Token 消费持续爆发，Token 运营迎来发展拐点

Token 运营深度研究

多模型时代，Token 运营有望成为连接模型供给与 AI 应用的重要基础设施。随着基础模型数量持续增加、不同模型在能力、成本、速度及稳定性上不断分化，下游开发者同时调用多模型及多 Provider 逐渐成为常态。Token 运营平台通过统一 API 聚合模型，并进一步提供模型选择、智能路由、Token 计量、统一结算及企业治理等服务，其角色正由简单的模型接入工具向 AI 流量分发与企业调用控制层演进。我们认为，随着 Token 调用规模和使用场景持续扩张，Token 运营价值有望进一步提升。

- **行业空间：Token 调用高速增长，Token 运营市场仍处渗透早期。**Gartner 预计，全球 Token 运营市场支出将由 2026 年的约 2.51 亿美元增长至 2027 年的 4.29 亿美元，2028 年有望达到约 10 亿美元，对应三年 CAGR 接近 100%；当前仅约 15%的软件工程负责人已经部署相关平台，31%正在部署、22%仍处评估阶段，行业渗透率仍有较大提升空间。我们测算，2026 年全球经 Token 运营平台承载的模型调用 GMV 约 43-60 亿美元，约占全球生成式 AI 模型支出的 15%–21%。国内需求同样快速放量，2025 年我国全年 Token 调用量约 21100 万亿，2026 年 3 月日均调用量进一步突破 140 万亿。
- **商业模式：轻资产特征明显，流量规模、路由数据与企业粘性构筑核心壁垒。**Token 运营商通常不直接承担基础模型训练及大规模 GPU 资本开支，目前主要形成手续费、订阅制及“订阅+Credits”等收费模式。相比统一 API 本身较低的技术门槛，平台真正的竞争优势来自规模化运营后形成的正向循环：模型及 Provider 越丰富，能够吸引更多开发者；Token 调用规模提升后，调用数据的持续积累又有助于优化模型选择和路由效率，并通过预算、权限、日志、安全及治理等企业功能进一步提高客户迁移成本。
- **海外案例：OpenRouter 由模型聚合器快速成长为 AI 流量分发入口。**OpenRouter 通过统一 API 连接不同模型及 Provider，并逐步建立模型路由与 Provider 路由双层体系。截至 2026 年 9 月，平台已接入 500+活跃模型、80+ Provider，服务超过 1000 万开发者和企业，每日处理 Token 超过 10 万亿。平台采用 Credits 统一结算，并主要收取 5.5%的平台服务费；随着越来越多用户将模型和 Provider 选择交由平台完成，OpenRouter 掌握的不再只是模型接入入口，而是参与 Token 流量在不同模型与推理服务商之间的实际分配。
- **国内布局：由模型接入向 Token 供给及应用场景延伸。**腾讯、硅基流动重点布局模型聚合、推理及 Token 供给；易点天下、迈富时、吉宏股份等从营销、企业服务及跨境电商场景切入，将模型调用、Token 计量和 Agent 调度嵌入业务流程。荣信文化依托内容及 IP 资源探索垂类知识资产的 Token 化调用，范式智能则从企业 AI 基础设施向模型服务、API 调用及 Agent 平台延伸。**我们认为，国内 Token 运营已由单一模型接入逐步向“基础设施+企业应用+垂类内容”多路径发展，随着 Agent 和企业 AI 应用深入业务流程，Token 计量、调度及场景化变现能力的重要性有望进一步提升。**
- **风险提示：AI 应用及 Token 调用增长不及预期；多模型调用及智能路由需求不及预期；行业竞争加剧；商业化进展不及预期；数据安全与合规风险。**

推荐（维持）

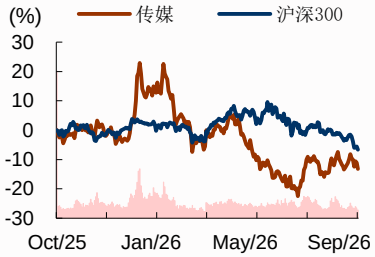
TMT 及中小盘/传媒

行业规模

		占比%
股票家数（只）	159	3.0
总市值（十亿元）	1568.1	1.4
流通市值（十亿元）	1441.1	1.5

行业指数

%	1m	6m	12m
绝对表现	-4.9	-8.1	-14.5
相对表现	0.5	-1.9	-7.4



资料来源：公司数据、招商证券

相关报告

1、《AI 应用研究感思录（二十六）—META 引领 AGENT 普及，国内 AGENT 业务有望跟上》2026-09-27

2、《AI 应用研究感思录（二十五）—AI 影视产业链走访反馈：行业正在积极迎接新变革》2026-09-20

3、《AI 应用研究感思录（二十四）—AI 影视持续火爆，IP 公司有望显著受益》2026-09-12

顾佳 S1090513030002
gujia@cmschina.com.cn
谢笑妍 S1090519030003
xiexiaoyan1@cmschina.com.cn

仅供内部参考，请勿外传！

正文目录

一、行业分析：AI 产业链的“路由层”.....4
1、多模型供给走向繁荣，Token 运营层应运而生.....4
2、Token 调用加速放量，Token 运营市场步入高速增长期.....5
二、商业模式与竞争壁垒：流量、数据与企业粘性构筑护城河.....6
1、轻资产运营，多元收费模式打开商业化空间.....6
2、技术壁垒有限，规模、数据与企业粘性构筑竞争优势.....7
三、OpenRouter：全球模型路由平台的商业化样本.....9
1、OpenRouter：从模型聚合器快速跃迁至全球 AI 流量入口.....9
2、OpenRouter 的核心资产：模型选择权正在成为新的流量分配权.....10
四、中国市场：Token 运营从基础设施向应用场景延伸.....11
1、腾讯：TokenHub 构建统一模型入口，云厂商加速布局 Token 运营.....11
2、易点天下：依托营销流量入口，构建 Token 服务体系.....12
3、硅基流动：推理基础设施型 Token 供应商，打造规模化“Token 工厂”.....13
4、荣信文化：内容资产 Token 化，探索“内容+AI”商业化新模式.....14
5、迈富时：由 AI 应用平台向“全栈 Token 工厂”延伸.....15
6、吉宏股份：Giikin AI+ FDE 智能创业平台开启 AI 赋能商业化探索.....16
7、范式智能：从企业 AI 平台向 Token 基础设施延伸.....17
五、风险提示.....18

图表目录

图 1：Token 运营产业链结构图.....4
图 2：OpenRouter 分场景模型调用偏好热力图.....5
图 3：全球 Token 运营市场规模及增速（亿美元）.....5
图 4：企业 Token 运营平台部署阶段分布.....5
图 5：全球主要 Token 运营平台月度模型调用规模对比.....6
图 6：2025 年中国 MaaS 市场月度日均 Token 消耗（亿）.....6
图 7：OpenRouter 前 10 Provider 月度 Token 流量.....8
图 8：OpenRouter Provider 月度 Token 流量集中度.....8
图 9：OpenRouter Auto Router 升级前后成本—性能变化.....8

仅供内部参考，请勿外传！

图 10：OpenRouter 发展历程及战略重心 9
图 11：OpenRouter 通过统一 API 实现多模型接入 10
图 12：OpenRouter 的产品体系 10
图 13：模型供给覆盖主流模型厂商 11
图 14：易点天下 AI Drive 2.0 数智营销体系 12
图 15：AI 应用持续放量，运营效率显著提升 13
图 16：TokenLX 多模型与算力智能调度平台 13
图 17：硅基流动打造“Token 工厂” 13
图 18：硅基流动平台已接入多款主流大模型 13
图 19：荣信智趣&火山引擎 AI 行业解决方案产品架构 14
图 20：迈富时企业级 Agent 平台能力架构 15
图 21：迈富时 AI-Agentforce 智能体场景矩阵 15
图 22：ClawForce 支持多类企业业务场景落地 15
图 23：Giikin AI+ FDE 智能创业平台架构 16
图 24：范式提供“一站式 AIOS 解决方案” 17

表 1：Token 运营主要商业模式及收费模式 7
表 2：OpenRouter 融资及战略交易情况 9
表 3：TokenHub 核心功能 11



一、行业分析：AI 产业链的“路由层”

1、多模型供给走向繁荣，Token 运营层应运而生

多模型长期共存，推动 Token 运营成为连接模型供给与 AI 应用的重要中间层。大模型发展初期，模型数量有限，开发者通常直接向单一模型厂商采购并调用 Token；随着模型供给快速丰富、不同模型在能力与价格上的差异扩大，开发者逐一寻找、比较和接入模型的成本明显提升，第三方 Token 运营平台由此应运而生。从产业链来看，上游为 GPU、云计算及基础模型/推理服务商，中游 Token 运营平台汇集不同模型供给，下游连接 AI Coding、AI 办公、营销、游戏等应用及 Agent，并最终触达企业和个人用户。其商业逻辑与携程等 OTA 平台具有一定相似性：OTA 通过聚合分散的酒店和机票供给，为消费者提供统一的搜索、比较和交易入口；Token 运营则通过聚合分散的模型供给，为用户提供统一的模型选择和调用入口，降低寻找、比价和切换模型的成本，因此下游用户是平台化过程中最直接的受益方。

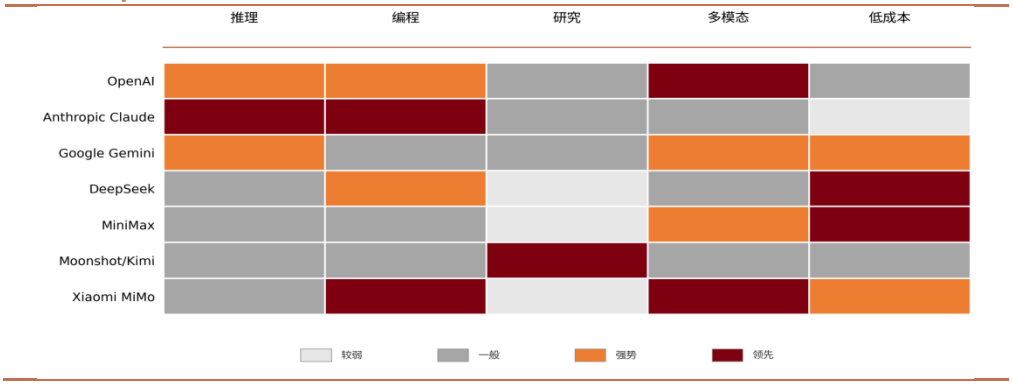
图 1：Token 运营产业链结构图



资料来源：OpenRouter，Vercel，Portkey，招商证券

随着模型数量增加，Token 运营的核心价值正由“连接模型”向“帮助用户选择模型”升级。当前不同模型在代码、研究、翻译、金融等任务中的表现和价格仍存在明显差异，且模型更新和价格调整频繁，用户长期绑定单一模型的成本不断提升。根据 OpenRouter 披露，截至 2026 年 9 月，其平台已接入 500 余个模型、80 余家推理服务商，覆盖 OpenAI、Anthropic、Google、DeepSeek、腾讯、智谱、MiniMax 等主要厂商；从平台分场景调用数据看，不同任务对应的主流模型选择亦明显不同，代码生成、研究、翻译、金融等场景并不存在一个长期占优的单一模型。由此，Token 运营平台的核心价值逐渐从“多接几个模型”转向帮助用户在大量供给中完成选择：一方面，通过统一入口减少重复接入和管理成本；另一方面，根据不同任务在效果、价格和稳定性之间进行匹配，并在模型服务异常时切换至其他可用供给。随着模型数量继续增加、Token 消费持续爆发，用户的选择成本和管理复杂度将进一步提升，Token 运营平台作为模型消费入口的价值有望同步放大。

图 2：OpenRouter 分场景模型调用偏好热力图

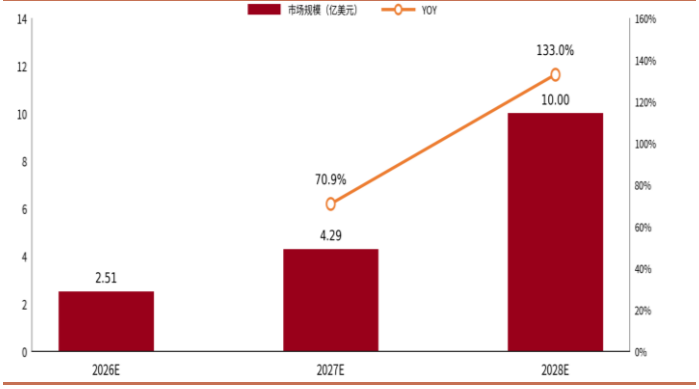


资料来源：OpenRouter，招商证券

2、Token 调用加速放量，Token 运营市场步入高速增长期

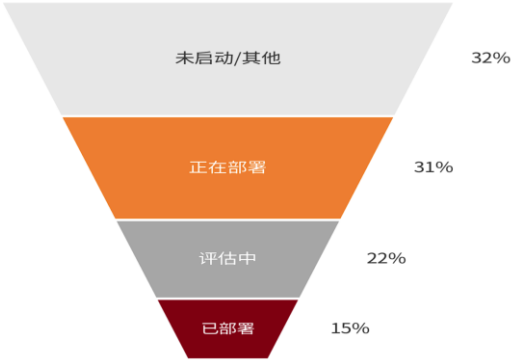
Token 运营已由早期开发工具逐渐形成独立的基础设施市场，2026 年全球市场收入约 2.5 亿美元。据 Gartner 预测，2026 年全球 Token 运营市场支出约 2.51 亿美元，2027 年将进一步增长 70.9%至 4.29 亿美元，预计到 2028 年市场规模达到约 10 亿美元，对应三年 CAGR 接近 100%。当前，Token 运营业务的企业渗透率仍处早期阶段，据 Gartner 调查仅 15%的软件工程负责人已经部署了类似 OpenRouter、Vercel 的 token 运营平台，另有 31%正在部署、22%处于评估阶段。伴随多模型调用及 Agent 应用持续普及，行业仍具备较大的渗透率提升空间。

图 3：全球 Token 运营市场规模及增速（亿美元）



资料来源：Gartner，招商证券

图 4：企业 Token 运营平台部署阶段分布

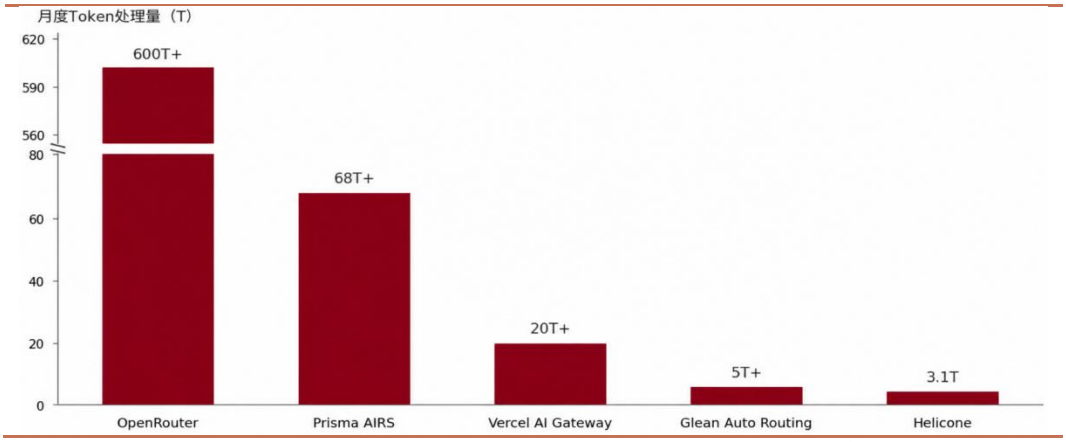


资料来源：Gartner，招商证券

从模型调用交易规模来看，我们测算 2026 年全球经 Token 运营平台承载的模型调用 GMV 约 43-60 亿美元。OpenRouter 目前月处理 Token 已超过 600 万亿；根据平台 Top 10 模型实际 Token 结构及当前 API 价格测算，其平均模型调用成本约 0.5–0.7 美元/M Token，对应 OpenRouter 单平台年化模型调用 GMV 约 36–50 亿美元，而 Prisma AIRS、Vercel AI Gateway 等其他几大主流 Token 运营平台公开披露的月度 Token 处理量合计约相当于 OpenRouter 的四分之一。假设各平台单位 Token 调用成本一致，并综合考虑市场还存在其他未披露规模的小型平台，我们预估 2026 年全球 Token 运营平台 GMV 区间为 43–60 亿美元。结合 Gartner 预计 2026 年全球基础及专业 GenAI 模型支出约 283 亿美元，我们估算当前约 15%–21%的模型消费已经经由第三方 Token 运营平台完成。

仅供内部参考，请勿外传！

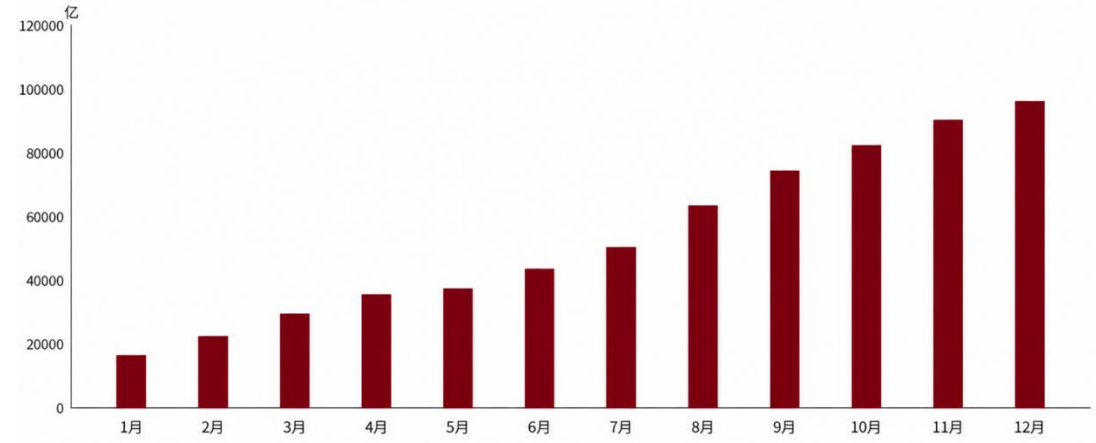
图 5：全球主要 Token 运营平台月度模型调用规模对比



资料来源：OpenRouter，Palo Alto Networks，Vercel，Glean，Helicone，招商证券

国内 Token 消费同样进入快速放量期，为 Token 运营提供持续扩张的底层交易池。国家数据局数据显示，我国 2025 年全年 Token 调用量约 21100 万亿，2026 年 3 月日均调用量进一步突破 140 万亿，较 2024 年初增长超过千倍；IDC 数据显示 2025 年以来中国企业级 MaaS 市场月 Token 调用量一路高涨，全年总调用量 1944 万亿，同比增长约 16 倍，并预计 2026 年全年 Token 消耗量将达到 40000 万亿次。此外，腾讯旗下的 WorkBuddy 等 Agent 产品采用“订阅+Credits”的 Token 商业化模式，其 7 月单月收入已达到亿元规模。我们认为，伴随国内 Token 消费量持续增长以及 Agent 对多模型调用、成本控制和统一结算需求提升，当前约 186 亿元的 MaaS 市场将为 Token 运营层提供持续扩大的可服务市场空间。

图 6：2025 年中国 MaaS 市场月度日均 Token 消耗（亿）



资料来源：IDC 中国，招商证券

二、商业模式与竞争壁垒：流量、数据与企业粘性构筑护城河

1、轻资产运营，多元收费模式打开商业化空间

Token 运营平台本质上是连接模型供给与 AI 应用需求的中间层，通过聚合不同模型及推理服务商，为开发者和企业提供统一的模型调用入口。相比基础模型厂商需要承担 GPU 采购、模型训练和大规模推理集群建设，Token 运营商通常不直接参与模型研发，主要通过平台服务获取收入，商业模式具备较强的轻资产特

仅供内部参考，请勿外传！

征。其主要成本包括：1）在模型聚合模式下，需要向上游模型及推理服务商支付 Token 调用费用，并承担一定的预充值及资金占用；2）需要投入平台基础设施建设，包括模型接入、调用管理、数据存储及交易结算等运营成本；3）随着企业客户需求提升，还需增加模型调度、安全治理、客户服务及商业拓展等投入。

根据收费方式不同，目前行业商业模式可进一步划分为以下三类：手续费模式、订阅制模式以及订阅+Credits 混合模式，分别对应 Token 交易抽佣、企业软件订阅以及软件服务与 Token 消费结合的变现方式。

表 1：Token 运营主要商业模式及收费模式

商业模式	代表平台	收费模式	Take Rate
手续费模式	OpenRouter	按模型调用金额收取平台服务费	约 5%–8%
订阅制模式	Portkey	企业客户支付固定软件订阅费	约 5%-10%
订阅+Credits 混合模式	WorkBuddy	订阅+Token 额度+超额收费	约 5%-10%

资料来源：OpenRouter，Portkey，WorkBuddy，招商证券

1）手续费模式：纯手续费模式直接参与 Token 交易，并按照客户模型消费金额收取固定比例平台服务费，收入随平台承载的模型调用规模同步增长。OpenRouter 是目前最典型的代表，其并不对上游 Provider 公布的 Token 价格额外加价，而是在用户购买平台 Credits 时收取 5.5%的平台费；对于使用自身模型厂商 API Key 的 BYOK 客户，在超过一定免费额度后按照对应模型调用成本收取 5%的平台费。该模式收入逻辑较为直接，即平台收入≈经平台完成的模型调用 GMV×有效 Take Rate。该模式的核心在于尽可能扩大平台 Token 流量和交易规模，并通过路由效率、模型覆盖和稳定性提高用户留存。

2）订阅制模式：订阅制平台不依赖 Token 价差或交易抽佣，而是将统一 API、路由、监控及企业治理能力作为软件服务销售，收入与客户数量、请求量及功能等级挂钩。以 Portkey 为例，平台通过提供统一模型接入、调用监控、自动切换及企业治理等能力，向客户收取固定订阅费用，例如其 Production 版本定价 49 美元/月。该模式下，客户向模型厂商支付 Token 费用，向 Token 运营平台支付软件订阅及管理费用，两者相互独立。因此平台收入不直接受模型 Token 价格下降影响，商业模式更接近传统企业 SaaS。收入稳定性较高，也可通过企业治理、安全合规等高级功能提升单客户价值。

3）订阅+Credits 混合模式：混合模式将软件订阅与 Token 消费打包，通过固定订阅费提供基础 Credits，并对超额 Token 需求进一步收费，在提高收入可预测性的同时保留 Token 用量增长弹性。腾讯 WorkBuddy 是这一模式的典型代表。其个人 SaaS 标准版目前定价 70 元/月，包含 4000 Credits，高级版 140 元/月，包含 9000 Credits，当套餐内 Credits 耗尽后，付费会员还可按 50 元/1000 Credits 购买加量包，Credits 实际消耗则取决于所调用的模型、Token 数量及任务复杂度。与 OpenRouter 主要承担模型路由和分发不同，WorkBuddy 进一步将 Token 能力封装进具体 Agent 产品，因此平台不仅获得模型调用层价值，还能够获取应用层订阅收入。

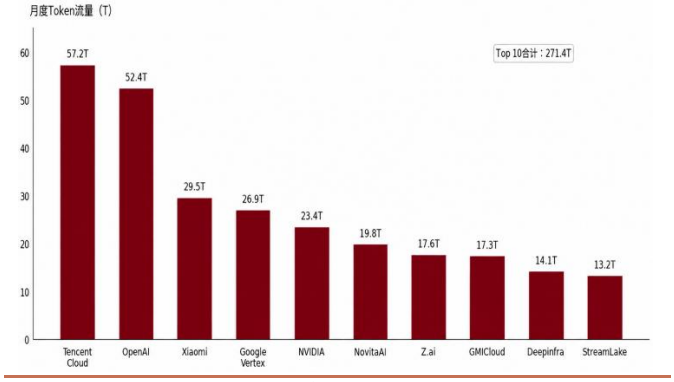
2、技术壁垒有限，规模、数据与企业粘性构筑竞争优势

统一 API 本身技术门槛有限，真正的竞争壁垒来自规模化运营后形成的正向循

仅供内部参考，请勿外传！

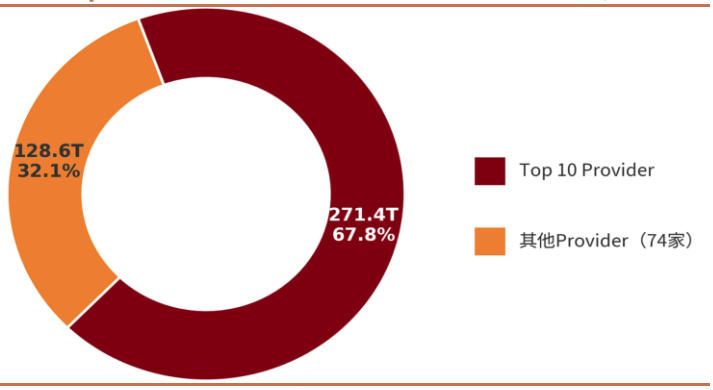
环。对下游开发者而言，平台接入的模型和 Provider 越丰富，一次 API 接入能够覆盖的任务场景越多，使用价值越高；而用户和 Token 调用量增加后，又能够吸引更多模型厂商和推理服务商主动进入平台获取流量，形成“更多模型—更多用户—更多调用—更多模型”的正向循环。目前 OpenRouter 已接入 80 余家 Provider 并服务超过 1000 万开发者，其路由机制会综合价格、延迟、吞吐量及稳定性分配请求，表现更好的 Provider 能够获得更多流量。平台因此由单纯连接供需双方，逐步获得对模型流量的实际分配能力。

图 7：OpenRouter 前 10 Provider 月度 Token 流量



资料来源：OpenRouter，招商证券

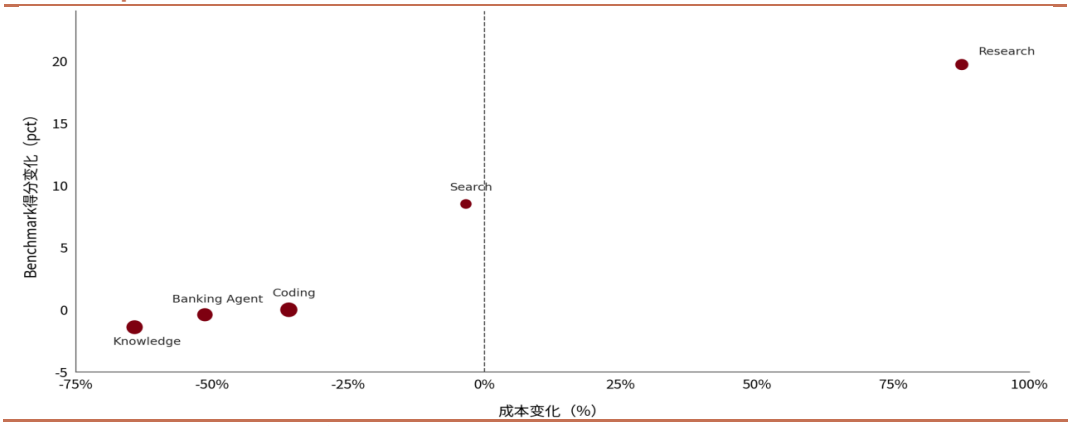
图 8：OpenRouter Provider 月度 Token 流量集中度



资料来源：OpenRouter，招商证券

调用规模进一步沉淀真实业务数据，推动模型选择能力持续优化。在业务早期，Token 运营平台主要依靠公开 Benchmark 进行模型推荐。随着调用规模扩大，平台逐步积累真实业务场景下的模型表现和成本数据，并通过数据反馈持续优化模型选择效率。以 OpenRouter 为例，公司于 2026 年 8 月升级 Auto Router 系统，通过利用大规模真实调用数据优化模型选择。根据 OpenRouter 官方披露，Auto Router 可利用每周超过 55 万亿 Token 对应的真实使用数据进行模型推荐。随着平台调用规模扩大，更多真实业务数据将进一步提升模型选择准确性，形成“更多调用—更多数据—更优匹配—更高用户价值—更多调用”的正向循环，并逐步拉开与简单 API 聚合工具之间的差距。

图 9：OpenRouter Auto Router 升级前后成本—性能变化



资料来源：OpenRouter，招商证券

企业 AI 应用规模化后，平台进一步成为企业模型调用与治理入口，客户迁移成本随之提升。随着企业同时使用更多模型和 Provider，Token 运营平台不仅承担模型接入，还逐步承接预算管理、权限控制、调用监控、安全治理及成本分析等功能。企业一旦围绕平台形成模型体系、调用流程和治理机制，更换平台需要重新配置相关系统，客户粘性随业务复杂度提升而增强。因此，Token 运营行业的竞争壁垒并不主要来自单一技术能力，而是由**规模效应、真实调用数据和企业**

治理能力共同构成。

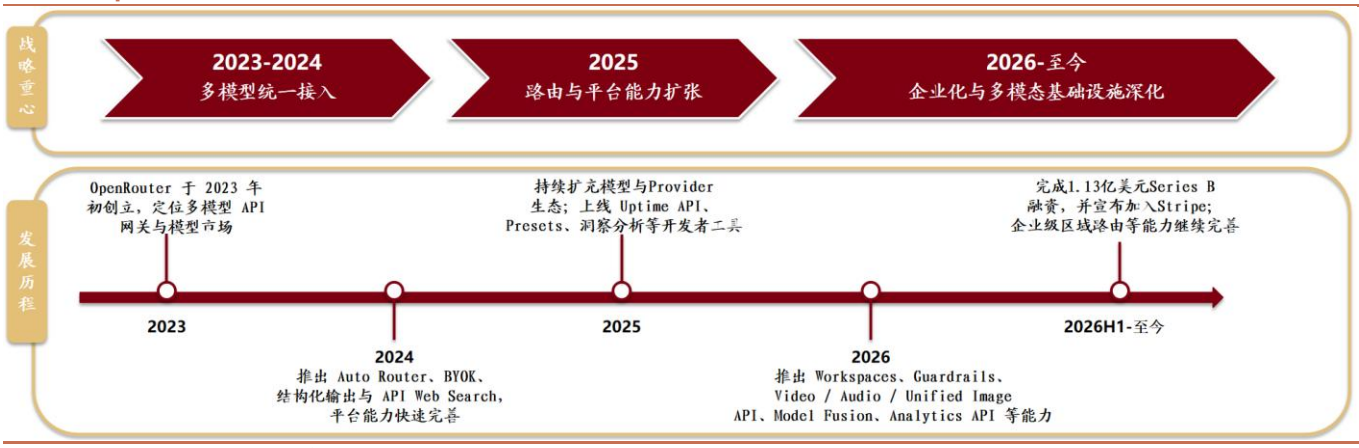
三、OpenRouter：全球模型路由平台的商业化样本

1、OpenRouter：从模型聚合器跃迁至全球 AI 流量入口

由模型聚合走向多模型基础设施，平台能力持续扩展。OpenRouter 于 2023 年创立，基于“人工智能将长期呈现多模型共存”的判断切入，定位为大模型 API 聚合与路由平台。公司通过统一接口连接主流模型与多家推理服务商，为开发者和企业提供模型发现、调用、比价、故障切换、成本管理与可观测性等能力。伴随模型供给持续丰富、调用复杂度提升以及企业级需求增长，OpenRouter 持续扩展面向生产场景的平台能力，目前已形成统一 API 接入、企业级控制与智能路由三大能力体系，并逐步构建面向多模型时代的 AI 基础设施平台。

调用规模快速放大，平台已进入规模化发展阶段。自成立以来，OpenRouter 平台推理量每年至少增长 10 倍；根据公司 2026 年 9 月披露，OpenRouter 已接入 500+活跃模型及 80+ Provider，服务超 1000 万开发者和企业，每日处理 10 万亿+ Token。2026 年 5 月，公司完成 1.13 亿美元 Series B 融资；同年 8 月，公司宣布加入 Stripe。随着模型数量、Provider 供给及下游开发者持续扩张，OpenRouter 已由早期的多模型接入工具逐步成长为连接模型供给、推理资源与下游 AI 应用的重要中间层。

图 10：OpenRouter 发展历程及战略重心



资料来源：OpenRouter、招商证券

回顾发展历程，OpenRouter 自成立以来已先后完成 Seed、Series A 及 Series B 三轮机构融资，资本实力与平台规模同步提升。OpenRouter 于 2025 年连续完成 Seed 及 Series A 两轮融资，其中 2 月完成 1250 万美元 Seed 轮融资，由 Andreessen Horowitz（a16z）领投；4 月完成 2800 万美元 Series A 融资，由 Menlo Ventures 领投，两轮合计融资约 4000 万美元，彼时公司估值约 5 亿美元。2026 年 5 月，公司进一步完成 1.13 亿美元 Series B 融资，由 Alphabet 旗下 CapitalG 领投，NVentures、ServiceNow Ventures、MongoDB Ventures、Snowflake Ventures、Databricks Ventures 等产业资本参与，投后估值约 13 亿美元。此后 2026 年 8 月 OpenRouter 宣布加入 Stripe，资本化进程持续推进。

表 2：OpenRouter 融资及战略交易情况

仅供内部参考，请勿外传！

时间	事件	金额	主要投资方/交易方	估值
2025 年 2 月	Seed 轮	1,250 万美元	a16z 领投	早期机构融资
2025 年 4 月	Series A	2,800 万美元	Menlo Ventures 领投	Seed+Series A 约 5 亿美元
2026 年 5 月	Series B	1.13 亿美元	CapitalG 领投；NVentures、ServiceNow Ventures 等参与	投后估值约 13 亿美元
2026 年 8 月	Stripe 收购	交易对价大于 70 亿美金	Stripe	-

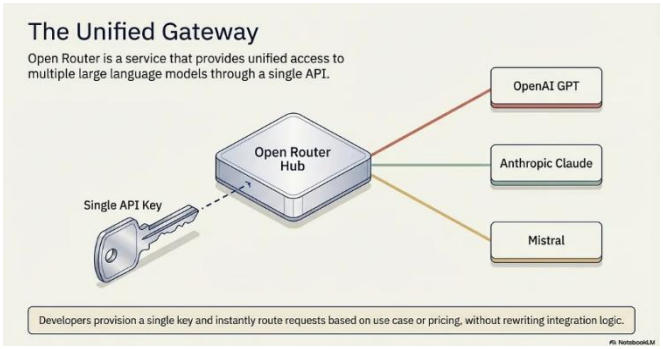
资料来源：OpenRouter，Bloomberg，Reuters，招商证券

2、OpenRouter 的核心资产：模型选择权正在成为新的流量分配权

由模型聚合切入，OpenRouter 正从统一接入平台向 AI 流量分发入口演进。平台早期主要解决多模型时代的“接入”问题，通过统一 API 聚合不同模型和 Provider，降低开发者重复接入、账户管理和计费结算成本。随着模型与 Provider 供给不断丰富，用户需求进一步从“能否调用”转向“调用哪个模型、由哪个 Provider 承接”。OpenRouter 由此将能力从统一接入延伸至模型路由和 Provider 路由，根据任务需求、价格、速度及稳定性等因素决定请求去向。当越来越多用户将模型和 Provider 选择交由平台完成时，OpenRouter 掌握的不再只是模型入口，而开始参与 Token 流量在不同模型及推理服务商之间的实际分配。

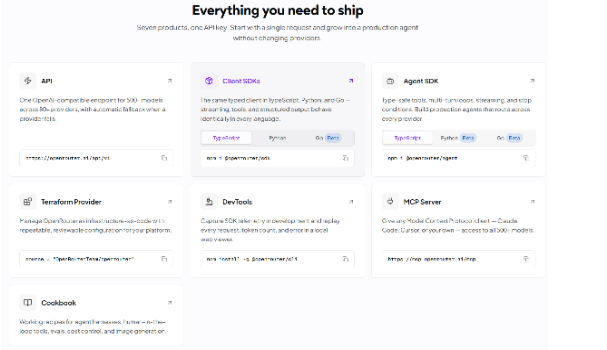
目前 OpenRouter 已形成以统一 API 为入口，模型路由能力为核心，并向企业管理及数据分析延伸的产品体系。1）接入层面，用户通过单一 API 即可访问不同模型及 Provider；**2）路由层面，**用户既可自主指定模型，也可通过 Auto Router 完成模型选择，而同一模型又可在多个 Provider 之间按照价格、吞吐量、延迟等维度进行排序，并在主要 Provider 不可用时自动切换至其他可用 Provider，必要时进一步切换至备用模型，从而提升调用稳定性。**3）面向企业客户，**平台进一步提供预算、权限、Guardrails、Activity 和 Analytics 等管理工具，将模型调用、成本控制和治理逐步纳入统一平台。由此，OpenRouter 的角色正由“模型调用工具”向“AI 请求分配和管理入口”演进。

图 11：OpenRouter 通过统一 API 实现多模型接入



资料来源：Medium，招商证券

图 12：OpenRouter 的产品体系



资料来源：OpenRouter，招商证券

商业模式清晰，Token 调用规模快速增长验证 AI 流量分发需求。OpenRouter 采用 Credits 体系完成平台内的统一支付与结算，用户购买 Credits 后即可调用不同模型及 Provider，平台在购买 Credits 时收取 5.5%的平台服务费，最低 0.8 美元；对于使用自有模型厂商 API Key 的 BYOK 用户，在超过一定免费额度后

仅供内部参考，请勿外传！

按照对应模型调用成本收取 5%的平台费。因此，OpenRouter 的收入主要由平台承载的模型调用规模及有效 Take Rate 决定，伴随 Token 流量扩大，平台服务费收入具备同步增长基础。2026 年 5 月，公司披露过去六个月周 Token 处理量由 5 万亿增长至 25 万亿，增长 5 倍；6 月初月度 Token 路由量进一步达到 100 万亿。目前 OpenRouter 已接入 80 余家 Provider 并服务超 1000 万开发者。

四、中国市场：Token 运营从基础设施向应用场景延伸

1、腾讯：TokenHub 构建统一模型入口，云厂商加速布局 Token 运营

TokenHub 聚合多模型供给，打造腾讯云统一大模型服务入口。2026 年 4 月，腾讯云正式上线大模型服务平台 TokenHub，并逐步将原腾讯混元、知识引擎等大模型服务入口统一迁移至新平台。TokenHub 在腾讯混元基础上，引入第三方模型，覆盖文本、图像、视频、语音及 3D 等多模态能力，并通过统一 API 降低不同模型之间的接入和迁移成本。TokenHub 提供模型体验、调用管理、用量统计及 API Key 管理等功能，并已适配 WorkBuddy、Claude Code、Cursor、Codex 等 Agent 和 AI Coding 工具，进一步连接下游应用场景。截至 2026 年 9 月，平台已上线 59 个模型，覆盖腾讯混元、DeepSeek、智谱、Kimi、MiniMax 等主流厂商，并保持较快的模型更新频率。

图 13：模型供给覆盖主流模型厂商



资料来源：TokenHub，招商证券

表 3：TokenHub 核心功能

核心功能	说明
模型广场	汇聚多家厂商的主力大模型，提供多维度筛选（类型、服务商、体验状态）、模型对比和详细信息查看，帮助快速选择合适的模型使用。
体验中心	在线体验大模型能力，支持文本对话（深度思考、流式输出），便于在接入 API 之前充分评估模型效果。
在线推理	创建和管理模型的推理服务实例，获取专属 API Endpoint。支持按需选择计费方式——按 Token 计费，满足从测试验证到生产部署的全阶段需求。
模型监控	展示模型、服务性能相关指标，如：TTFT（Time To First Token）、TPOT（Time Per Output Token）、RPM（Request Per Minute）。
用量统计	展示模型、服务性能计费用量指标，如：输入 Token、输出 Token、TPM（Tokens Per Minute）、插件调用次数。
API Key 管理	集中管理 API 访问密钥，支持精细化的权限控制（全部模型及服务或限定范围），便捷地启停切换和调用统计，保障接口调用的安全性与可控性。

资料来源：TokenHub 官网，招商证券

TokenHub 收费体系目前主要包括按量后付费和预付费订阅套餐（Token Plan）两类模式。按量后付费模式下，用户无需提前预留资源，根据实际 Token 消耗付费，其中语言模型按照输入、输出及缓存命中 Token 分别计价，不同模型对应不同单价，并按小时结算；该模式使用灵活，更适合调用量波动较大或处于业务起步阶段的用户。对于 Agent、AI Coding 等高频场景，则通过 Token Plan 提供预付费调用额度，并以统一积分体系对不同模型的 Token 消耗进行计量。

2、易点天下：依托营销流量入口，构建 Token 服务体系

从应用侧切入 Token 运营，易点天下将模型调用、Token 计量与路由分发嵌入具体营销业务流程。易点天下从真实营销场景切入，将模型调用、Token 计量与路由调度嵌入业务流程。公司底层通过 Cycor 聚合云厂商及大模型资源，由 AI Hub 统一承接模型调用、Token 计量和路由分发，再通过 AdClaw 将模型能力调度至不同 Agent 和 Skills，最终执行创意生产、广告投放、数据分析及运营优化等具体任务。公司 Token 运营的核心在于将多模型资源统一接入后，对 Token 消耗进行计量和路由，再通过 Agent 及工具链落地至实际营销任务。

图 14：易点天下 AI Drive 2.0 数智营销体系



资料来源：易点天下 2026 半年报，招商证券

Token 消耗快速增长，真实业务场景已形成规模化验证，对外商业化有望进一步打开增量。2026 年上半年，公司 AI 已覆盖约 50%的核心执行类工作，单运营人均效率提升 40%以上；2026Q2 日均 Token 消耗量较 Q1 环比增长超 3 倍，显示 AI 调用需求正在快速放量。具体业务层面，AI 生成素材占内部总创意产出 50%以上，AdsGo.ai 累计成功发布广告 1276 条；广告平台业务中，智能 DSP 核心客户日均广告消耗突破 10 万美元，电商场景算法转化率提升超 20%，智能运营体累计节省近 200 小时高价值人工耗时。

在内部业务持续验证模型调用与 Token 运营能力的基础上，公司进一步推出 AI 模型与算力智能调度平台 TokenLX，并面向出海企业及开发者开放内测。TokenLX 通过统一控制平面接入全球 190+主流大模型，并提供多模型智能路由与算力弹性调度，根据不同任务匹配模型和算力资源，同时提供数据本地化、数据主权及隐私保护等治理能力。我们认为，TokenLX 的推出标志着公司正在将自身营销业务中积累的模型调用、Token 管理及推理请求治理能力进一步产品化，由服务内部营销场景向外部企业及开发者的多模型调用与算力调度需求延伸。

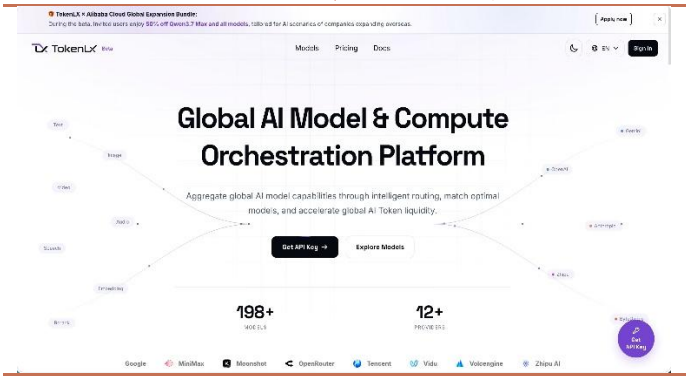
仅供内部参考，请勿外传！

图 15：AI 应用持续放量，运营效率显著提升



资料来源：易点天下 2026 半年报，招商证券

图 16：TokenLX 多模型与算力智能调度平台



资料来源：公司官网，招商证券

3、硅基流动：推理基础设施型 Token 供应商，打造规模化“Token 工厂”

硅基流动从 Token 供给侧切入，通过推理优化将算力和模型资源转化为标准化 Token 服务。公司定位于算力、模型与上层应用之间，不开发基础模型，也不直接布局终端应用，而是通过适配不同算力和主流大模型，提升推理效率，并以标准 API 向开发者和企业提供模型调用服务。其核心逻辑是将 GPU 算力进一步转化为稳定、低成本、可计量的 Token 产能，因此公司将自身定位概括为“Token 工厂”。

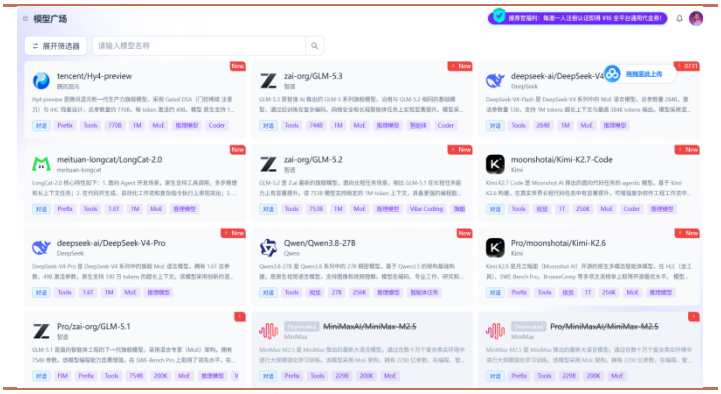
围绕不同客户的 Token 使用需求，公司已形成公有云 API、预留实例、私有化部署及算力联合运营等多种交付方式。公有云 API 面向调用需求灵活的开发者和企业；预留实例主要服务 Coding Agent 等高频、高稳定性场景；私有化部署满足金融、能源等对数据安全要求较高的客户；算力联合运营则帮助运营商、智算中心等提升已有 GPU 资源利用率。由此，公司提供的不只是模型 API，而是围绕 Token 生产、调度和交付的一整套推理基础设施。

图 17：硅基流动打造“Token 工厂”



资料来源：公司官网，招商证券

图 18：硅基流动平台已接入多款主流大模型



资料来源：公司官网，招商证券

用户及 Token 吞吐持续放量，中立 Token 供应模式已获得规模化需求验证。截至 2026 年，硅基流动平台已服务超过 1000 万名用户和 1.3 万家企业客户，日均 Token 吞吐量超过 1 万亿；根据 IDC 相关市场追踪，公司已进入中国公有云 MaaS 市场前四，也是第一梯队中唯一创业公司。企业端高频 Agent 场景进一步体现 Token 需求弹性。公司披露，部分 Coding Agent 客户 2026 年以来单日 Token 消耗已达到千亿量级；与贵州移动联合建设的智能服务平台在试运营期间日

仅供内部参考，请勿外传！

Token 消耗也已超过 50 亿。随着 Agent 及企业级 AI 应用从测试走向规模化生产，客户需求正由“能否调用模型”进一步转向“能否持续、稳定、低成本获得 Token”，推理效率和规模化 Token 供给能力有望成为公司新的核心竞争力。

4、荣信文化：内容资产 Token 化，探索“内容+AI”商业化新模式

深耕少儿内容二十余年，优质内容资产成为 AI 时代的重要资源基础。荣信文化长期聚焦少儿图书及儿童内容领域，在少儿科普、低幼启蒙等细分市场积累了较为丰富的内容及 IP 资源。2026 年以来，公司进一步推进“内容+IP+AI 应用”战略，将传统内容资产向 AI 可调用的数据资产转化，并与火山引擎开展合作，探索“内容+AI 的 Token 运营”模式。与传统模型聚合平台不同，公司切入 Token 运营的核心基础并非模型或算力，而是长期积累的垂类优质内容，通过 AI 技术重新释放内容资产价值。2026H1，公司实现营业收入 2.18 亿元，同比增长 20.81%，同期研发投入同比增长 84.24%，AI 剧/漫剧、智能硬件及 Token 运营等新业务持续推进。

将内容结构化、向量化形成垂类知识库，以“知识库+AI Token”实现内容调用。公司已对积累的少儿内容及 IP 进行结构化打标、向量化处理，将原本以图书、IP 等形式存在的内容转化为可供大模型检索和调用的知识库，并依托火山引擎 HiAgent 企业 AI 应用平台对外输出。具体来看，B 端可面向出版机构、教育平台及 AI 应用开发者提供内容调用服务；C 端则可应用于 AI 点读笔、AI 玩偶、APP 等产品，同时也可向第三方智能硬件提供内容能力。公司实现的并非简单采购模型 Token 后再销售，而是将通用模型 Token 与自身垂类知识资产结合，形成更具场景价值的内容 Token。

图 19：荣信智趣&火山引擎 AI 行业解决方案产品架构



资料来源：荣信文化公众号，招商证券

在商业化层面，公司提出以 Token 消耗量作为计价单位，可根据不同产品形态采用按调用收费、收入分成等模式，商业化推进计划分为产品整合及应用闭环、硬件应用与试点扩大、上下游生态接入三个阶段。我们认为，荣信文化提供了不同于 OpenRouter 的另一类 Token 运营路径：OpenRouter 主要解决模型及 Provider 之间的调用与流量分发，而荣信文化则依托自有内容资产构建垂类知识库，将通用模型能力与优质内容结合，并以 Token 消耗量作为服务计量基础。其核心在于将传统内容资产转化为可被 AI 持续检索、调用和商业化的数字化资

仅供内部参考，请勿外传！

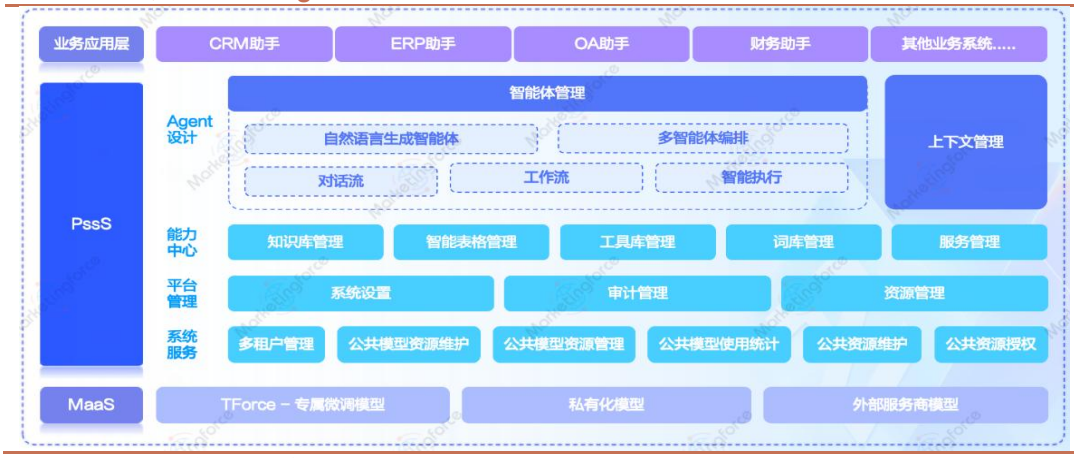
产，从而拓展内容 IP 在 AI 时代的变现路径。

5、迈富时：由 AI 应用平台向“全栈 Token 工厂”延伸

迈富时以企业级 Agent 和 AI 应用为主要交付载体，将模型能力嵌入企业实际业务流程。公司长期覆盖营销、销售、客服及客户运营等数字化场景，2026 年以来进一步提出“全栈 Token 工厂”战略，构建企业级 AI Agent 体系，产品已覆盖 GenAI OS、AI-Agentforce、KnowForce 等平台，并向营销、销售、客服、研发及经营分析等具体业务环节延伸。与单纯提供模型 API 不同，公司将底层模型能力、企业数据和业务流程结合，使 Token 调用直接服务于具体业务任务。

多模型融合与调度能力嵌入 Agent 底层，为不同业务任务动态匹配模型能力。迈富时将多模型管理与调度能力嵌入企业 Agent，根据任务复杂度、调用频率及业务需求选择和组合不同模型，并结合企业数据库、知识库及业务流程，支撑 Agent 完成营销、销售、客服、外贸、研发及经营分析等具体任务。公司将这一过程定义为将底层通用 Token 转化为能够解决具体业务问题、产生业务结果的“场景 Token”。2026 年 9 月，公司进一步升级企业级智能体操作系统 ClawForce，围绕统一 AI 应用入口、零代码创建智能体、Skill 沉淀、多智能体协作编排、能力分发及全链路监控等环节，强化企业对 AI 能力的统一管理，其中支持 Token 计数统计及智能体执行过程监控，进一步提升 Agent 规模化运行和管理能力。

图 20：迈富时企业级 Agent 平台能力架构



资料来源：公司官网，招商证券

图 21：迈富时 AI-Agentforce 智能体场景矩阵



资料来源：公司官网，招商证券

图 22：ClawForce 支持多类企业业务场景落地



资料来源：ClawForce 官网，招商证券

由软件订阅向基于使用量的 Token 经济模式延伸，AI 调用量与商业价值进一步挂钩。随着 AI Agent 逐步进入企业实际业务流程，其持续执行任务并产生 Token

仅供内部参考，请勿外传！

消耗，迈富时正探索由传统软件订阅模式向按使用量收费模式演进。公司部分 AI 原生产品已明确采用按智能体实际业务工作量消耗 Token 计费的方式，对于高复杂度、高专业要求的场景调用推理能力更强的模型，高频标准化任务调用轻量模型，以提升效率、降低使用成本。2026H1，公司 AI 应用付费用户达到 27269 家，同比增长 25.9%，月均 ARPU 同比增长 80.0%。我们认为，迈富时 Token 运营的核心在于以 Agent 作为交付载体，将模型调用、企业数据及业务场景结合，使 Token 消耗与具体业务任务的执行更加紧密地联系起来。

6、吉宏股份：Giikin AI+ FDE 智能创业平台开启 AI 赋能商业化探索

AI 能力由内部经营场景沉淀，推动跨境业务智能化升级。公司长期积累跨境电商业务数据和运营经验，并持续迭代 Giikin AI+平台。逐步将 AI 能力融入选品、素材生成、广告投放、客服、物流履约等核心环节，形成覆盖跨境经营全流程的智能化体系。随着 AI 能力逐步渗透业务流程，公司实现整体运营效率提升 80%以上，其中内容与素材制作效率提升 770%以上，广告投放 ROI 提升 90%以上，智能客服成本降低 85%以上。2026H1，公司跨境电商业务实现收入 24.51 亿元，同比增长约 15.82%，归母净利润 1.00 亿元，同比增长 80.92%，AI 驱动的运营效率提升已逐步体现在经营表现中。

依托 Giikin AI+能力积累，公司布局 Giikin AI+ FDE 智能创业平台战略，探索 Token 驱动的 AI 商业化模式。2026 年 9 月，公司发布“Giikin AI+FDE 智能创业平台”，围绕模型与算力底座、Agent 矩阵及 FDE 场景交付构建能力体系，将原本服务内部跨境业务的 AI 能力进一步面向企业和创业者开放。其中，TokGate.io 作为底层大模型与算力服务平台，连接不同模型资源，并通过 Token 形式提供模型调用服务；上层 Agent 矩阵则覆盖选品、营销、视频生成、客服、履约等跨境经营场景，使底层模型调用进一步转化为具体业务任务和经营服务。

图 23：Giikin AI+ FDE 智能创业平台架构



资料来源：吉宏股份公众号，招商证券

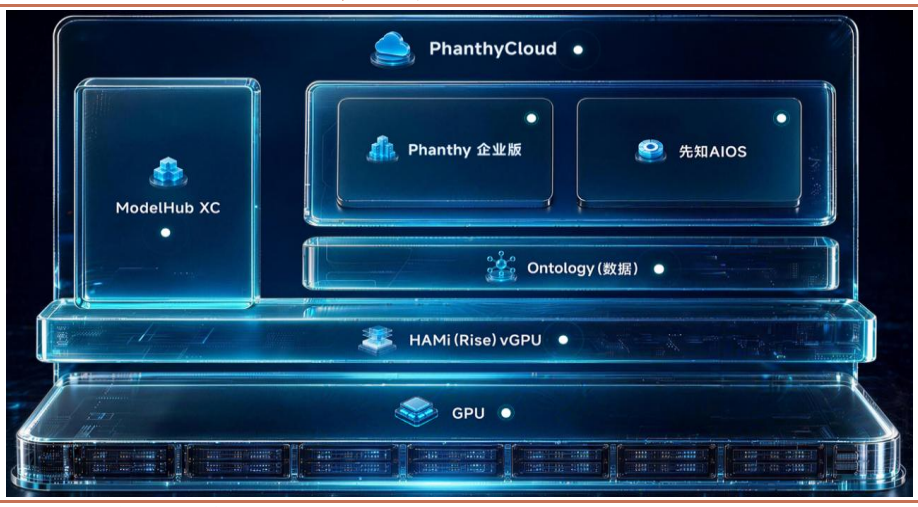
TokGate.io 作为公司 AI 对外商业化战略的第一步，率先探索 Token 运营模式落地。TokGate.io 通过统一账户、API 和治理体系连接多模型与算力资源，并面向出海企业、品牌团队及创业者提供模型调用、AI 应用出海及基础算力服务。在收费模式方面，TokGate.io 采用“订阅费+Token 消耗收费”的模式，平台收取基础订阅会员费，并按 Token 消耗量收费；模型定价上，海外模型按照官方价格体系定价，国产模型基于与模型厂商合作价格进行动态调整。

仅供内部参考，请勿外传！

7、范式智能：从企业 AI 平台向 Token 基础设施延伸

范式智能长期布局企业 AI 基础设施，已形成覆盖算力、模型管理、数据治理及 AI 应用开发的能力体系。公司通过先知 AIOS、Phanthy 等平台，为企业提供算力资源管理、模型部署与管理、数据治理及 AI 应用开发能力，服务大型企业、运营商及垂直行业客户，为企业级 AI 应用规模化落地提供基础支撑。

图 24：范式提供“一站式 AIOS 解决方案”



资料来源：公司官网，招商证券

依托既有基础设施能力，公司进一步向 API 调用、模型服务及 Agent 场景延伸。公司适配开源大模型及国产算力资源，通过 API 向企业提供模型调用服务。目前相关服务覆盖大型企业、运营商、垂直行业客户以及第三方聚合平台，满足不同客户在 AI 应用部署和模型调用过程中的需求。同时，公司进一步推进 Agentic AI 业务，将大模型能力与企业业务流程结合，通过智能体完成知识检索、业务分析、流程自动化等任务，推动模型调用由单次交互向持续任务执行转变。

API 业务成为公司 Token 商业化的重要落点。随着企业客户从模型部署向规模化应用转变，模型调用需求快速增长，2026 年上半年，公司 API 业务实现收入 4.64 亿元，同比增长 860.8%，客户 Token 调用量同比增长约 620%，API 服务逐渐成为连接底层模型能力与企业应用场景的重要入口。收费模式方面，公司通过 API 接口向企业提供模型调用服务，客户根据实际模型使用量计费。随着 Agent 应用逐步深入企业业务流程，API 调用频率和 Token 消耗规模有望持续提升，进一步强化范式智能作为企业 AI 基础设施服务商的商业价值。

仅供内部参考，请勿外传！

五、风险提示

1、AI 应用及 Token 调用增长不及预期。AI 应用落地速度或使用频率不及预期，Token 调用规模增长放缓，影响行业发展及商业化进程。

2、多模型调用及智能路由需求不及预期。对单一模型依赖提升，或模型厂商自身路由能力持续增强，第三方多模型聚合及智能路由需求可能低于预期。

3、行业竞争加剧风险。云厂商、模型厂商及第三方平台持续布局 Token 运营，价格竞争加剧或服务趋于同质化，对平台费率、客户获取及盈利能力形成压力。

4、商业化进展不及预期风险。客户付费转化、调用规模或续费情况不及预期，相关收入贡献可能低于预期。

5、数据安全与合规风险。Token 运营涉及企业数据等信息，若相关监管趋严，可能增加合规成本并影响产品推广。

仅供内部参考，请勿外传！

分析师承诺

负责本研究报告的每一位证券分析师，在此申明，本报告清晰、准确地反映了分析师本人的研究观点。本人薪酬的任何部分过去不曾与、现在不与，未来也将不会与本报告中的具体推荐或观点直接或间接相关。

评级说明

报告中所涉及的投资评级采用相对评级体系，基于报告发布日后 6-12 个月内公司股价（或行业指数）相对同期当地市场基准指数的市场表现预期。其中，A 股市场以沪深 300 指数为基准；香港市场以恒生指数为基准；美国市场以标普 500 指数为基准。具体标准如下：

股票评级

强烈推荐：预期公司股价涨幅超越基准指数 20%以上

增持：预期公司股价涨幅超越基准指数 5-20%之间

中性：预期公司股价变动幅度相对基准指数介于±5%之间

减持：预期公司股价表现弱于基准指数 5%以上

行业评级

推荐：行业基本面向好，预期行业指数超越基准指数

中性：行业基本面稳定，预期行业指数跟随基准指数

回避：行业基本面转弱，预期行业指数弱于基准指数

重要声明

本报告由招商证券股份有限公司（以下简称“本公司”）编制。本公司具有中国证监会许可的证券投资咨询业务资格。本报告基于合法取得的信息，但本公司对这些信息的准确性和完整性不作任何保证。本报告所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。报告中的内容和意见仅供参考，并不构成对所述证券买卖的出价，在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。除法律或规则规定必须承担的责任外，本公司及其雇员不对使用本报告及其内容所引发的任何直接或间接损失负任何责任。本公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行业务服务。客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突。

本报告版权归本公司所有。本公司保留所有权利。未经本公司事先书面许可，任何机构和个人均不得以任何形式翻版、复制、引用或转载，否则，本公司将保留随时追究其法律责任的权利。