目　录

AI网关

AI网关

AI 网关的作用与市场趋势

一、什么是 AI 网关

AI 网关（AI Gateway / LLM Gateway）是位于应用与大模型服务之间的中间层，负责统一入口、智能路由、流量治理与安全管控，类似微服务架构中的 API 网关之于大模型调用。它对上提供统一 API，对下屏蔽数百个模型提供商的差异（认证方式、请求格式、计费模式均不同），让应用无需适配 SDK 细节，也无需关心底层模型由谁提供。

二、核心作用

1. 统一接入

一个端点对接所有模型：统一的 OpenAI 兼容接口（ `/v1/chat/completions` ）即可同时调用 OpenAI、Claude、Gemini、DeepSeek、Kimi、GLM 等上千种模型，应用零改造即可切换供应商。

2. 智能路由与容灾（高可用）

- 健康检查与熔断：自动识别限流（429）、5xx 故障并踢出问题提供商，毫秒级切换
- 自动故障转移（Fallback）：主模型失败/配额耗尽时，自动降级到备用模型链
- 多级路由策略：优先级、权重、轮询、最低成本、最高配额余量、最佳缓存命中率等（如 OmniRoute 提供 19 种策略）

3. 成本治理

- 配额与预算管控：按团队/项目/Key 设置每日预算与速率限制，防超支
- 路由降级链路：订阅账号 → API Key → 低价模型 → 免费档，把成本压到最低
- Token 压缩与语义缓存：压缩工具输出（可省 15–95% token）、缓存重复请求前缀，直接降低账单
- 成本归因：精确统计每次调用花在哪个模型多少钱

4. 可观测性

统一日志、指标、链路追踪与用量看板；每次响应携带路由决策、成本、延迟等元数据，故障可回溯、用量可审计。

5. 安全与合规

- 密钥集中托管、AES 加密存储，业务代码不再接触裸 Key
- Prompt 注入防护、内容过滤、敏感信息（密钥/手机号/身份证）脱敏
- 细粒度访问控制、审计日志，满足企业合规要求

三、市场主要产品

类型	代表产品	特点
开源自托管	LiteLLM（BerriAI）	社区最活跃的通用网关，Python 生态，代理 + SDK 双形态
开源自托管	One-API / New-API	中文社区最流行的网关，多模型、多 Key 计费，Docker 一键部署
开源自托管	Higress AI Gateway（阿里）	企业级入口网关的 AI 扩展，K8s/云原生场景
开源自托管	OmniRoute	面向编码 Agent 的免费网关，290 提供商、内置免费额度与 MCP/A2A，本地优先
商业托管	OpenRouter	模型转售市场，一个 Key 买所有模型，按量计费（含抽成）
商业托管	Portkey	企业级网关平台，强于治理、护栏与观测
商业托管	Kong AI Gateway	老牌 API 网关的 AI 扩展，企业集成能力成熟
商业托管	Cloudflare AI Gateway	部署在 CDN 边缘节点，缓存、限流与全球低延迟
观测类	Helicone	专注于 LLM 可观测性与分析

四、市场趋势（2026）

1. 从「接得通」到「接得稳」

多模型接入已不再是难点，竞争转向韧性：三级自愈（提供商熔断 → 连接冷却 → 模型锁定）、防惊群、429 退避等生产级能力成为标配。

2. 成本焦虑推动「自动省钱」

网关从被动代理转向主动省钱：基于实时定价的成本最优路由、订阅/API/低价/免费的分层降级链路、聚合免费额度最大化（OmniRoute 聚合约 15 亿免费 token/月）、Token 压缩与缓存复用成为卖点。

3. Agent 原生成为新战场

面向 Claude Code、Codex、Cursor、Cline 等编码 Agent的网关快速崛起：内置 MCP Server、A2A（Agent-to-Agent）协议、一键配置 30+ 工具、对工具调用密集场景做专项 token 压缩。这回答了「后端早就 Agent 化了，网关如何 Agent 化」的命题。

4. 安全与合规权重上升

数据不出域（本地优先、自托管）、Prompt 注入防护、凭据脱敏、密钥 AES 加密成为标配；企业客户对审计、RBAC、OIDC 的要求驱动网关向治理平台演化。

5. 两条路线分化：模型市场 vs 基础设施