目　录

OmniRoute 项目介绍

OmniRoute 项目介绍

OmniRoute 项目介绍

项目定位

OmniRoute 是一个 MIT 开源的免费 AI 网关，口号是「Never stop coding」。它把 290 个 AI 提供商（其中 90+ 免费档）、1200+ 模型聚合到一个本地端点后面，让 Claude Code、Codex、Cursor、Cline、Copilot 等所有 OpenAI 兼容工具都能直接跑在免费或低价模型上。

- 仓库地址：https://github.com/diegosouzapw/OmniRoute
- npm 包：`omniroute`，Docker 镜像：`diegosouzapw/omniroute`
- 作者：Diego Souza（巴西），创始开源赞助方为 Kimi（月之暗面）

核心承诺（六大支柱）

1. 一个端点：所有请求走 `http://localhost:20128/v1`，兼容 OpenAI / Claude / Gemini / Responses API
2. 自动故障转移：毫秒级跨 290 个提供商降级，永不停机
3. Token 压缩：RTK + Caveman 双层堆叠压缩，节省 15–95% token（工具密集型会话平均约 89%）
4. $0 起步：90+ 免费档、40+ 永久免费，无需绑卡；聚合统计约 15.3 亿免费 token/月
5. 所有工具可用：一个配置适配 33+ 编码 Agent
6. 生产级能力：熔断器、TLS 指纹隐身、内置 MCP（104 个工具）、A2A、记忆、护栏、评估，25000+ 测试

旗舰功能：Combos（模型链路由）

Combo 是一串模型，OmniRoute 在其间自动路由：配额耗尽、提供商故障或成本飙升时，自动静默切换到下一个模型。

零配置 `auto`

直接设 `model: auto` 即可，另提供变体：

Model ID	优化目标
`auto`	均衡默认（LKGP，记住上次成功的提供商）
`auto/coding`	代码生成质量优先
`auto/fast`	低延迟优先
`auto/cheap`	每 token 成本最低
`auto/offline`	配额余量最多优先
`auto/smart`	质量优先 + 10% 探索

19 种路由策略

可混搭：`priority`、`fill-first`、`weighted`、`round-robin`、`p2c`、`least-used`、`random`、`strict-random`、`cost-optimized`、`headroom`、`reset-window`、`reset-aware`、`context-relay`、`context-optimized`、`cache-optimized`、`lkgp`、`auto`、`fusion`（多模型投票 + 裁判合成）、`pipeline`（链式流水线）。

三级自愈韧性

- Level 1 提供商熔断器：仅 408/5xx 触发，半开探测 + 惰性恢复
- Level 2 连接冷却：指数退避（×2），429 尊重 Retry-After，防惊群
- Level 3 模型锁定：单个模型 $^{429}/_{404}$ 只锁该模型，不波及连接

其他亮点

- Tier 级联降级：订阅账号 → API Key（DeepSeek、Groq）→ 低价（GLM、MiniMax）→ 免费档
- MCP / A2A / CLI：内置 MCP 服务器与 Agent-to-Agent 协议，`setup-*` 一条命令配置 12+ 编码工具
- 安全：每路由 prompt 注入检测、凭证脱敏护栏、本地优先 + AES-256-GCM 加密存 Key、可选 OIDC 登录
- 可观测：每个响应带 `X-OmniRoute-Decision` / `X-OmniRoute-*` 头，展示路由决策、成本与用量
- 媒体生成：图像 / 视频 / 音频 / OCR / 语音翻译
- 远程模式：scoped token 驱动远端实例，支持 VPS 部署

运行环境

Node.js 项目，存储用 `node:sqlite`（Node 22+）或 WASM 版 sql.js。部署形态：

- npm 全局安装：`npm i -g omniroute`
- Docker：AMD64 + ARM64 多架构
- Electron 桌面端：Windows / macOS / Linux
- Android Termux：手机 $^{24}/_{7}$ 运行，无需 root
- PWA、OpenCode 插件、源码运行

Docker 部署（本机已实测）

```
docker run -d --name omniroute --restart unless-stopped --stop-timeout 40 \
  -p 127.0.0.1:20128:20128 -v omniroute-data:/app/data diegosouzapw/omniroute:latest
```

- Dashboard：`http://127.0.0.1:20128/dashboard`
- 常用地址：`/dashboard/free-tiers`（免费额度汇总）
- 装完零配置即可用：

```
curl http://localhost:20128/v1/chat/completions \
  -H "Content-Type: application/json" \
  -d '{"model":"auto","messages":[{"role":"user","content":"Hello!"}]}'
```