到年底升至约\$3万-\$4万。
- ii. **某薪酬与人力资源服务企业（欧洲业务）** 给每名开发者设了用量上限，起因是 AI 编程工具厂商不断把默认模型升级到新版本，开发者若未留意默认设置，就可能切换至高价模型；有正当理由可以申请追加额度，否则回到人工开发。支出仍在上升：以年初月度支出为100，当前约 200、年底接近250（受访者估计）；下一财年 AI 预算按与今年持平的水平编制。

b. 部分企业按岗位和应用价值区别管理，而非统一限制所有用量。

- i. 某二手车交易平台（印度业务）按用户层级设置 Token 池并每日补充。工程、数据科学和 AI 团队基本不设用量上限，运营人员则有额度限制；模型选择同时按任务难度区分，常规工作使用小模型，复杂或模糊任务才调用前沿模型。
- ii. **某制药企业**对通用开发与编码工具设限并跟踪 Token 消耗，要求衡量价值；已上线、价值可衡量的生产智能体，在安全要求、防护措施和任务边界明确后不限制部署，仅优化 Token 消耗、算力使用及端到端流程。
- iii. **某电信运营商**对顶级前沿模型仅作有限开放；订阅内用户自选模型与思考档位，额度过早耗尽后自行调整，追加安排未谈及。

c. 广告媒介企业收紧 API 用途审批后，相关支出有所回落。

- i. 某广告媒介代理企业去年基本放开 API 密钥的使用，受访者所在团队曾有人忘记关闭模型，一夜产生\$数千费用。现在密钥仍绑定到个人，但申请须写明用例和服务客户，并由非直属上级的高级管理人员背书；超出约定额度后自动计费、照常付费，反复超限又没有合理用例的才会被停用。访谈前约 6 个月 API 支出估计降10%-20%，订阅支出随新员工开通账号增加约10%，总额大致持平。受访者将 API 支出下降归因于用例审批收紧和原定的高阶平台计划取消，并预计未来 6 个月整体持平。

d. 企业也通过项目审批和模型评估控制投入，管理方式不限于用户额度。

- i. **某多元化控股企业**的订阅内含 Token 额度，超出后按 Token 计费，编程用量大的团队用额度更高的高级套餐；核心 AI 与 IT 预算由公司集中管控，团队和事业部可以削减其他开支，以增加 Token 用量。过去约 6 个月支出增约30%（受访者估计）。
- ii. 某数字营销技术与服务企业在生产 API 侧追踪调用成本与质量，并定期复评模型。公司根据模型在实际用例中的表现和费用作出选择，新增需求主要来自客户和流量增长。
- iii. **某支付技术服务企业**的订阅席位固定收费，订阅席位内 Token 不另计费。过去约 6 个月支出由零起步；API预计约 3 个月后将开始使用，受访者担心缺乏规划会出现没有回报的 Token 消耗，预计约 6 个月后将月度 AI 支出较当前增约50%，新增部分全部来自 API 自建项目。

'they don't care how many tokens can be burned because we... we are not being charged for it, they're okay about it.'

1. 成本优化与需求扩张并行，部分费用下降并未带来 AI 总支出的持续回落。

数字营销企业更换模型后，费用一度明显下降，随后随客户流量增长回升。工业软件企业通过议价压低一项

API 采购费用，但内部应用推广仍推动整体投入增加。二手车交易平台持续优化缓存和模型配置，支出同时随地域和工作流覆盖扩大而增长。相比之下，广告媒介企业通过收紧用途审批、清理闲置权限并取消平台计划减少 API 费用，费用变化来自使用管理和项目调整。

a. 采购议价降低部分费用，内部应用推广继续带来新增需求。

- i. 某工业软件企业按预测用量预购 API 配额，一项采购经议价由约\$25万降至约\$5万，质量不变。与此同时，广义 AI 月度支出在过去 6 个月由约\$17万增至约\$20万，内部智能体与跨部门应用继续扩展。当前外购 AI 模型与工具约\$4万-\$5万/月，其中 API 约\$3万。

b. 缓存和模型组合降低特定任务成本，应用推广仍推动总需求增加。

- i. 某二手车交易平台（印度业务）采用缓存、上下文记忆、去重、量化和批处理，并按任务复杂度组合大小模型。受访者估计，特定工作负载的低成本方案费用可低至较高成本方案的约十分之一。随着应用向更多地区推广，未来 6 个月总支出仍预计增长约30%。

c. 更换模型曾明显降低费用，客户流量增长随后带动支出回升。

- i. 某数字营销技术与服务企业约一年前通过模型选型，将此前约\$10万的月度 Token 费用削减了一半以上。此后，支出随客户和流量增长逐步回升，近几个月每月增长约2%-6%。公司继续按实际任务效果评估模型，认为最高价格的模型并不总是最佳选择。

'sometimes the... the most expensive model, or the quote-unquote best model, is not the answer. Maybe it's the cheap one who behaves better with our use case.'

1. 模型网关和路由有望减少高价调用，节省幅度仍取决于实际迁移效果。

- a. 某薪酬与人力资源服务企业（欧洲业务）计划通过大模型网关，将合适的任务分配给更小或开放权重模型。受访者估计，这类模型的价格可为当前较贵方案的三分之一至六分之一；网关与路由优化合计可能使总支出下降约50%。这仍是潜在节省，尚未形成实际结果。

2. 模型选择重视任务匹配，低价模型的成本优势仍需通过评估和迁移兑现。

数字营销企业定期比较模型在自身用例中的表现，电信运营商在应用设计阶段固定模型，制药企业按任务复杂度与成本选择能够胜任的模型。另一些应用按请求内容分流，例如将对话送至云端、设备控制交给本地小模型。对既有应用而言，更便宜的新模型只有通过评估并替换原有调用，才会减少实际费用；统一 API 入口本身并不等于自动选用低价模型。

a. 按任务复杂度配置模型，可减少常规工作对高价模型的依赖。

- i. 某工业软件企业把 Fable 用于代码库分析、复杂推理和代码审查，内部工具逐步转向 Llama、Mistral 等开放权重模型，并测试自托管 DeepSeek。API 调用通过中间件管理，模型选择同时考虑数据隐私、定制能力、成本和供应商依赖。
- ii. 某电信运营商在设计阶段按应用固定模型，经云厂商模型平台上的 AI 网关按应用监测 Token - 费用。编码和需要汇总多个来源信息、进行规划的生产任务用 Opus；文本转 SQL 经评测用 Sonnet 或 Haiku，摘要等结果较易预期的任务也用较低档位模型。受访者认为按意图或关键词路由未必能达到预期效果，公司也没有需要大量智能体互相调用的用例。

- iii. 某制药企业按任务效果、消耗和复杂度评估模型，选择能够胜任的低成本方案。分类、抽取等常规工作可用较小模型，复杂推理和专用智能体则采用更强模型；逼近推理极限的任务仍更多处于实验阶段。

b. 部分应用按请求内容分流，以小模型处理常规任务。

- i. 某二手车交易平台（印度业务）以小模型处理常规任务，复杂或含义不明确请求才交给前沿模型。受访者举例，聊天机器人中开放权重与闭源模型的使用比例可达约80%与20%。
- ii. 某多元化控股企业的对客智能硬件产品按意图分流：对话请求送云端模型，控制指令交设备端小模型处理。内部智能体的路由则向供应商采购，引入新工具时通常比较4-5款同类产品（非正式流程），根据输出效果作出选择。

c. 高价模型带来成本压力，部分企业计划通过网关自主选模。

- i. 某电商运营服务企业的 API 支出约70%用于Opus，尚无统一的模型路由机制。公司未量化通过更换模型可以节省的费用。
- ii. 某薪酬与人力资源服务企业（欧洲业务）认为，供应商默认使用高价模型推高了费用。公司计划自建大模型网关，并在发放密钥前登记用例，按任务选择模型；合适的小模型或开放权重模型，价格可低至较贵方案的三分之一至六分之一。

五、前沿模型

前沿模型主要用于规划、研发和复杂任务，常规生产负载仍大量使用中端或小模型。工业软件企业将 Fable 用于代码分析与审查，电信运营商将最高档模型用于规划，数字营销企业则以中端模型处理客户沟通。模型能力提升并不意味着企业会全面升级：薪酬服务企业曾短期启用Fable，后因积分消耗过快而停用；支付企业认为现有模型已能满足主要任务。进一步采用仍取决于业务需求、数据质量、实际表现和成本。

样本企业	Fable / Astra 占比	Fable / Astra 使用情况
某数字营销技术与服务企业	生产 API 为0%；订阅内用于探索和规划，费用未单列。	①采用，限订阅内：深度探索、代码任务和首轮规划；②生产 API 以Terra / Sol级中端模型为主，另用新版DeepSeek，主要任务是客户沟通回复生成；受访者认为此类任务用前沿模型是浪费
某广告媒介代理企业	约15%	①前沿模型约占 AI 支出15%；②API主要用于数据科学工作，模型选择受客户平台及团队偏好影响；③未单列可归因于前沿模型的新增用例。
某多元化控股企业	0%（公司未使用）	①未采用最新前沿模型，顾虑是成本较高、安全性尚未充分验证；②内部智能体与对客应用使用现有模型；③更重视模型对具体用例的适配及界面体验。

某工业软件企业	约5%（FUNDA估计，受访者未给出比例）	①研发团队用 Fable 进行代码库分析、复杂推理和代码审查；②以真实软件工程任务比较，表现优于人工基准才采用；③同时测试自托管DeepSeek，并关注多模态设计能力。
某制药企业	未披露（受访者不清楚）	①Fable / Astra的具体使用情况未明确；②复杂推理、专用智能体和困难的结构化任务采用更强模型，常规任务使用较小模型；③逼近推理极限的任务仍更多处于实验阶段，不能与已规模化的采购、工单和拜访准备智能体混为一谈。
某支付技术服务企业	0%	①未采用，也无需求：GPT-5（近期升至 5.1）与 Opus 、Sonnet 足够且稳定；②定价提案、文档撰写、数据查询与数据管道、代码重构已够用；③受访者认为约束在数据质量而非模型能力
某电信运营商	约10%	①采用，限规划与构思，顶级前沿模型有限开放；②效率类用例不用，受访者称成本高出数倍；③文本转 SQL 、摘要用 Sonnet 或 Haiku 即可，未发现必须用 Astra 的场景，未举出由前沿模型带来的新用例；④设想的订单排障智能体需至少 Opus 级
某薪酬与人力资源服务企业（欧洲业务）	0%（Fable短期启用后已停用）	①Fable曾向开发者开放约 14 天，因积分消耗过快停用；②供应商默认选用高价模型仍带来成本压力，公司已设置开发者用量上限并计划建设模型网关；③数据保留政策尚未造成可观察到的用量变化，但合规要求延缓了部分实施。
某二手车交易平台（印度业务）	未量化（受访者未给出）	①已测试或使用Fable，具体占比未量化；②常规任务使用小模型，复杂或含义不明确的任务使用前沿模型，未明确识别出新的独特用例；③数据保留政策未造成明显减用，持续进行权限审计、个人信息脱敏和模型监测。
某电商运营服务企业	0%（未用 Fable 5.1 或Astra）	①主要依靠Opus，约占 API 支出的70%；②未明确识别出由模型升级直接带来的新用例，受访者个人感受是 Opus 推理表现较数月前减弱；③通过合同约定数据保护要求，数据保留问题未导致整体使用下降。

来源：FUNDA企业调研；FUNDA测算。

注：本节前沿模型主要指Fable / Astra，不等同于企业使用的全部高价模型。占比按表内范围理解，数字营销企业的0%仅指生产API；未量化不代表未使用。部分比例为 FUNDA 估算。零数据保留（ZDR）相关变化只反映受访企业观察到的影响。

1. 前沿模型更集中于探索、研发和复杂任务，常规应用仍重视成本与稳定性。

数字营销企业在订阅中使用前沿模型做探索和规划，生产 API 仍以中端模型为主。工业软件企业将 Fable 用于复杂研发任务，电信运营商的常规 SQL 查询与摘要则使用较低档位模型。二手车交易平台按任务难度组合大小模型。这些做法表明，企业更关注模型能否以合适的成本完成任务，而非单纯追求最高能力档位。

a. 固定订阅额度让员工能够尝试前沿模型，生产调用则更直接面对成本约束。

- i. 某数字营销技术与服务企业的前沿模型用量基本都来自订阅工具：员工日常用前沿模型做深度探索和代码任务，受访者本人先用它做首轮规划，再换便宜模型执行；订阅按订阅席位固定收费，无法单独拆分这部分用量对应的支出。生产 API 以 Terra / Sol 级中端模型为主、另用新版 DeepSeek，主要任务是客户往来沟通的回复生成；受访者认为这类系统化任务用前沿模型并不划算。

'We use it a lot on the personal applications, on the subscriptions, because it is available to us for a cheaper rate'

1. 某电信运营商将 Fable / Astra 用于规划与构思，前沿模型约占 AI 支出 10%。生产应用更注重成本效益：文本转 SQL 和摘要采用 Sonnet 或 Haiku，编程等错误代价较高的任务使用 Opus。受访者认为，现阶段尚无必须由 Fable / Astra 完成的业务任务。

2. 研发和数据科学，是前沿模型较明确的采用场景。

- a. 某工业软件企业的研发团队用 Fable 做代码库分析、复杂推理与代码审查，采用标准是在真实软件工程中优于人工基准。前沿模型支出占比估算约 5%；公司同时使用开放权重模型支持内部工具，并未把全部研发和内部负载转向前沿模型。
- b. 某广告媒介代理企业的前沿模型支出约占 15%。其 API 主要服务于数据科学相关工作，模型选择通常先考虑客户使用的云平台，再考虑团队偏好；前沿模型并非公司统一指定的默认方案。

3. 薪酬服务企业因高价模型带来的费用压力，加强用量限制并推进自主选型。

薪酬服务企业认为，供应商默认调用高价模型使成本偏高，已对开发者设置额度，并计划通过自建网关选择更适合的模型。该企业短期试用 Fable 后因积分消耗过快将其停用，说明能力提升仍须经受成本约束。二手车交易平台则按任务难度配置模型，以小模型处理常规工作。对高价模型的需求因此更取决于实际任务价值和企业选型，而非是否被纳入工具的默认配置。

a. 薪酬服务企业因积分消耗过快，停用了试用中的 Fable。

- i. 某薪酬与人力资源服务企业（欧洲业务）曾向开发者开放 Fable 约 14 天，随后因积分消耗过快停用。公司还认为，供应商默认使用高价模型削弱了企业对费用的控制，计划通过大模型网关按用例选择模型。合适的小模型或开放权重模型可显著降低单价，但网关带来的整体节省尚待实施验证。

b. 按任务难度配置模型，有助于把前沿能力用于更需要它的工作。

- i. 某二手车交易平台（印度业务）用小模型完成大部分常规任务，复杂或含义不明确的工作才使用前沿模型。企业 AI 工具已覆盖约 75%-80% 的员工，后续增长主要依靠工作流和地区扩展；公司按岗位设置额度，工程、数据科学与 AI 团队基本不设用量上限，运营人员则受限额管理。

c. 相关企业通过合同和合规管理应对数据保留风险，目前未见明显减用。

- i. **某电商运营服务企业**称合同已约定供应商不得暴露公司数据，所以不会因此减少使用，但也承认公司无法核查供应商是否履约。公司并未使用Fable 5.1，这一表态针对整体使用。
- ii. 某薪酬与人力资源服务企业（欧洲业务）尚未观察到数据保留政策对使用量的明显影响，但合规要求延缓了部分应用的实施。模型是否适用，仍需按具体用例完成审查。
- iii. 某二手车交易平台（印度业务）没有因数据保留政策明显减少使用。公司约每两个月审计权限、个人信息脱敏及模型表现，持续监控智能体的访问和操作范围。

4. 新模型能否带来新增投入，取决于是否解决业务问题，数据准备和场景适配仍是关键条件。

现有模型已能处理起草、摘要、数据查询和客户沟通等许多日常任务。支付企业更关注数据质量，多元化控股企业重视用例调校和界面体验，电商运营企业需要先完成数据准备与内部产品建设。工业软件企业则关注多模态设计，制药企业探索药物设计和临床方案优化，电信运营商设想跨系统排障。上述需求能否形成新增投入，取决于模型能力与业务实施的共同推进；广告媒介企业也将重大模型能力提升列为未来 API 支出可能增加的条件。

a. 数据质量、业务适配和使用条件，影响更强模型的实际价值。

- i. 某支付技术服务企业认为，现有模型已经能完成定价提案、文档撰写、数据查询、数据管道和代码重构等任务，尚无采用Fable / Astra的需求。相比继续升级模型，受访者更重视数据质量，认为基础数据不可靠会直接限制 AI 系统的效果。
- ii. 某广告媒介代理企业的模型选择受到客户环境和法务审批约束。即使模型能力提高，客户仍需确认其数据可以在相应环境中使用；受访者认为，数据保护与法律边界更清楚后，开放权重模型的使用才更有机会增加。
- iii. 某电商运营服务企业未使用 Fable 5.1 或Astra，主要依靠Opus，约70%的 API 支出用于该模型。后续店铺运营自动化、智能采购和内部工作流扩展，仍需数据准备和产品建设配合。
- iv. 某多元化控股企业未采用最新前沿模型，主要顾虑是成本较高及安全性尚未充分验证。公司认为，企业应用的效果更取决于对具体用例的调校和界面体验；现有内部智能体、对客聊天机器人和智能硬件已采用不同的模型与部署方式。

b. 多模态设计、药物研发和跨系统执行，仍对模型能力提出更高要求。

- i. 某工业软件企业关注多模态能力，希望根据文字、图片等输入帮助客户生成三维设计。对这类应用，能否适配设计任务、安全处理数据并保持透明，与模型能力本身同样重要。
- ii. 某制药企业将更强模型用于复杂推理、专用智能体和困难的结构化任务，也通过专有数据微调模型支持药物开发。药物设计与优化仍是难点：生物和化学体系复杂且不断变化，模型能力提升并不自动转化为研发成果。逼近推理极限的任务仍更多处于实验阶段，而采购、工单和拜访准备等工作已实现规模化应用。
- iii. 某电信运营商设想以智能体跨系统诊断订单问题：读取订单管理及后端系统日志，对照操作规程与架构文档定位故障，再执行修复。受访者认为这需要 Opus 或更强模型；相比之下，日历、文件与客户信息查询等个人助理任务并不需要最高档前沿模型。

六、开放权重与闭源模型

部分企业已将开放权重模型用于生产 API、内部工具和客户项目，主要看重任务适配、数据控制和成本。工业软件企业计划继续提高开放权重模型承担的工作负载比例，薪酬服务企业准备通过网关引入较小及开放权重模型，数字营销企业已在生产中使用 DeepSeek。企业既采用云平台托管，也采用自托管。成本优势取决于具体任务及部署方式：部分任务的模型费用明显较低，但 GPU、托管和维护费用可能抵消这一优势。

样本企业	开放权重与闭源模型使用情况	对开放权重模型的态度
某数字营销技术与服务企业	①生产 API 以中端模型为主，也使用新版DeepSeek；②开源相关支出约\$1,000-\$2,000/月；③按月度总额折算约占3.75%（FUNDA估算），开放权重推理费用未单独拆分。	偏积极：①根据具体用例的效果和成本选用，不只看价格；②对通用基准测试宣传保持保留态度，重视自身评估；③FUNDA估计 2026 年底、2027 年底相关支出占比分别约5%、6%。
某广告媒介代理企业	①以闭源为主，员工订阅席位与 API 均与模型厂商直接签约；②受访者所知的开放权重使用只有 1 个客户指定项目，在本地笔记本电脑上运行；③受访者称总体以闭源为主，开放权重用量很少（FUNDA估计占比2%）	偏积极：①认可开放权重模型便于调整和微调；②认为闭源厂商更容易通过法务审批；③若数据保护与法律边界更清楚，支出可能更多转向开放权重模型；④FUNDA估计 2026 年底、2027 年底占比2%、3%
某多元化控股企业	①当前模型使用基本为闭源，开放权重支出占比记为0；②内部主要使用办公套件 AI 助手，对客应用使用云端模型；③智能硬件的控制指令由设备端小模型处理，模型权重属性未明确。	未评级（受访者未表态）：未谈开放权重的成本、风险或采用计划；FUNDA估计 2026 年底、2027 年底占比仍为0
某工业软件企业	①API工作负载中闭源约60%、开放权重约40%；②开放权重相关支出占外购 AI 模型与工具费用约13%（FUNDA估算），与工作负载比例不同；③另有约\$5万-\$6万/月的自托管基础设施及租赁机房费用，支持模型托管、测试和数据存储。	积极：①选型考虑数据隐私、定制能力、总成本和供应商依赖；②计划下一年将开放权重工作负载提高至约60%，5年目标约80%并自托管；③FUNDA估计 2026 年底、2027 年底开放权重相关支出占比约15%、20%，不能与负载目标直接比较。

某制药企业	①开放权重用量未量化，未点名具体模型；②受访者负责的用例大多托管在一家公有云上，调用云平台智能体框架提供的模型，无本地或私有部署；③内部 AI 助手基于第三方闭源大模型	偏积极：①按任务选择模型，分类、抽取和重复转换可用较小开放权重模型，复杂推理用更强模型，专有数据场景使用微调模型；②主要通过云平台提供的框架使用模型，开放权重费用未单列。
某支付技术服务企业	全部为闭源：支出均为办公套件 AI - 助手订阅席位，经其使用 OpenAI - 与 Anthropic 模型；尚未使用 API，目前不自托管	未给出明确态度：目前未使用开放权重模型，也未提出具体采用计划；未表达对模型来源或数据保留的看法。
某电信运营商	全部为闭源：生产 API 经两家云厂商的模型平台调用闭源模型，无本地或自建推理；敏感提示先经云厂商模型平台上的 AI 网关脱敏	积极：①认为开放权重模型能力持续提升，约 270 亿参数的小模型已有竞争力；②当前 API Token 仍受模型厂商补贴，现价下 API 更便宜；③正在评估，预计长期内开放权重与闭源模型并存、逐步偏向开放权重，未给时间表或占比；④障碍是缺少本地推理的硬件与技能
某薪酬与人力资源服务企业（欧洲业务）	目前使用闭源模型，开放权重占比为0；拟在模型网关上线后引入较小及开放权重模型。	积极：①网关上线后以云端部署为主，按用例完成合规审查；②敏感负载可能使用小规模本地模型；③尚未给出年底开放权重占比目标。
某二手车交易平台（印度业务）	①尚无全公司统一比例，方向性估计开放权重与闭源约各半；②聊天机器人场景可达到约80%开放权重、20%闭源，未明确分母；③通过云平台和自有 GPU 混合部署。	积极：①按任务在成本、准确率和延迟之间权衡；②在特定工作负载下，低成本方案的费用可低至较高成本方案的约十分之一；③尚无固定的全公司开放权重比例目标。
某电商运营服务企业	全部为闭源：只用一家厂商（Claude）的企业版订阅席位与API	中性：①若现有模型回报不足，可能采用混合方案，开放权重使用或从 2027 年开始；②因业务跨地区，云端部署的可能性高于本地部署；③具体模型与比例尚未确定。

来源：FUNDA企业调研；FUNDA测算。

注：支出占比、工作负载和特定场景的模型比例不可直接比较。数字营销企业约3.75%及工业软件企业约13%的支出占比为 FUNDA 估算；后者以外购

AI 模型与工具费用为基础，另有自托管基础设施费用。工业软件企业40%、60%和80%指工作负载；二手车企业的约各半及80:20未明确分母。未来比例属于规划或估计，不代表已经实现。

1. 开放权重模型已用于部分生产调用与内部工具，网关建设有望支持更多模型替换。

数字营销企业已在生产 API 中采用DeepSeek；工业软件企业将开放权重模型用于内部工具和私有数据，并计划扩大相关工作负载；二手车交易平台按任务需要混用开放权重与闭源模型。薪酬服务企业当前仍使用闭源模型，计划在网关上线后引入更多选择。开放权重对闭源支出的替代，首先取决于企业能否自行选择模型，以及相应工具和部署方案是否就绪。

a. 高吞吐、标准化和私有数据任务，是开放权重模型的重要应用场景。

- i. 某工业软件企业将闭源模型用于复杂推理等任务，开放权重模型（如Llama）用于高吞吐处理、数据抽取转换和私有数据处理，DeepSeek仍在自托管测试中。开放权重模型目前承担约40%的 API 工作负载，计划明年提高至约60%、5 年内提高至约80%。

- ii. 某二手车交易平台（印度业务）按任务配置小模型与前沿模型。受访者举例，聊天机器人可采用约80%的开放权重模型与20%的闭源模型；复杂或含义不明确的任务仍需更强模型。上述比例仅针对聊天机器人场景。
- iii. 某制药企业按照任务选择模型：分类、抽取、快速问答和重复转换可由较小的开放权重模型承担，复杂推理和专用智能体使用更强模型，专有数据场景则使用微调模型。相关应用主要通过云平台框架部署，开放权重的具体支出未单列。

b. 通过自身用例评估后，开放权重模型可替代部分中端闭源调用。

- i. 某数字营销技术与服务企业的员工订阅是两家闭源厂商的企业版，开放权重用在生产 API。公司每4-6个月用自有评估集重测模型，生产 API 以中端闭源模型为主，最新 DeepSeek 也在生产中使用；受访者举例称，评估显示新版 DeepSeek 优于当时所用的 GPT 模型后就做了替换。

c. 订阅工具的模型范围和现有供应商安排，会影响开放权重采用。

- i. 某支付技术服务企业的 AI 支出全部是办公套件 AI 助手的订阅席位，模型只能在助手提供的 OpenAI - 与 Anthropic 模型之间自动或手动选择，尚未使用 API；受访者预计约 3 个月后才开始用 API 开发产品。
- ii. 某电商运营服务企业的 API 及按量费用约占总支出的85%-90%，目前主要使用 Claude 企业版和-Opus API，尚无统一的模型路由机制。公司计划根据回报情况考虑混用开放权重模型，具体模型尚未确定。

d. 企业正通过网关和自托管工具，扩大自主选择模型的空间。

- i. 某薪酬与人力资源服务企业（欧洲业务）目前仍使用闭源模型，计划在大模型网关上线后引入较小及开放权重模型，以云端托管为主。网关迁移后的 AI 支出预计更多体现为 API 费用，约\$4万/月，处于当前约\$3.5万-\$5万的月度支出范围内；此后能否进一步降费取决于模型路由和用例优化。
- ii. 某工业软件企业正将部分内部工具转向自托管开放权重模型，以满足私有数据处理和定制需求，并降低对单一云平台的依赖。公司计划扩大开放权重工作负载，同时仍需承担 GPU、托管、维护与人员能力建设等成本。

2. 自托管能增强数据与模型控制，但不必然比云端 API 更便宜。

工业软件企业将部分模型与私有数据放在租赁机房内，既满足数据驻留要求，也支持模型托管和测试；电信运营商则认为现阶段 API 更便宜，自建推理还受到硬件与技能约束。二手车交易平台混用云平台和自有GPU，薪酬服务企业计划优先采用云端方案。开放权重采用并不等于新增自建机房，资金流向取决于负载规模、数据要求和企业运维能力。

a. 部署、维护和团队能力，决定模型低价能否转化为整体成本优势。

- i. 某工业软件企业租赁美国和欧洲的机房容量，每月相关基础设施费用约\$5万-\$6万，包括托管、电力和硬件维护，支持模型运行、测试及数据存储。自托管有助于满足数据驻留和内部财务数据处理需求，也提供更多定制和硬件选择；但公司认为，GPU、托管和维护可能使其成本高于价格更可预测的云端方案。

- ii. **某电信运营商**认为约 270 亿参数的小模型已有竞争力，但按当前成本估算，API 仍更便宜；自建推理所需的硬件、模型架构和推理服务技能，公司尚不具备，云上租 GPU 偏运营支出、本地部署偏资本支出。

'Currently, I believe a lot of the... the API token is still subsidized by the... by the AI labs, right? ... Because it seems to be cheaper if you are consuming API, but I think it will not be forever like that'

- 1. 某二手车交易平台（印度业务）通过云平台和自有 GPU 混合部署模型，在成本、准确率和延迟之间按任务取舍。受访者称，在特定工作负载下，低成本方案的费用可低至较高成本方案的约十分之一。

2. 使用低价模型可以减少部分调用费用，整体节省仍取决于适用任务的比例。

- a. 某数字营销技术与服务企业通过定期评估，将适合的生产任务交给更低成本模型。约一年前的模型调整使月度 Token 费用明显下降，之后客户和流量增长又带动支出回升。公司当前开源相关支出约\$1,000-\$2,000/月，但没有单独拆出开放权重模型对历次降费的贡献。
- b. 某薪酬与人力资源服务企业（欧洲业务）认为，合适的小模型或开放权重模型价格可低至较贵方案的三分之一至六分之一。网关和路由优化可能使总支出降低约50%，但这是组合优化的潜力，并非开放权重模型单独带来的节省。

3. 数据要求影响部署和选型，但并不必然导向自托管或开放权重。

工业软件企业以私有部署满足部分客户的数据驻留要求，电信运营商仅将本地或租用 GPU 列为敏感负载的备选，薪酬服务企业则计划以云端为主、为高敏感任务保留小规模本地模型。广告媒介企业的模型选择还需要客户和法务批准。数据控制、部署位置与模型是否开放权重是不同问题，企业会根据具体用途组合选择。

a. 数据驻留推动部分私有部署，其他企业仍将本地推理作为备选。

- i. 某工业软件企业在租赁机房中处理部分客户要求驻留的数据，以及内部财务等私有信息。开放权重模型便于调整与微调，也降低了对单一供应商的依赖；公司计划扩大相关负载，同时继续使用公有云服务，并未全面转向自托管。
- ii. **某电信运营商**目前经云厂商模型平台上的 AI 网关对敏感提示脱敏后调用闭源API；受访者认为部分负载敏感，需要考虑开放权重模型，超敏感负载可能要本地部署或在云上租GPU，尚无时间表。
- iii. 某薪酬与人力资源服务企业（欧洲业务）计划在网关上线后优先使用云端模型，并按用例完成合规审查。对于敏感数据负载，公司保留小规模本地部署的可能，当前尚未采用这类方案。

b. 处理客户数据时，模型选择还需通过客户与法务审批。

- i. 某数字营销技术与服务企业的部分客户排斥某些来源的模型，另有客户因隐私法规限制而不能使用某些模型。受访者称，多数情况下各客户使用同一套模型；据其了解，部分客户已有单独安排，可能包括回退方案，具体安排不由其负责。
- ii. **某广告媒介代理企业**的受访者认为闭源厂商更容易通过法务审批；去年即使用闭源厂商的API，也难以明确数据是否始终保留在所用环境内。受访者所知唯一的开放权重项目由客户指定，为确保数据不离开公司而部署在本地笔记本电脑上，客户的数据要求决定部署位置，不决定是否用开放权重。公司也有不涉及客户数据的内部用例，受访者把它列为开放权重支出可能增加的方向之一。

'because everything has to go through legal, you know, the clients have to also approve that, you know, that, you know, nothing is going out'

1. 合规审查按具体用例开展，采用计划仍需逐项落实。

- a. 某薪酬与人力资源服务企业（欧洲业务）计划引入较小及开放权重模型，但每个用例仍须经过合规审查。模型成本下降并不会免除这一要求，应用推进速度仍受实施与治理安排影响。

七、应用场景

企业 AI 应用已覆盖客户沟通、提案生成、采购、工单和数据查询等业务流程。制药企业已将采购、工单登记与销售拜访准备智能体规模化部署，数字营销企业的 AI 直接支持客户沟通，支付企业则希望借助 AI 提高客户经理的商户覆盖能力。后续应用推广将继续扩大现有工具的使用范围，店铺运营、多智能体协作和药物研发等场景也有望带来新的需求。

注：文中区分已使用、规模化部署、开发测试及未来设想；尚未说明成熟度的用例不视为已经投产。应用范围包括生成式 AI 及其他 AI 方法，涉及具体模型时另行说明。

1. 客户应用扩大覆盖，有望持续带动生产 API 需求。

数字营销企业以 AI 处理客户沟通，控股企业在对客产品中部署聊天机器人和智能硬件，二手车交易平台用于客户互动与运营流程，薪酬服务企业则在客户产品中加入聊天和客服功能。随着相关业务扩大，生产 API 调用也有望增加。

a. AI 已进入客户沟通和产品服务流程，具体任务仍按需要保留人工处理。

- i. **某数字营销技术与服务企业的核心产品**用 AI 为客户的呼入与呼出沟通生成回复，由工程部门的 AI 应用团队通过 API 调用中端模型完成，另以成本最低的小模型对调用记录评分，评估模型表现。其他场景也在探索，但进度较慢，公司将重点放在已有收入的现有产品上。
- ii. **某多元化控股企业**在对客产品中嵌入聊天机器人，按产品需求分别使用两家厂商的模型：需要结构化回答时使用一家厂商，需要简洁自然的回答时使用另一家。公司另运营 AI 呼叫中心，用于减轻人工负担，是否用大模型未说明。
- iii. 某二手车交易平台（印度业务）将 AI 用于客户互动、客户服务和运营工作流，并以收入、客户体验与效率评估生产项目。不同任务按复杂度选择模型；后续推广重点是扩大工作流和地区覆盖，而不只是增加员工工具权限。

- iv. 某薪酬与人力资源服务企业（欧洲业务）已在客户产品中加入聊天机器人和客服功能；公司同时希望通过模型网关和更多 workflow 应用扩大 AI 使用。

b. 多模态能力为设计软件带来新的产品方向。

- i. 某工业软件企业希望将文字、图片等输入转化为三维设计，帮助客户更便捷地生成设计模型。相较于单纯提高文本推理能力，这类需求更依赖多模态能力与设计工作流的结合。

c. 电商运营企业因成本偏高，暂未采用 AI 直播。

- i. 某电商运营服务企业与几家 AI 直播服务商沟通后认为，相比真人直播，现有方案价格仍偏高，回报提升也不明确，因此尚未采用。未来如提供这项功能，主要是满足品牌客户的尝试需求，暂不以盈利为目标。

2. 内部 AI 已从单项辅助扩展到采购等完整流程，跨部门应用仍在继续增加。

制药企业的采购智能体已覆盖端到端流程，拜访准备和事件工单也已规模化部署；支付企业的智能体生成销售提案，电信运营商把自然语言查询接入数据系统。后续，二手车交易平台将推进多智能体与运营自动化，电商运营企业计划改造店铺运营和采购流程。应用能否扩张，既取决于模型能力，也取决于数据、系统接入和业务流程建设。

a. 提案、采购、工单和数据查询，已成为内部 AI 的具体业务用途。

- i. **某支付技术服务企业**的销售团队广泛使用 AI 编写提案：将提案模板作为智能体和技能配置到办公套件 AI 助手中，指定一份按卡组织、借记与贷记、境内与国际卡组合整理的定价数据文件，AI 按模板生成给企业客户的正式提案，列明客户需支付的费用及可能节省的金额。
- ii. **某制药企业**有三类智能体已从实验转入面向大量用户的规模化部署，均接入内部 AI 助手：拜访准备智能体在销售代表拜访医疗专业人士之前，从客户全景数据平台调取信息，提供行动建议、沟通话术及对方关注点，销售代表无需再自行查找资料；采购智能体根据员工请求执行端到端采购流程；事件登记智能体根据一句自然语言指令生成工单。内部 AI 助手本身用于写邮件、回复消息，员工还可在该助手部署执行专门任务的智能体。
- iii. 某多元化控股企业将办公套件 AI 助手接入内部系统，用于会议室预订、报告和营销内容生成、人员与制度查询，以及财务分析。营销图像和内容使用增长较快，但基数较低；公司仍重视具体用例调校，而非全面追逐最新模型。
- iv. 某工业软件企业正在扩展法务、财务和人力资源应用，包括追踪客户新闻、分析客户财务信息以支持续约，以及生成入职培训内容。部分内部工具使用开放权重模型，涉及私有或财务数据的任务在租赁机房中处理；这类应用扩展也是后续用户增长的来源。
- v. **某电信运营商**已在生产中上线查询智能体：员工用自然语言提问，智能体借助语义层把问题转成 SQL，经评测，使用较低档位模型即可。市场部门是仅次于 IT 的第二大使用部门，在做流失预测、营销活动自动化、图像生成和活动 workflow；流失预测是否用大模型未说明，可能是传统机器学习。公司已用自建应用替代一套外购营销自动化平台（见第三节）。财务部门的对账与数据测算正由传统商业智能与报表系统转向大模型，目前用量尚少。

b. 模型网关建设有望降低内部智能体的接入门槛。

- i. 某薪酬与人力资源服务企业（欧洲业务）计划通过大模型网关统一模型接入、登记用例并分配权限，随后扩大智能体工作流。现阶段的重点是建立工具与治理基础，新增应用数量和实际支出仍取决于后续落地。

c. 运营自动化继续向更多环节扩展，数据准备和系统整合仍需推进。

- i. 某二手车交易平台（印度业务）已有约15-20个项目投入生产，超过 50 个在推进，另有超过 10 0 个处于构思或早期阶段。后续重点包括更复杂的多智能体系统、运营工作流自动化和地区扩展。公司已向较大比例员工提供 AI 工具，新增需求更取决于这些应用的推广，而非单纯增加权限。
- ii. 某电商运营服务企业计划将 AI 用于店铺运营自动化、智能采购和更多内部工作流。目标是让同一团队承担更多店铺与运营任务，但这依赖数据准备和内部产品建设；受访者以 3 人由支持 1 家店铺增至 3 家说明潜在效率，尚不能视为已经实现的收益。
- iii. 某广告媒介代理企业正在开发结案复盘自动化工作流，将投放数据整理、分析和客户汇报连接起来，减少员工反复拼接演示材料的工作。这是受访者当前岗位的首要任务，目标是让团队把更多时间用于解释投放效果。团队还在开发一个指标推荐智能体，根据市场和预算等条件建议应衡量的投放指标；公司已有提示词库和内部智能体。
- iv. 某电信运营商的受访者设想一个订单排障智能体，跨多个后端系统诊断失败订单并执行修复，是否立项未说明；任务细节与模型要求见第五节。

d. 药物研发、注册写作和生产质量管理，为制药企业提供更多 AI 应用方向。

- i. 某制药企业关注 AI 在靶点发现、基因组和单细胞数据分析、生成式化学、临床方案优化及数字孪生中的应用。医学与注册文件可由 AI 起草，再由专业人员审阅；制造环节则包括偏差识别和处置。药物设计与优化仍是较难的方向，受访者认为，生物与化学体系的复杂性使其难以仅靠模型升级解决。

3. 监管与业务可靠性要求，使核心支付和薪酬作业采用 AI 更为审慎。

支付企业强调，涉及交易安全的错误可能造成难以追回的损失，因此对实时交易判断保持审慎，主要将 AI 用于辅助工作和事后分析。薪酬服务企业的 AI 更多用于开发、通用知识工作和客户产品，薪酬处理岗位直接采用仍有限。两类企业并非没有 AI 需求，而是需要根据具体流程的错误成本、监管要求和人工责任逐步推进。

a. 支付流程更重视人工把关，薪酬处理岗位的直接采用仍有限。

- i. 某支付技术服务企业认为，当前深度采用 AI 的主要约束是监管与流程风险，而非预算。公司对实时交易判断保持审慎，AI 更多用于交易之外的辅助流程和事后分析，复杂过程仍保留人工参与。

'if it's a complex process, we always need human in the loop, along with AI. AI can be an aid in that case, but it can't replace humans.'

- 1. 某薪酬与人力资源服务企业（欧洲业务）的薪酬处理岗位直接采用 AI 仍较有限；当前主要应用集中在开发、研究写作、演示材料以及客户聊天和客服产品。

2. 健康问答通过专属检索与专业数据，约束输出内容。

- a. **某多元化控股企业**的对客产品里有健康问答。受访者认为通用模型在健康话题上容易给出过长的回答，风险大，因此把健康问答部署在云端专属区域，自建检索增强系统，并接入专业数据库，要求回答可靠、简短。

免责声明

免责声明与使用条款

分析师独立性声明

本报告所载观点均真实反映分析师对相关证券及发行人的个人独立观点。报告酬劳之任何部分，均不会直接或间接与本报告所载具体建议或观点相关联。本公司恪守严格合规流程，以确保所采用数据来源及信息之完整性、安全性及合法性。

关于 FUNDA

Consensus AI Technology Pte. Ltd. (“本公司”或

“FUNDA”)系一家注册于新加坡的独立研究机构。本公司不隶属于任何证券交易商、投资银行或资产管理公司，亦未在任何司法管辖区注册为投资顾问或经纪交易商。本公司与本报告所涉公司之间不存在任何投资银行或承销业务关系，亦不接受上市公司的赞助或资助。

非投资建议与报告中立性

本报告内容及其所载信息仅供参考，不构成任何投资建议或商业决策依据。报告中观点仅代表相关观点持有者个人在特定时间点的看法，不代表 FUNDA 的立场。作为独立运营的平台，FUNDA

保持中立第三方地位，不对上述观点的准确性、完整性或适用性作出任何明示或暗示的保证。

- 非邀约：** 本报告不构成买卖或持有任何证券或金融工具之邀约、招揽、推荐或建议，亦不构成任何法律、税务或会计建议。
- 第三方数据：** 本报告中引用的第三方数据均来自公开或授权渠道，本公司不对其准确性或完整性作出任何明示或默示保证。
- 不参与决策：** 本公司不以任何直接或间接形式参与客户或读者之投资决策过程。请读者根据自身投资目标及财务状况做出独立投资决策，并在进行任何交易前咨询专业财务或法律顾问意见。

时效性声明

鉴于市场环境与技术趋势的快速变化，本报告反映的是截至发布日之前的市场情况，相关信息可能随时间推移而失效。FUNDA 无义务对报告内容进行实时更新。

知识产权、保密与商业使用限制

本公司提供的一切报告、数据、模型及内容均构成本公司独家专有知识产权，受版权法和知识产权法保护。

- 权利归属：** 除非书面协议另有明确规定，因本报告产生或与之相关之一切知识产权（包括著作权）均独家归属于本公司。
- 授权使用：** 机构客户可在其内部合规政策允许之范围内使用本报告，但严禁向任何第三方披露或共享。
- 禁止商业使用：** 未经本公司事先明确书面同意，任何个人或实体不得复制、转载、分发、转售、重新包装或以其他方式将本报告或其任何部分用于任何商业或盈利目的。
- 法律追索：** 任何未经授权之商业使用、再分发或侵犯本公司知识产权之行为，本公司将依法追究其一切法律责任及追索权。

前瞻性陈述与免责声明

本报告可能包含前瞻性陈述，包括预测、估算及行业趋势分析。该等陈述受制于可能导致实际结果与预期产生重大差异之风险及不确定性。对于因使用或依赖本报告内容而产生之任何直接、间接或附带损失，本公司概不承担任何责任。