

LXRouter 托管式 Token 供应链白皮书

副标题：Token 是 AI 时代的水电煤——从市场格局到三层调度

版本：V1.0 · 2026 年 9 月

北京新灵犀科技有限公司（LingxiNova · 灵犀新星） 官网：<https://xinlx.com> | 产品入口：<http://lxrouter.xinlx.com/> 邮箱：bd@xinlx.com | 电话：+86-139-3096-5726

执行摘要

Token 市场的资源层集体背离买方立场，差价是运营者的属性而非平台的属性——我们的答案是三层调度：供给调度、请求调度、任务调度。

Token 是 AI 时代的水电煤，但它是一个特殊的商品：同一模型市场价 5~8 折并存是常态而非乱象——开源模型的价差由部署者成本结构差异决定，闭源模型的价差由又长又乱的 API 分销路径决定。资源层的三类玩家（模型原厂、Token 工厂、互联网大厂）各有自己的"必须"，利益与买方结构性错位，没有人站在买方立场替客户管理整个供应链。

中转站的集体误区是把差价当成平台的属性。实际上差价跟着运营者走：同一个渠道，有人谈到 5 折、有人只能 8 折；同一个池子，有人焊成稳定供给、有人裸奔转售。客户真正买不到的不是聚合，而是"有人替我运营、替我担责"。

LXRouter 的定位：**面向生产负载的托管式 Token 供应链平台**——不卖市场最低价，卖"保证可用前提下的低价"；资源管理、成本管理、服务管理是收费对象，差价只是入场券。产品上落地为三层调度：L1 供给调度（稳定相对低价的 Token 服务）、L2 请求调度（M 系列，让每个请求更合理）、L3 任务调度（Auto 系列，让每个任务更合理）。市场合作上三层模式各走各路：直客做深、大合作伙伴做通、城市合伙人做广。

本白皮书同时完整公开支撑这一定位的平台与技术体系：统一协议接入、同模型多厂商比价路由、请求复杂度分类器、双跑金标数据飞轮、全链路稳定性工程与分毫不差的计费计量体系——三层调度不是营销话术，而是运行在生产环境、经过真实流量长期验证的工程系统。

第一章 · Token 是什么

Token 是 AI 时代的水电煤——标准品又不是标准品，价差是它的常态。

无论上层是聊天机器人、编程助手、智能体还是行业应用，向下追溯，成本与体验的源头都收敛到同一件事：以什么价格、什么稳定性、什么效率，把 Token 送进模型、拿回结果。正如工业时代的工厂不必自建电厂、信息时代的企业不必自建机房，AI 时代的组织也不应被迫直面一个混乱、不透明、随时断供的 Token 市场——基础设施的意义，就是把底层的复杂性收走，让上层专注自己的业务。但要承担这个角色，首先要诚实地认识这个商品本身。Token 有三个与众不同的商品属性。

它是标准品，又不是标准品。 同一个模型，不同来源的 Token 价格可差 5~8 折，真实性上可以是原版、蒸馏、量化甚至旧版冒充——买家难以检测，感知到问题时只会静默流失。

它的价格在持续通缩。 模型原厂不断下调官方价，每一次降价都重塑整个市场的价格体系。去年成立的差价，今年可能不存在了。

它的供给天然冗余且混乱。 同一个模型，原厂在卖、云厂商在卖、Token 工厂在卖、无数中转站在卖，价格、稳定性、合规性各不相同。

价差为什么乱：两种供给模式各自的必然产物

开源模型必然产生价差。 权重开放意味着任何人都能部署，部署者的算力来源、利用率、工程水平千差万别——成本结构不同，售价自然拉开档次。这是开源的商业属性决定的，不会消失。

闭源模型同样必然产生价差。 闭源模型只以 API 形式存在，分销路径天然又长又乱：官方直销、大客户年框、渠道代理、资源包转售、共享号池，每一层折扣和让利空间都不同。路径越多，水越深。

所以"同一个模型，5 折到 8 折并存"不是阶段性的乱象，而是 Token 这个商品的常态。承认这一点是后面所有判断的前提：价差不会因为市场成熟而消失，等待"价格统一"等于等一件不会发生的事；真正的问题从来不是"有没有便宜的 Token"，而是"谁把便宜且可靠的 Token 持续地组织出来"。

第二章 · 资源层格局

三类资源玩家各有自己的"必须"，利益与买方结构性错位。

模型原厂：品牌专卖店——卖的是模型能力本身

定价是战略武器。最极端的是 DeepSeek：开源模型贴推理成本定价，让第三方托管算不过账——这是原厂中独一无二的打法，因为它同时手握权重和 API 两头。其他原厂走官方定价 + 量级折扣的常规体系。

低段价格被锁在量级门槛之内：大客户直接谈年框，普通客户和中小渠道够不到，原厂也不会让低段价格流到市场上伤害自己的直销体系。

能力边界清楚：只卖自己的模型，无法替客户解决多模型组合，更没有动力做跨厂冗余。

Token 工厂：批发市场——卖的是算力成本优势

自部署开源模型把成本压到官方价之下，是合法低价供给；但只覆盖开源模型，而开源恰是原厂价格战的重灾区，两头挤压。

这个细分已被做过两遍（国际 OpenRouter 厂商市场、国内硅基流动），同样的权重同样的 API，唯一变量是单价——商品化程度最高、利润最薄的角落。

工厂的低价是它自己的成本优势，对任何下游平台无依赖；该角落验真成本也最高（量化 / 旧 checkpoint 冒充）。

互联网大厂：超级市场——卖的是云生态入口

自部署 + 聚合的货架逻辑，平台永远倾向推自家模型，立场天然不中立。

企业级重量：IAM、对公合同、账期——对月消耗几万元的中小团队既是过度设计，也意味着没有存在感。

定价锚定官方价体系，折扣通过返点和代金券发给大框架客户，中小客户拿到牌价。

三类玩家对比

玩家	卖的本质	结构性盲区
模型原厂	模型能力	只有自己的模型，无跨厂冗余，低段价格锁在量级门槛内
Token 工厂	算力成本	只覆盖开源模型，商品化最彻底，验真成本最高
互联网大厂	云生态入口	立场不中立，企业级重量，小客户无存在感

资源层的共同盲区： 原厂必须推自己的模型，工厂必须卖自己的算力，大厂必须导自己的云——每一个资源方的利益都与买方结构性错位。而买方要的（跨厂冗余、中立比价、统一账单、出事有人担责）恰是任何单一资源方给不了也不愿给的。这不是没人想到，是没人能做——**只有不绑定任何一家上游的中立方，才有资格站在买方这一边。**

第三章 · 中游的三种玩法与集体误区

聚合、折扣、联盟——三种玩法都试过了，集体误区是把差价当成平台的属性。

聚合平台：把货架做到全球，解决不了价格

OpenRouter 证明了"一个 API 接入全球模型"的产品形态成立，也证明了纯撮合解决不了价格：能上架的正规供给天然就是官方价（实际成交 $\approx$ 官方价 +5% 抽佣），真正便宜的灰产供给无法在阳光下的平台存活。对中国客户还有三堵墙：海外主体结算与发票、数据出境合规、人民币支付不便。

性价比平台与自部署中转站：把折扣做到极致，把稳定性赔进去

开源网关一键部署让供给零门槛，竞争只剩折扣一个维度——而纯价差客户的忠诚度约等于 2% 的差价。灰市上"稳定"与"便宜"负相关：最便宜的货（共享号、来路不明池）恰恰最不稳定。更致命的是结构性威胁：凶手不是同行而是上游通缩——官方价每降一次，纯差价模式的生存空间就被削掉一层。

联盟模式为什么在结构上不成立

市场上有一种被反复讨论的方向："任何服务商入驻、客户买到全球最低价、平台统一结算"的联盟模式。从市场结构角度推演，这个模式存在四条死结：

1. **按价排序会把价格战制度化**——当排序规则公开以价格为先，竞争就被压缩到单价一个维度，供给端被迫无限内卷；
2. **逆向选择让平台只剩灰产供给**——正规供给拼不过灰产的成本，劣币驱逐良币，阳光下的货架最终被灰色货源占据；
3. **抽佣的那一刻平台就不再是最低价**——平台自身要生存就必须抽佣，而抽佣直接加在终端价格上，"全球最低价"的承诺自我瓦解；
4. **平台承担全球 KYC / AML / 外汇合规，成为最显眼的靶子**——统一结算意味着合规责任集中，监管风险由平台独扛。

OpenRouter 就是这个设想的活体样本：它做成了"聚合"，却始终做不出"最低价"。

差价是运营者的属性，不是平台的属性

差价从来不是平台的属性，而是运营者的属性——差价跟着人走，不跟着平台走。

平台只是货架和管道，货架本身不产生差价：同一个上游渠道，有人能谈到 5 折、有人只能拿到 8 折，差的是拿货能力和谈判筹码；同一个池子，有人能焊成"低价且稳定"的供给，有人只能裸奔转售，差的是多池冗余、健康熔断、死池秒切这些运营出来的确定性。一键部署的中转站没有差价、只有转发。**客户真正买不到的从来不是聚合，而是"有人替我运营、替我担责"。**

第四章 · 我们的答案：三层调度

不卖资源，卖运营；不卖模型列表，卖三层调度。

第一层 · 供给调度：稳定的、相对低价的 Token 服务 (L1)

这一层卖的不是资源本身——资源谁都能接——而是运营和服务：多池冗余（单一上游故障客户无感）、健康熔断与死池秒切、主动拨测与可验证的可用性、把混乱市场中的低价段焊成稳定供给。

SLA 和低价的来源是同一个东西：池子多。池子多，冗余就高、可用性就高；池子多，比价空间就大、成本就低。这也是为什么“稳定”和“低价”在别处往往负相关、在这里却能同时成立——它们不是两个被艰难平衡的目标，而是同一份供给运营能力的两个结果。

第二层 · 请求调度：M 系列，让每一个请求更合理 (L2)

不是所有请求都需要最贵的模型。M 系列（LingxiNova-M1 / M2 / M3）对每次请求做复杂度判定，简单请求走性价比模型、复杂请求走强模型，保守原则宁可错升不可错放。

红线是不可逾越的：绝不对客户做无感知的模型降级或替换——M 系列是客户显式选择的产品，调度逻辑与实测效果公开：生产实测满意度 87%，定价为对标旗舰的 6 折。护城河不在某个阈值参数，而在**双跑金标的数据飞轮**：每一次调度决策都有真实结果回流校验，判定模型在持续生产中越用越准。

第三层 · 任务调度：Auto 系列，让每一个任务更合理 (L3)

真实负载的单位不是请求，是任务：同一智能体任务的上百次调用，上下文大部分重叠、难度高度相关。Auto 通过上下文相似度识别“同一个任务”，为整个任务锁定合适的模型档位——任务内不再逐次分类、缓存命中滚雪球、避免跨模型横跳导致的质量漂移。

我们还做不到完全的“按任务计费、按效果计费”，但 Auto 是在朝这个方向走——目前这个方向上没有同类玩家。

三层调度对比

层	调度对象	回答的问题	收费的本质
L1 模型中转	供给（选哪个服务商）	这个模型从哪儿来最稳、最值	运营与服务
L2 M 系列	请求（选哪个模型）	这次调用值多少钱的模型	决策引擎
L3 Auto 系列	任务（选哪个档位）	整件事值多少钱的模型	决策引擎

三层共用同一个技术承诺：同一端点、同一 Key，改一个参数即换层。客户不需要更换接入方式，就能从"买到稳定供给"升级到"每个请求更合理"，再升级到"每个任务更合理"。三层不是三个产品，而是同一条供应链上逐层递进的运营深度——客户按自己的成熟度选择停在哪一层，随时可以向前走。

第五章 · 平台与技术体系

第四章回答"卖什么"——三层调度的产品承诺；本章回答"靠什么兑现"——支撑承诺的工程系统。

LXRouter 的全部技术设计服从三条主线：**热路径零数据库**——路由决策、余额扣减、限流、熔断全部走 Redis 与进程内缓存，数据库只做冷存储与批量落库；**一切外部依赖都有降级路径**，且降级永远不影响计费与路由的正确性；**保守主义工程取向**——宁可保守不可用，不可错误扣费、错误降档。

5.1 统一协议接入：一套 API 挂接多厂商

接入碎片化是大模型应用开发者的第一痛点：每家厂商一套 API 规范、一套鉴权、一套错误码，多模型切换成本极高。LXRouter 以统一 OpenAI 协议对外提供服务，用户对接一套 API，即可调用后台挂接的数十家模型厂商。平台同时支持 OpenAI、Anthropic (Claude)、Gemini 三大协议族的"原生透传 + 格式转换"双形态，共 19 个端点文件：对生态成熟的协议提供原生端点避免转换损耗；所有协议在内部收敛为统一调用表示，路由与计费逻辑只需实现一次。

多厂商接入的现实是：每家对协议的实现都有偏差。LXRouter 用**声明式 YAML 兼容层**解决，而不是在代码里堆特例——参数白名单过滤、重命名、互斥组、默认值填充的六步处理管线，加上消息级字段剥离与思考模式差异映射，按文件修改时间自动热重载。新厂商的参数适配改配置即生效，零代码、免重启。转发质量上，平台在边界吸收客户端常见脏数据（双重编码修复、孤儿工具调用移除等），并把数十家厂商的错误码收敛为统一错误分类（限流/鉴权/余额/参数/服务不可用）——这既是对外一致的错误体验，也是驱动自动换厂商决策的基础。

作为中间层，LXRouter 恪守两条标准化原则：**做减法不做加法**——剥离客户端发出的非标准字段残留，但不发明数据、不向上游注入非标准字段；**漏配后果不对称**——默认全量回传上游字段，拒收的模型通过配置剥离，把"漏配"的最坏后果锁在一次可热修复的报错上，而不是线上数据丢失。

5.2 同模型多厂商比价路由

同一模型挂接多个服务商后，路由层在它们之间按**价格、健康度、速度**三因子加权择优：

- **价格因子**按相对成本动态加权，并支持服务商动态成本规则（如时段折扣）——便宜的供给获得更高的选中概率，而不是简单一刀切选最低价；
- **健康因子**来自实时成功率统计，失败率、低流量全灭、慢响应三类熔断规则把故障供给自动移出候选；

- **速度因子**以首 token 延迟 (TTFT) 为基准动态调整，慢的供给降权。

为避免流量在厂商间无谓漂移，系统对同一调用方保持"粘性"，以保住上游前缀缓存的命中率；但当候选最优成本比当前厂商低出阈值（默认 5%）时，整体迁移到更便宜的厂商——**既保持缓存亲和，又不被粘性锁死在高价上**。整个路由决策在缓存命中时延迟小于 5ms。客户无需做任何事，拿到的始终是池子里"此刻最值"的那一份供给——这就是"比价"从口号变成机制的地方。

5.3 请求复杂度分类器与智能路由

M 系列的核心是一台请求复杂度分类器：对每次请求输出 1~5 级的绝对难度分 cx （任务要求的认知步骤数 $\times$ 输出开放度），再按各档位能力画像映射到轻量档或旗舰档。判分管线为五级漏斗：ACK 地板（空输入、纯确认语一律走强档） $\rightarrow$ 规则地板（报错堆栈等确定性难特征短路） $\rightarrow$ 记忆化缓存 $\rightarrow$ L2 小模型打分 $\rightarrow$ L3 边界复核。

管线遵守不可逾越的红线：**L3 复核只升不降**；"判易"只能由模型判分加复核做出；任何失败或超时**一律回退强档——宁可错升，不可错放**。 cx 边界即产品阈值（当前 M3 口径： $cx \leq 3$ 走轻量档， $cx = 4 \sim 5$ 走旗舰档），这个边界不是猜的，是用金标样本按充足率目标反推出来的。金标 263 例上，L3 复核后边界准确率 100%，而月判分成本约 1 元——判分精度靠输入信息工程（短输入全量喂入、长会话取"最新指令 + 执行现场快照"），而不是靠判分模型本身的昂贵。

实践还证明了一件反直觉的事：**单请求的难度无法被可靠预测**（轮级难度存在接近 50% 的随机成分），**但翻车可以被快速识别**。因此 M 系列的架构是"保守分类 + 反应式纠偏"：新会话首请求直接走旗舰档（起点强注）；一旦检测到上一轮翻车证据（空输出、用户重试、模型打转），本会话立即锁定旗舰档；轻量档连续承接达到阈值时，强制升级探测一次。上下文物理门独立于分类器——任何档位装不下的长上下文，自动顺延至更高档，物理约束优先于分类结果。

5.4 双跑金标与数据飞轮的技术实现

第四章所说的"双跑金标数据飞轮"不是修辞，是一套运行中的机制。对抽样请求，系统在返回主结果的同时静默调用另一档模型作为"影子"，再由大模型裁判对两份回答做盲评（better / equal / worse）。双跑分两个方向，各回答一个关键问题：

- **正向影子**（轻量档主跑、旗舰档做金标）回答"轻量档接不接得住"——**充足率**；
- **反向影子**（旗舰档主跑、轻量档做影子）回答"旗舰档流量里有多少其实可以省下来"——**可下放率**。

新后端模型上线前必须通过双跑验收：影子调用并行发送到候选后端，回收响应做保留度比对，**保留度 $\geq 95\%$ 才允许启用**——用极小的影子调用成本，换取降档决策的质量保证。任何路由参数变

更，必须用双跑金标数据回放验证充足率不退化才允许上线——所有策略调整过金标闸，禁止拍脑袋调阈值。裁判结论逐条回流样本库，成为分类器训练、阈值校准与后续路由策略的持续数据源：这是飞轮闭合的地方。

到 2026 年 9 月中，该体系已积累 **803 条带裁判结论的金标样本**，覆盖真实生产请求的完整分布（平均输入长度 3.5 万 token，含长文档与多轮工具调用场景）。

生产实测（2026-09-07 ~ 09-15 窗口，详见《LingxiNova-M3 技术报告》）：

指标	数值
轻量档承接率	31.5%（按请求数）
用户满意率（轻量档 vs 旗舰档）	87.0% vs 87.7%
整体推理成本 vs 全量旗舰直调	下降约 30%
请求成功率	97.8%
轻量档充足率（金标裁判口径）	74.2%
可下放率（反向影子）	67.5%

两个档位在用户感知质量上的差距小于 1 个百分点——这是 M 系列产品主张的核心证据：**用户不需要知道自己被调度到了哪一档**。调度节省的成本，大部分让渡给用户（对标旗舰 6 折定价），这是"效率红利共享"的定价逻辑，而不是成本加成逻辑。

响应性能上，轻量档同样占优：在平均输入 3.5 万 token 的长上下文场景下，轻量档非流式整体时延 p50 为 7.4s、p90 为 17.5s，旗舰档为 8.8s / 38.5s——轻量档 p90 快约一倍；轻量档缓存读命中率 55.8%，比旗舰档高出约 17 个百分点，长文档复用场景优势显著。

5.5 稳定性工程

"池子多"要变成"客户无感"，靠的是一套稳定性工程的落地：

- **多池冗余与熔断**：失败率、低流量全灭、慢响应三类熔断规则在供给故障时自动移出候选池；极端情况下全部熔断时，按最近平均 TTFT 选最快者兜底——矮子里拔将军，可用性优先。
- **健康熔断与死池秒切**：厂商错误统一分类后触发 fallback 链自动换厂商；平台**不做同厂商内重试**——同厂商重试大概率撞上同一故障，换池才是正解。
- **TTFT 守卫**：流式请求在时限内拿不到首字节即判定超时并切换厂商，首字节透明拼回流，客户端无感；TTFT 阈值随输入上下文长度自动放大，长思考模型不被误杀。

- **分层超时**：连接、池等待、流式停顿、超长生成分别有独立预算——384K tokens 级长生成合法，卡顿流快速失败，总时长以"持续出字"为准。
- **主动拨测与状态页**：对供给做主动拨测，平台可用率按"供给侧故障"口径对外公开——可用性是可验证的数字，不是口头承诺。
- **容量与降级矩阵**：单集群 25,600 并发；9 类组件故障均有预置降级路径——Redis 故障降级同步结算、分类器故障回退强档、配置解析失败回退旧配置——降级永远不影响主流程正确性。
- **可观测**：分钟级实时指标持久化，fallback 路由日志与错误日志独立落盘，虚拟模型逐请求记录路由决策明细——每一次调度、每一次换厂商都有据可查。

5.6 计费与计量体系

计费是交易平台的生命线。设计目标是三句话：**分毫不差、绝不阻塞请求、任何组件故障都不出错扣费。**

- **四层计费**：输入、输出、缓存读、缓存写四层分别计价——命中前缀缓存的 token 按缓存价单独低价计费，缓存红利按真实命中如实让利给客户；另支持图片、视频等非 token 计量模型的按量计费。
- **分位精度滚动累计**：单次调用费用可能远小于 1 分，逐次四舍五入会系统性多扣或少扣。平台按"用户 + 模型 + 价格签名"分段，以万分之一分为单位滚动累计，每次只扣整数分、余数滚动，段内误差恒小于 1 分。
- **原子扣款与零阻塞**：Redis Lua 脚本原子扣款、计费条目入队，请求路径数据库写入 0 次，后台批量结算落库；Redis 不可用时降级为同步结算——功能降级，正确性不降级。
- **预扣-结算-差额返还**：请求前按最大可能费用预扣防超支，请求后按实际上游用量精确结算，多扣返还、少扣补扣。
- **可审计**：每笔余额变动逐笔流水可查；调用记录含 60+ 字段（四层 token、路由决策、费用明细），用户余额在任何时刻都可由流水重放验证。计费审计制度化——定期逐项核对累计器口径与流水恒等，发现的差异项均记录并跟踪修复。

5.7 工程体系兜底

技术优势最终由工程体系兜底：796 项自动化测试覆盖路由选择、熔断、计费口径、累计器精度、兼容层管线等核心路径；分类器等核心逻辑的每次变更必须配套金标集回归；平台核心运营配置（厂商挂接、优先级、权重、定价、路由阈值）全部免重启热更新，厂商侧报错从发现到修复以分钟计；线上事故的复盘结论沉淀为设计原则，回写进兼容层与清洗层。亮点不是孤立的技巧，而是

可维护、可演进、可审计的系统工程——这正是"托管"二字的技术含义：客户把供应链的复杂性交给我们，我们把确定性交给客户。

第六章 · 合作模式

三类合作对象，三种相处方式：直客做深、大合作伙伴做通、城市合伙人做广。

直客：做深，以钱包份额为北极星指标

直客是有真实生产负载的终端客户：有自己的主营业务，Token 是刚性成本但不是全部成本，他们要的不是绝对最低价，而是"可用性 + 全系列模型 + 省心"的组合真实成立。LXRouter 通过 Token 工厂和合作商拿到的成本显著低于客户自采价格，以温和的成本加成定价即可支撑稳定合作。

因为市场价格混乱，客户真正买不到的不是低价，是"低且稳"。对直客，我们的北极星指标是**存量客户钱包份额**——不比拉新速度，比的是客户把多少 Token 消耗放心地交给我们。

大合作伙伴：双向的供应链关系

大合作伙伴是本身也对外提供托管服务的大体量服务商。我们不去抢份额，而是建立双向的供应链关系：

- **向下，承接自然调度流量**——合作伙伴不可能每个模型的成本都远低于我们，在成本持平或我们更优的模型上，流量调度过来即可成立；
- **向上，互通低价资源**——合作伙伴手里的低价资源也是我们"池子多"的一部分来源，我们同样向他们开放我们的优势供给。

这是供应链思维，不是竞争思维：在供给层互为上下游，而不是在客户端互为对手。

城市合伙人：渠道做广，利润主动让渡给渠道

面向全国市场的城市合伙人体系，核心原则只有一条：**对外标价必须为渠道留足利润空间，利润主动让渡给渠道**。标价体系为渠道设计——标价即低价会反过来咬死渠道空间，伤害整个体系。渠道拿走量贩的利润，平台拿走规模和管理费，共同把市场做广。

第七章 · 与 BXAGI 的关系：两层一盘棋

Token 层把智能管好送稳，应用层把智能用好沉淀住。

新灵犀的产品体系是"两层一盘棋"：Token 平台 LXRouter 管智能的供给，BXAGI 平台管智能的组织；两层各自成立，又彼此成全，都不绑定上游，都站在买方一边。

- **LXRouter 是供应链层。** 它解决"智能从哪来、多少钱、稳不稳"的问题——多池冗余、健康熔断、三层调度，让 Token 像水电煤一样随取随用、可度量、可担责。
- **BXAGI 是应用组织层。** 它解决"智能怎么进组织、怎么干活、怎么沉淀"的问题——把模型能力组织成可管理、可考核、可持续进化的"AGI 员工"，让智能真正嵌入企业业务流程。

两层之间的成全关系是天然的：BXAGI 的 AGI 员工需要稳定、可负担、可调度的智能供给，LXRouter 正是它的供应链底座；LXRouter 的三层调度能力，在 BXAGI 的真实生产负载中得到最充分的验证与打磨。两层又各自独立成立：LXRouter 面向一切需要 Token 供给的客户开放，BXAGI 也可以运行在任何合规的模型供给之上——**都不绑定上游，都站在买方一边**，这是两层共同继承的结构性立场。

关于 BXAGI 的完整产品体系，请参阅兄弟白皮书《BXAGI 企业级 AGI 员工平台白皮书 V3.0》(<https://xinlx.com/bxagi>)。对客户而言，两层可以分开采用，也可以整体采纳：先用 LXRouter 把智能供给管好送稳，再用 BXAGI 把智能用好沉淀住——这是新灵犀对"让 AI 给公司赋能"这一使命的完整工程化回答。

第八章 · 总结

别人卷的是折扣，我们建的是供应链。

- **市场盲区**：原厂、工厂、大厂各有"必须"，利益与买方结构性错位，没人站在买方立场管理供应链——只有不绑定任何上游的中立方才能站过去，这是 LXRouter 最重要的结构性优势。
- **价差常态**：开源由成本结构差异决定，闭源由分销路径又长又乱决定；差价是运营者的属性，不是平台的属性。
- **我们的定位**：面向生产负载的托管式 Token 供应链平台。不卖市场最低价，卖"保证可用前提下的低价"；资源、成本、服务的管理是收费对象，差价只是入场券。
- **我们的产品**：L1 供给调度（稳定相对低价的 Token 服务）、L2 请求调度（M 系列，让每个请求更合理）、L3 任务调度（Auto 系列，让每个任务更合理）。同一端点、同一 Key，改一个参数即换层。
- **我们的平台**：统一协议接入、同模型多厂商比价路由、复杂度分类器与双跑金标数据飞轮、全链路稳定性工程、分毫不差的计费体系——三层调度承诺背后的技术兑现。
- **我们的市场**：直客做深（钱包份额）、合作伙伴做通（自然调度 + 资源互补）、渠道做广（城市合伙人，利润让渡）。

SLA 和低价同源，都来自池子多；稳定和智能同源，都来自运营深。

Token 市场的混乱不会自行结束，但混乱恰恰是供应链能力价值最大的地方。当折扣战争逐一退潮，留下来的一定不是报价最低的玩家，而是把供给、成本、服务运营成确定性的玩家。

这就是新灵犀对 Token 供应链的回答。

关于新灵犀

北京新灵犀科技有限公司（品牌：LingxiNova · 灵犀新星），国家级高新技术企业。

- **使命：**让 AI 给公司赋能
- **品牌口号：**探索 AGI 的无限可能
- **产品体系：**两层一盘棋——Token 平台 LXRouter 管智能的供给，BXAGI 平台管智能的组织
- **地址：**北京市朝阳区望京SOHO T1B座1209
- **电话：**+86-139-3096-5726
- **邮箱：**bd@xinlx.com
- **官网：**<https://xinlx.com>
- **LXRouter 平台入口：**<http://lxrouter.xinlx.com/> | **官网产品页：**<https://xinlx.com/token>

文档声明

本文为公开资料，版权归北京新灵犀科技有限公司所有，欢迎完整转载并注明出处。文中数据来自生产实测与公开市场信息；涉及第三方产品与公司的描述基于公开信息整理，仅用于市场格局分析。本文不构成任何投资建议或商业承诺。