

中转站不收过路费

中转站不收过路费 —— 推理路由层的租金落在支付环节，所以买家是一家支付公司

推理路由层对 token 推理零加价，这不是竞争压出来的结果，而是写进挂牌价的产品承诺：平台费 5.5% 收在充值环节，加密通道 5%，BYOK 在每月 25,000 美元（企业版 200,000 美元）挂牌推理额度内为零。这层赚的是支付费不是路由费，费基是充值金额而非撮合价差。由此，能给这个资产出价的只能是收单方——云厂与模型商各自自建路由层且不缺流量。规模上，2026 年 5 月官方周吞吐 25 万亿 tokens、全年逾 1 千万亿，对 Google 全平台年化约 3.4%，是次干道与长尾试验场而非主干道；企业侧约 88% 的用量走直连合同。已上市公司里这一层不存在任何具备法定披露的收入口径，最硬的横向反证来自 Shopify：智能体交易与普通交易经济性完全相同、无新增费率。

推理路由层的租金落在支付环节，所以买家是一家支付公司

涌现资本产业研究部 · 2026 年 8 月 27 日

30秒结论卡

一句话判断：推理路由层对 token 本身的加价是零——这是它写在挂牌价上的产品承诺，不是竞争压出来的结果。它唯一的挂牌租金是 5.5% 的充值平台费，那是支付费，不是过路费。所以这一层的价值捕获从来不归路由，归计量与收单。

项目	内容
机会状态	● 「路由层将成为 AI 过路收费站」这一叙事 · ● 计量与收单层现任者 · ● 纯路由标的（上市载体不存在）
市场忽略了什么	这一层的收费结构是公开挂牌的：推理零加价、平台费 5.5% 收在充值、加密通道 5%、自带密钥在每月 25,000 美元挂牌推理额度内为零。费基是充值金额，与撮合价差无关
市场误解了什么	把「支付公司买下路由层」读成支付公司跨界买 AI 资产。方向是反的——路由层的收入本来就产生在支付环节，这是收单方买回本属于自己的那段
重点标的	SHOP ●（反证锚） · NET ●（有位置无口径） · V / MA ●（计量与追责层现任者） · DDOG ● · AKAM ●
最大反证	若路由层把平台费从充值环节移到 token 环节并成功维持（即出现挂牌加价），或自带密钥免费额度被显著下调而客户不流失，则「无 token 定价权」这一支柱失效，全篇需重写

投资要点

推理这条链上，token 的价差归拥有硅的人，token 的账单归收单的人，中间负责路由的那一段，按自己的挂牌价什么都不留。

判断	方向	依据
路由层对推理零加价，是产品承诺而非市场结果	●	官方定价页与常见问题页两处并存：推理定价直通上游、平台费只在购买额度时发生
唯一挂牌租金是支付费：5.5% 平台费、加密通道 5%、企业版另有费率折扣	●	定价页三档并列，企业档明示可议折扣，故全平台实收混合费率必低于 5.5%
公司自建了一条把费率打到零的合法旁路	●	自带密钥每月 25,000 美元（企业版 200,000 美元）挂牌推理额度内零费，超出部分 5%
费基方向不确定：量在涨、单价在跌，而它撮合的恰是最价格敏感的负载	●	费基正比于推理支出 = 用量 × 单价；商品档单价已被打到成本地板附近
这一层是次干道：年化约占全平台 token 的 3.4%，企业用量约 88% 走直连	●	周吞吐 25 万亿（2026 年 5 月官方）年化 1.3 千万亿，对 38.4 千万亿
已上市公司里这一层无任何法定披露的收入口径	●	有金额的读数全在旁边：云基础设施、安全、可观测性
新流量层拿不到溢价，在支付侧独立复现	●	智能体交易与普通线上交易经济性完全相同、无新增费率、无单独定价
计价单位并未收敛到成果制，反而两处回退	●	企业订阅回到纯席位、编码工具从计次退回按用量、代码助手从请求额度退回消耗额度

一、钱流：这一层到底在哪一步收钱

推理路由的商业模式在公开定价页上是完全透明的，三档并列，没有需要推演的部分。

收费项	免费档	按量档	企业档
推理加价	0	0	0
平台费（充值时收取）	不适用	5.5%	可议折扣
加密通道费	不适用	5.0%	可议折扣
自带密钥免费额度（按挂牌推理成本计）	不适用	25,000 美元/月	200,000 美元/月
超出免费额度后	不适用	5%	5%
在架模型 / 供应商	25+ 免费模型	500+ / 80+	500+ / 80+

三件事由这张表直接读出，都不依赖任何估算：

1. **加价为零是承诺，不是现状。**推理价格直通上游供应商，客户支付的费率与直连供应商完全相同。这意味着「上游降价挤压中转层转售毛利」这条常见推理链在这里不成立——它本来就没有转售毛利可被挤压。

- 2. 收费点在充值，不在调用。5.5% 是一笔支付手续费，费基是充值金额，与撮合了多少 token、撮合了哪家模型无关。把它称作路由费是口径错位。
- 3. 企业档明示可议折扣，故混合实收费率必然低于 5.5%。方向确定，幅度未披露——这是本篇唯一需要读者接受的不确定项，而它只影响幅度不影响方向。

费基的方向是本篇最该被追问的一件事。5.5% 的费基正比于推理支出，而推理支出 = 用量 × 单价。用量在高速增长，但商品档单价正被打到边际成本附近；这一层撮合的恰恰是价格最敏感、最愿意为省钱而比价的那部分负载。用量与单价方向相反，费基的合成方向不可先验断定为向上。

二、价值链五层：路由被放在哪一格

- 需求源（应用与智能体开发者）
- 一级供应商（前沿模型商 / 三大云的模型目录）
 - 真卡口（先进封装、高带宽存储、供电与并网）
 - 可投代理（拥有硅的平台、计量与收单的现任者）
 - 已拥挤 / 未定价（路由与网关：流量真实，租金归属不在此层）

路由层落在链条的最后一格，且落在那一格里「未定价」的一侧——不是因为它没有流量，而是因为它的流量与它的收入之间没有定价接口。它对流量的计量能力是真实的，对流量的收费权是零。

三、卡点六指标：非卡点，且四特征中最硬的一条被官方口径直接否掉

指标	评分	依据与置信度
短缺强度	弱	供给侧极度充裕，开源自托管方案零许可费可得（LiteLLM 57,354 星标、10,923 复刻，2026-08-27 实测）
替代难度	低	三大云各自内建模型目录与路由，独立路由层不是稀缺资源
认证周期	无有效证据	企业规模化后是留在路由层还是转直连，正反两向均无具名一手案例
集中度	低	在架 80+ 供应商，切换成本主要是配置成本
定价权	零	官方挂牌推理加价为零，且自带密钥旁路把费率打到零。这是本篇唯一的确证级读数
可投性	无上市载体	唯一纯载体已被未上市公司收购，二级市场无可量化敞口

四象限归属：假卡点很热 → 回避。需要说明的是各条的证据强度并不相同：定价权不成立是官方挂牌口径确证，替代难度与短缺强度不成立是产业推论，认证周期一条证据为空。定价权这一条足以独立支撑结论，但不应替其余三条背书。

四、买家身份就是最强的口径证据

2026 年 8 月 19 日，Stripe 与 OpenRouter 双方同日发布公告，确认已达成收购协议；截至本报告日交易尚未交割，双方均拒绝披露交易对价。

交易主体的身份，比任何未披露的金额都更能说明这一层的价值归属。如果路由层的租金真的产生在路由环节，出价方应当是云厂或模型商——但这两类主体各自已经自建路由能力，且都不缺流量。真正为这个资产出价的是一家收单公司，而这层资产的挂牌收入恰恰全部产生在充值这一步。**这不是一家支付公司跨界买 AI 资产，是收单方把本来就属于支付的那段收入买回来。**

买方的规模锚只有一个可用读数：2025 年经其处理的总支付额 1.9 万亿美元（约占全球生产总值 1.6%，反算全球生产总值约 118.75 万亿美元，量级自治）。这是全维度唯一通过三道关的一手财务锚。

同样值得记录的是**卖方在公告里披露了什么、没披露什么**。公告披露了在架模型数、供应商数、开发者数；没有披露收入、抽成额、毛利。一个被出价数十亿美元量级的资产，在自己的公告里选择只讲流量不讲收入——这个披露选择本身，是关于「这层捕获了多少价值」的信息。

五、规模：次干道，不是主干道

2026 年 5 月，公司在融资公告中给出唯一一次精确的周吞吐披露：半年内周处理量由 5 万亿 tokens 增至 25 万亿，全年有望处理逾 1 千万亿，服务 800 万以上开发者、在架 400 余个模型。到 8 月的收购公告，用户口径更新为 1,000 万以上。

按此年化，1.3 至 1.5 千万亿 tokens/年。同期 Google 全平台月处理量为 3.2 千万亿（年化 38.4 千万亿），**得出份额约 3.4% 至 3.9%。**

这个分母偏宽——它包含搜索、视频、办公套件等大量从不产生第三方账单的内部自用量，因此 3.4% 应读作份额下界而非中枢。但方向不因分母而改变：企业采购视角下，前三家模型商合计约 88% 的企业用量走直连合同；而路由层平台上开源权重模型占比约三成。两套口径的抽样框不同（采购合同视角 vs 单平台结算日志视角），不可相减，但方向一致——**规模化的、有合规与保供要求的负载天然去中介化，留在路由层的是试验期、长尾与价格敏感负载。**

六、横向反证：新流量层在支付侧同样拿不到溢价

路由侧的零加价，在支付侧有一条完全独立的复现。

Shopify 在 2026 年第二季度业绩中给出公司口径：**智能体交易与普通线上店铺交易的经济性完全相同，没有新增费率，没有单独定价。**同季营收 36 亿美元、同比增长 34%，商品交易总额 1,160 亿美元、同比增长 32%，商户解决方案 29 亿美元、订阅解决方案 8.02 亿美元（子项合计与总额自治）。

这条读数的价值在于它的独立性：一个来自推理侧的定价页，一个来自电商侧的业绩披露，两者在完全不同的产业环节上得出同一个结论——**AI 带来的新流量形态，本身不构成向现有费率之上加价的理由。**卡与收单网络的季报同样只把智能体支付列为产品创新条目，未给出任何笔数、交易额或收入口径。

七、计价单位：并未收敛到成果制，反而出现两处回退

「计价单位正从 token 迁向成果」是一条流传较广的判断。公开挂牌价给出的图景要复杂得多，且有两处方向相反的移动：

主体	计价单位现状	方向
企业订阅（前沿模型商）	纯席位制，标准席位 25 美元/用户/月，2026-08-10 起新增 125 美元高级席位	席位制强化
编码工具	由「固定 500 次快速请求」改为与接口成本挂钩的用量计费，公布自有每百万 token 费率	退回 token
代码助手	由高级请求计次（含 300 / 1,500 次）改为按模型费率折算的额度制	退回消耗量
客服智能体	按结果计费 0.99 美元/次解决，49 美元/月基础包含 50 次解决，超出按次计	成果制

最关键的一处不在表内：**成果制挂牌最明确的那一家，其智能体跑在自研专有模型上**，且已于 2026 年 6 月 15 日被平台厂商以约 36 亿美元签约收购，带来 3 万余家企业客户。也就是说，成果计价的那一端不但没有把 token 账单交给中转层，反而把第三方 token 账单从买方账单上消灭了——它自研了模型，并被一家平台公司收编。

计价单位迁移对中转层构成的是威胁而非机会：售卖面越往成果与席位走，买方账单上的 token 行项越淡，中转层「帮你省 token 钱」这个卖点就越不可感知。

八、大公司受益链：需求源到底把钱付给了谁

环节	主体	是否真的收到这笔钱
需求源	智能体与应用开发者、企业 AI 预算	—
直接受益	前沿模型商与三大云（直连合同承接约 88% 企业用量）	是（未上市或不单列）
二阶受益	拥有硅的平台、计量与收单网络	是
隐形受益	推理边缘承载、中间层可观测性	部分（相邻层口径）
伪受益	以「AI 入口 / 推理过路费」为叙事但挂牌零加价的路由与网关	否——挂牌价已写明不收
可投代理	收单与计量层现任者、边缘网关（待变现率披露）	待验证
不可投但监控	交易双方（均未上市）	不适用

这条链的关键一格是「伪受益」：流量真实经过路由层，但按其自身挂牌价，经过并不产生留存。**流量份额与收入份额之间隔着一个变现率，而这一层的挂牌变现率是零。**

九、公司卡

SHOP (\$150.29)

项	内容
位置	智能体交易的收单与履约方，本篇的横向反证锚
为什么重要	它给出了「新流量形态是否配得上新费率」这个问题的唯一一条上市公司口径答案
为什么被低估	不适用——本篇不以低估立论，它在此是证据不是标的
证据	2026Q2 营收 36 亿美元 +34%、交易总额 1,160 亿美元 +32%、商户解决方案 29 亿 / 订阅解决方案 8.02 亿；智能体交易与普通交易经济性完全相同、无新增费率、无单独定价
估值状态	不在本篇结论范围内
催化	若公布智能体交易独立费率档，第六节反证锚作废
反证	智能体交易占比若仍极小，则「同费率」是尚未定价而非不能定价

NET (\$284.89)

项	内容
位置	边缘网关，处在智能体流量的必经路径与计量位置上
为什么重要	二级市场里最接近这一层的位置，且已具备计量能力
为什么被低估	不成立——市销率与增速的组合已使其估值处于透支区间
证据	2026Q2 营收 6.961 亿美元 +36%、non-GAAP 每股收益 0.29 美元；同季 GAAP 经营亏损 2.057 亿美元（含重组费用 1.507 亿）；全年指引上调至约 28.67 亿美元；智能体流量已超全网活动一半
估值状态	已透支（市销率高位，增速回落至 30% 以下将触发压缩）
催化	首次给出智能体相关收入或变现率口径
反证	流量份额是先行读数，与计费收入之间无已披露的转换关系；本篇结论正提示这段转换在路由环节挂牌为零

十、情景与触发器标的表

情景	核心假设	结果	标的动作
熊（本篇基准之下）	挂牌零加价维持，自带密钥旁路扩大，计价单位继续向席位与成果迁移	路由层长期不产生可披露收入，叙事标的估值回落	回避入口叙事标的
基准	挂牌结构不变，费基随用量增长但被单价下行部分抵消	价值捕获稳定归于计量与收单层	维持底仓，等边缘网关变现率披露
牛	边缘网关首次披露智能体变现率，或收单方推出智能体独立费率档	出现第一个可量化敞口，该层从触发器升为候选	该标的上调兑现度后再议仓位

触发器清单（非推荐清单）

标的	现价（8/27）	这一层的位置	可量化口径	状态
SHOP	\$150.29	智能体交易收单方	智能体交易费率与普通交易相同（公司口径）	● 反证锚
NET	\$284.89	边缘网关与计量位置	无产品级收入拆分；流量占比是先行读数	● 估值已透支
V	\$383.90	授权、验证与追责层	智能体支付仅列产品条目，无笔数/收入	● 底仓
MA	\$598.47	同上	同上；2026Q2 净收入 93 亿美元、报告口径 +14%	● 底仓
DDOG	\$227.67	中间层可观测性	AI 原生客户 750+ 家，年消费超百万美元 31 家、超千万美元 8 家	● 估值偏贵
AKAM	\$108.06	推理边缘承载	云基础设施收入 9,900 万美元 +39%	● 观察
FSLY	\$23.29	推理边缘承载	安全收入 4,170 万美元 +43%、占比 23%	● 观察

三点说明：其一，云基础设施与安全收入是**相邻层**读数，当作路由层敞口即层属错配；其二，流量占比与收入份额之间隔着未披露的变现率；其三，年内已签多年期合同承诺属合同期总额，不可年化。

本篇不给出任何路由层标的的**营收、流水或估值倍数**——公司零披露，第三方估算的分子与分母出自同一家机构，互证性为零。

十一、数据分歧与口径

以下分歧真实存在于各机构之间，不作取舍合并。

1. **份额差一个数量级**。按全平台 token 分母算得约 3.4%；持股方口径则称该平台约相当于 Google token 运行速率的 15% 至 30%、Azure 模型工厂的 50% 以上。两者相差近一个数量级，唯一解释是

后者采用第三方接口口径而非全平台口径。分母定义未被任何一方公开定义，故本篇份额只作量级、不作点估计。

- 2. **利益相关披露。**本主题可得的分子（平台吞吐、开源占比）与分母（企业 AI 支出总额）几乎全部出自该平台的两家股东机构。数字未必有误，但「两个独立来源」这一关实质未过，据此不作估值推导。
- 3. **周吞吐口径：**官方融资公告为 25 万亿/周（2026 年 5 月），另有结算周记录为 28.9 万亿/周，相差约 15%，故分子写作区间。流传的「每日 10 万亿」与官方年度口径不相容（年化将达官方自述总量的 5 倍以上，构成部分大于整体），不予采用。
- 4. **接口侧分母敏感性：**前沿模型商官方可确认口径为每分钟 60 亿 tokens（2025 年 10 月开发者大会），媒体口径为每分钟 150 亿，年化相差 2.5 倍，份额读数随之上移约 0.5 至 1 个百分点。
- 5. **交易对价：**官方未披露；各家援引匿名信源的报道值介于 70 亿至 100 亿美元之间且互不一致，另有报道在官宣日之前即以「已完成」表述发布。本篇不取中值、不取众数、不给点估计，亦不据此计算任何估值跳升倍数。
- 6. **在架模型计数不可跨平台比较：**各平台对「一个模型」的定义（是否计入版本、量化变体、微调、社区上传）互不相同，同一家厂商的两份文档亦出现 1,900+ 与 10,000+ 两个数字。任何跨平台模型数量排名都不成立。

十二、风险与反证

「风险」在此指会推翻本篇结论的具体条件，不是免责表述。

若出现以下情形	被推翻的支柱	处置
路由层把收费点从充值移到 token，并出现挂牌加价且客户不流失	「无 token 定价权」——本篇地基	全篇重写
自带密钥免费额度被显著下调、或该旁路被取消	「自建去中介旁路」；反向读数 = 定价权在增强	结论转向，需复议
自带密钥免费额度被显著上调	支柱不变，指向以免费换留存 = 定价权进一步弱化	加强结论
任一上市公司首次给出路由/网关的收入口径披露	「无可量化敞口」	该标的自触发器升为候选
边缘网关披露智能体流量的变现率并成规模	「流量份额不等于收入份额」	上调该标的兑现度
收单方公布智能体交易的独立费率档	「新流量层拿不到溢价」——第六节反证锚	第六节作废，需复议
交易最终未能交割	不影响定价结构结论，但削弱「买家身份即证据」这一读法	第四节降级为推论

本篇不构成任何买卖建议，第十节为触发器清单而非推荐清单。

十三、来源置信度表与数字三道关自评

来源置信度表（一）· 本篇地基

结论 / 数字	来源	日期	口径	置信度
推理零加价；平台费 5.5%、加密 5%；自带密钥 25,000 / 200,000 美元额度、超出 5%；500+ 模型 / 80+ 供应商	平台官方定价页与常见问题页（当日直接复核，逐字命中两处）	2026-08-27	挂牌价	A
收购已签署、尚未交割、对价未披露；1,000 万 + 用户	交易双方同日官方公告	2026-08-19	一手公告	A（事实） / C（金额）
收单方 2025 年总支付额 1.9 万亿美元	买方官方年度披露	2025 年度	一手	A
智能体交易与普通交易经济性相同、无新增费率；营收 36 亿、交易总额 1,160 亿美元	电商平台季度业绩	2026Q2	上市公司披露	A
云基础设施 9,900 万美元 +39%；安全 4,170 万美元 +43%；AI 原生客户 750/31/8；边缘网关营收 6.961 亿 +36%	四家上市公司季度业绩	2026Q2	上市公司披露	A
开源路由项目 57,354 星标 / 10,923 复刻	代码托管平台接口实测	2026-08-27	实测	A
周吞吐 25 万亿 tokens、全年逾 1 千万亿、800 万+ 开发者、融资 1.13 亿美元	公司融资公告及官方通稿	2026-05-26	公司自报，未经审计	B

来源置信度表（二）· 参照与分歧项

结论 / 数字	来源	日期	口径	置信度
Google 月处理 3.2 千万亿 tokens	公司主题演讲	2026-05	全平台含内部自用，仅作分母上界	B
企业生成式 AI 支出 370 亿美元（应用 190 + 基础设施 180；基础模型接口 125 亿）	创投机构年度调研	2025 年度	该机构为标的股东	B-
前三家模型商合计约 88% 企业用量	同上调研	2025 年度	采购合同视角，与结算日志视角不可相减	C
份额 15%—30% 的对照口径	同上机构评论	2026	分母未定义	C
前沿模型商接口每分钟 60 亿 tokens	开发者大会	2025-10	官方口述	C+
客服智能体 0.99 美元/次解决；平台厂商约 36 亿美元签约收购	厂商计费文档 / 收购方官方公告	2026-06-15	挂牌价 / 一手公告	C+ / A
交易对价 70 亿至 100 亿美元	多家媒体援引匿名信源，互不一致	2026-08	无一手来源	C

数字三道关自评

- 量级一致性：**周吞吐 $25 \text{ 万亿} \times 52 = 1.3 \text{ 千万亿/年}$ ，与公司自述「全年逾 1 千万亿」自洽；月 $3.2 \text{ 千万亿} \times 12 = 38.4 \text{ 千万亿/年}$ ， $1.3 \div 38.4 = 3.4\%$ 。电商平台子项 $29 + 8.02 \approx 37 \text{ 亿}$ 与营收 36 亿量级自洽（含其他项与四舍五入）；企业支出 $190 + 180 = 370 \text{ 亿}$ 、基础模型接口 125 亿 < 基础设施层 180 亿，无部分大于整体。流传的「每日 10 万亿 tokens」年化后达公司自述年度总量的 5 倍以上，触发部分大于整体，已弃用。
- 单位显式化：**全文金额均为美元；1.9 万亿美元 = trillion 非「亿」；千万亿 = quadrillion；卡组织净收入 93 亿美元采用报告口径 +14%，不与汇率中性口径 +12% 混用。
- 两源与公告优先：**本篇地基四条（挂牌费率、交易事实、收单方支付额、电商平台费率口径）均取自官方挂牌页或法定披露；份额类数字的分子分母存在同一股东来源问题，已在第十一节第 2 条显式披露，并据此不作任何估值推导。

未上市主体（交易双方、前沿模型商）无法定披露义务，其运营与财务数字置信度上限为 C，仅「双方自发公告确认交易存在」这一事实按一手公告计。本篇不含任何由未披露数字推算而来的倍数、营收或估值。

行动清单

- 现在买什么：**本篇不产生新的买入标的。路由层无上市纯载体，相邻层的读数属于相邻层的逻辑。
- 重点观察：**边缘网关的变现率首次披露（流量份额转收入份额的时点）；收单网络是否出现智能体交易独立费率档。

- **只监控不买：**以「AI 过路费站」「推理入口」为叙事的标的——这一层的挂牌价已经写明它不收过路费。
 - **绝不追：**任何以未披露交易对价推算出的估值倍数为依据的交易。
 - **下次看这五个数：**①自带密钥免费额度的方向（上调或下调）②平台费收费点是否移动 ③周吞吐的下次官方披露 ④边缘网关智能体收入口径 ⑤收单方智能体支付的首个笔数或交易额披露。
 - **哪个信号出现要重写结论：**出现挂牌加价，或自带密钥旁路被取消。
-

涌现资本产业研究部

本报告基于公开挂牌价、上市公司法定披露与公司自发公告，为信息聚合与产业结构分析，不构成投资建议。文中未上市主体的运营数字未经审计，交易对价未经任何一方确认。市场有风险，决策需独立判断。