

行业新视角

AI大模型Token中转服务行业研究

报告日期：2026-08-28

研究部

姓名：李承儒

SFC：BLN914

电话：0755-21519182

Email: licr@gyzq.com.hk

行业背景：Token成为AI时代硬通货

- 生成式AI浪潮席卷全球，大模型能力主要以API形式对外输出，**Token是衡量与交易AI服务的“硬通货”**，其获取与使用效率直接决定应用开发的成本、性能与体验。
- 直接调用全球各家模型API，面临四大挑战：
 - 网络与地域限制：Claude等国际顶尖模型存在直接访问障碍
 - 支付与结算壁垒：海外模型普遍要求国际信用支付，部分地区用户与企业不便
 - 技术集成复杂：各厂商接口协议、认证方式、错误代码各异，多模型集成维护成本高
 - 成本与稳定性波动：定价动态调整、官方服务宕机频现，缺乏成本优化与风险规避机制

应运而生：连接上下游的数字立交桥



- **定位：**精准切入上游供给与下游需求之间的市场缝隙，构建中间技术与服务层
- **对用户：**一站式AI能力货架——统一API Key按需调用全球多种模型
- **核心价值：**解决网络限制、支付壁垒、集成复杂、成本波动四大痛点

海内外头部Token中转平台

海外龙头

- **OpenRouter**: 整合GPT-4、Claude、Mistral等20余种主流模型，提供统一API、智能路由、性能基准与开发者工具
- **Portkey**: 专注企业级可观测性、DevOps功能与治理层，路由为核心能力之一
- **Martian**: 基于机器学习的智能路由算法，自动实现成本与性能的最优化
- **Vercel AI Gateway**: 深度集成Vercel前端开发生态，简化Web应用的AI集成体验

国内相关企业

- **七牛云AI**: “国内直连型”聚合平台，支持OpenAI/Anthropic双协议，覆盖超130个模型
- **硅基流动**: 主打国产开源模型，主打免费额度和极低的成本，是预算有限或仅需使用国产开源模型用户的理想选择
- **EasyRouter**: 猎豹移动2026年5月推出的，聚合GPT、Claude等40余个大模型，以便捷的方式满足国内开发者出海需求

海内外头部Token中转平台估值快速攀升

OpenRouter 估值短时间内从 13 亿跃升到70 亿+



在大模型厂商调用量快速攀升且API接入成本居高不下的背景下，OpenRouter以中间层路由的轻量模式卡位，一端对接OpenAI、Anthropic、Gemini、Mistral等多家模型供应商，另一端服务全球开发者。三年时间做到82人团队、估值站上10亿美元门槛，成为AI基础设施层不可忽视的玩家。2026.08.16 Bloomberg曝Stripe已敲定收购协议。中间曾出现100亿美元传闻价但未落地，最终协议价仍是B轮估值5倍以上。

生意本质：典型的长尾服务

Token中转站可定性为大模型的销售渠道，更多承担API出口、调用加速、模型整合的作用，并不掌握大模型本身的核心能力。也就是说，它处在产业链的中游偏末端，属于大厂之外的补充性服务。Token中转站在技术层依赖开源方案、门槛较低，很多顶尖产研人员也更倾向于做自有API而非中转站。门槛低意味着参与者众多、单点规模有限，呈现典型的长尾格局。

- **渠道环节而非核心层：**定位“大模型的销售渠道”，承担API出口、调用加速、模型整合，不掌握大模型核心能力
- **门槛低、竞争分散：**技术依赖开源方案，顶尖产研人员更倾向做自有API，参与者众多、单点规模有限
- **头部体量有限：**头部中转项目月流水约500万元、毛利率近50%，属“走量”的中间生意，并非平台级核心业务
- **重服务、轻产品：**区域政企与制造业项目依赖现场调研、系统联调、培训、运维，本质是多模型、多账号、多成本的统一治理平台

Token消耗量爆发式增长，Token中转商市场空间仍在攀升

- 中国Token吞吐量：2024年约142.5万亿次 → 2025年约2,426.3万亿次，同比增长1,602.6%
- 驱动因素：AI使用形态由单次问答转向连续任务执行、复杂目标拆解、多步骤工具调用与多模态交互，调用频次、上下文长度、并发需求同步提升
- 远期展望：2026年预计跃升至约5千万亿次；2030年预计达约53.2百亿亿个，2025-2030年CAGR约638.3%
- Token供应市场将释放长期、高确定性的增长潜力

大模型公司ARR高速增长

公司	口径	2024	2025	2026H1	2025→2026H1 增量
OpenAI	ARR	\$6B	\$20B+	\$40B（年化，7月底）	2026/1 披露2x增长
Anthropic	ARR（gross）	~\$1B	~\$9B	\$65B（7月底）	IPO 前投资者披露7x增长
智谱 Z.ai	总收入	¥3.1亿*	¥7.24亿	¥18.1亿	半年已超2025全年2.5x
智谱 Z.ai	MaaS API ARR	¥0.28亿	¥17亿	>\$10亿（整体ARR）	未披露
MiniMax	总收入	\$30.5M	\$79.0M	~\$115M	半年已超2025全年
Gemini	订阅收入	\$0.3B	\$1.2B	未单列	Cloud Q2'26 \$24.8B, +82% YoY

模型调用量及用户仍然有指数级增长

公司	核心调用量指标	增长轨迹（近期）	关键增速
OpenAI	API 吞吐	~2亿→150亿 token/分（2024末→2026初）	~75x
OpenAI	ChatGPT 用户	WAU 300M(24末)→900M+(26初)	翻倍/年
Anthropic	API 用量	月 API 调用 250亿+；token 量 3–5x YoY	+300% YoY
Anthropic	Claude.ai MAU	59M(Q1'25)→220M(Q1'26)	3.7x
Anthropic	Claude Code	人均 1,549 API calls/月（4月）	编程调用 QoQ +300%
智谱 Z.ai	Token 调用量	6个月 +15x	涨价83%后调用量仍 +400%
MiniMax	M2 系列日均 Token	2025/12→2026/2 +6x；Coding +10x	低基数爆发
Gemini	API 吞吐	7→10→16→22 B token/分（25/9→26/6）	环比 +60%→+37.5%（放缓）
Gemini	AI Mode	MAU 突破 1B+	DAU 同比 3x

行业盈利模式与盈利水平



- **模式选择：**轻资产平台以差价与订阅为主；运营商倾向交易分成
- **企业采购：**大型企业年采购从数万元到数十万元不等

业务形态演进：从灰色通道到企业级智能中枢

1.0 灰色通道

代理翻墙 · 共享账号

运营主体

个人或小组团队运营

痛点

稳定性、安全性无保障

2.0 聚合分销商

低买高卖 · 平台化

平台

统一API+计费+密钥管理

模式

标准化服务赚差价

3.0 企业级智能中枢

增值服务 · B端深耕

核心能力

智能调度 · 成本优化 · 数据分析

客群

转向B端企业客户

驱动演进：监管收紧使灰色通道出清；企业级需求升级；增值服务构建护城河



Token中转站，这个诞生于AI大模型时代需求的新兴产业，在经历了短暂的野蛮生长后，正迅速步入一个以“合规”为基石、以“价值”为导向的全新发展阶段。它已从一个简单的“代理通道”，进化为承载多模型调度、成本优化、安全管控和开发者生态的关键基础设施。当前，该产业正处在一个快速增长与转型并存的阶段。

未来发展趋势展望

展望未来，我们认为Token中转站产业有以下几个发展趋势：

趋势一：全面合规化与品牌化

未来只有那些拥有完备合规资质、能够提供透明、可信服务的品牌，才能在市场中立足。用户在选择服务商时，会把合规性作为首要考量因素。行业的竞争将从价格战转向品牌、信誉和安全能力的综合比拼。

趋势二：价值链延伸与服务深化

单纯的API通道代理将不再具备竞争力。领先的平台会沿着价值链向上下游延伸，提供更深度的服务，包括模型微调与管理服务、Agent与 workflow编排、垂直行业解决方案。

趋势三：技术驱动的极致智能化调度

随着模型数量的增长和性能的日益细分，如何匹配最优模型组合，将成为一个复杂问题。未来的中转站将利用AI技术来管理AI调用，兼顾成本效益、性能指标、模型能力。

趋势四：市场集中度提升与生态整合

高昂的合规成本和技术门槛将不可避免地导致市场集中度的提升。少数头部平台将占据大部分市场份额。可能会出现大型云厂商收购或战略投资领先的API聚合平台，将其能力深度集成到自身的云原生AI服务中，形成“算力-平台-模型-应用”一体化的超级生态。

硅基流动（未上市，已交表）公司推出一站式大模型API云服务平台SiliconCloud，提供大模型推理服务，被业内称为“Token工厂”。

中科闻歌（1956.HK）正式发布企业级AI基础设施平台海汇TokSea。该平台以词元为基础单元，整合模型接入、专业模型调用、智能路由、权限管理、成本监控与运营优化等能力，帮助企业建立可管理、可度量、可优化的AI使用体系。

猎豹移动（CMCM.N）公司推出EasyRouter，通过统一接口调用海外GPT、Claude、Gemini、DeepSeek等40余个主流大模型。

Palo Alto Networks（PANW.O）通过收购Portkey切入中转赛道，提供企业级AI网关平台，主打可观测性与治理能力。

网易有道（DAO.N）正式发布企业级大模型聚合平台ThinkFlow，通过标准化API接口将DeepSeek、Kimi、Qwen等20余款主流大模型纳入统一调度。

七牛智能（2567.HK）提供DeepSeek、GLM、MiniMax、Kimi等主流大模型接入服务，兼容Codex、Claude Code、Gemini CLI等AI Coding工具。

➤ 监管与合规风险

行业监管政策持续收紧，不合规参与者面临出清；内容安全与数据合规体系建设成本高企，影响行业盈利预期。

➤ AI概念股票可能波动较大

当市场调整当中，由于缺乏历史估值对标，AI概念股票波动可能很大。

➤ 行业竞争格局恶化

行业进入门槛较低、参与者众多，价格竞争激烈，需要企业持续的资源投入。

➤ 上游依赖风险

中转站不掌握大模型核心能力，上游模型厂商定价与渠道政策变化，可能直接冲击现有商业模式。

免责声明

一般声明

本报告由国元证券经纪（香港）有限公司（简称“国元证券经纪（香港）”）制作，国元证券经纪（香港）为国元国际控股有限公司的全资子公司。本报告中的信息均来源于我们认为可靠的已公开资料，但国元证券经纪（香港）及其关联机构对这些信息的准确性及完整性不作任何保证。本报告中的信息、意见等均仅供投资者参考之用，不构成对买卖任何证券或其他金融工具的出价或征价或提供任何投资决策建议的服务。该等信息、意见并未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐或投资操作性建议。投资者应当对本报告中的信息和意见进行独立评估，自主审慎做出决策并自行承担风险。投资者在依据本报告涉及的内容进行任何决策前，应同时考虑各自的投资目的、财务状况和特定需求，并就相关决策咨询专业顾问的意见对依据或者使用本报告所造成的一切后果，国元证券经纪（香港）及/或其关联人员均不承担任何责任。

本报告署名分析师与本报告中提及公司无财务权益关系。本报告所载的意见、评估及预测仅为本报告出具日的观点和判断。该等意见、评估及预测无需通知即可随时更改。在不同时期，国元证券经纪（香港）可能会发出与本报告所载意见、评估及预测不一致的研究报告。

本报告署名分析师可能会不时与国元证券经纪（香港）的客户、销售交易人员、其他业务人员或在本报告中针对可能对本报告所涉及的标的证券或其他金融工具的市场价格产生短期影响的催化剂或事件进行交易策略的讨论。这种短期影响的分析可能与分析师已发布的关于相关证券或其他金融工具的目标价、评级、估值、预测等观点相反或不一致，相关的交易策略不同于且也不影响分析师关于其所研究标的证券或其他金融工具的基本面评级或评分。

国元证券经纪（香港）的销售人员、交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。国元证券经纪（香港）没有将此意见及建议向报告所有接收者进行更新的义务。国元国际控股有限公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见不一致的投资决策。

除非另行说明，本报告中所引用的关于业绩的数据代表过往表现。过往的业绩表现亦不应作为日后回报的预示。我们不承诺也不保证，任何所预示的回报会得以实现。

分析中所做的预测可能是基于相应的假设。任何假设的变化可能会显著地影响所预测的回报。

本报告提供给某接收人是基于该接收人被认为有能力独立评估投资风险并就投资决策能行使独立判断。投资的独立判断是指，投资决策是投资者自身基于对潜在投资的目标、需求、机会、风险、市场因素及其他投资考虑而独立做出的。

特别声明

在法律许可的情况下，国元证券经纪（香港）可能与本报告中提及公司正在建立或争取建立业务关系或服务关系。因此，投资者应当考虑到国元证券经纪（香港）及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突。

本报告的版权仅为国元证券经纪（香港）所有，未经书面许可任何机构和个人不得以任何形式转发、翻版、复制、刊登、发表或引用。

分析员声明

本人具备香港证监会授予的第四类牌照——就证券提供意见。本人以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

国元国际控股有限公司

香港中环康乐广场8号交易广场三期17楼

电 话：(852) 3769 6888

传 真：(852) 3769 6999

服务热线：400-888-1313

公司网址：<http://www.gyzq.com.hk>