

CẨM NANG CÀI ĐẶT & SỬ DỤNG OMNIROUTE AI GATEWAY

Cổng Kết Nối AI Đa Nhà Cung Cấp • 352+ Providers • Nén Token RTK • Quản Lý Hạn Mức Tự Động

★ **GitHub Repository:** <https://github.com/diegousouzapw/OmniRoute> (~6.800+ Stars)

👑 **Tác giả / Tổ chức:** Diego Souza (diegosouzapw)

Giấy phép: MIT

Chuẩn API: 100% tương thích OpenAI API (/v1/chat/completions, /v1/models)

Khả năng tương thích: Claude Code, Cursor, Windsurf, Cline, Continue, Aider, OpenClaw

Tính năng độc quyền: RTK + Caveman token compression (tiết kiệm 15-95% token), Auto-Fallback

1. OMNIROUTE LÀ GÌ & VẤN ĐỀ NÓ GIẢI QUYẾT?

Khi làm việc với các coding agent hiện đại (Claude Code, Cursor, Cline...), nhà phát triển thường xuyên gặp phải hai vấn đề nhức nhối: cạn kiệt hạn mức quota (Rate Limit) và chi phí token khổng lồ khi phải mua nhiều gói API riêng biệt từ OpenAI, Anthropic, Google, DeepSeek, Groq, OpenRouter...

OmniRoute hoạt động như một AI Gateway trung gian đặt ngay trên máy hoặc máy chủ cá nhân của bạn, quy tụ hơn 352 nhà cung cấp và 1.300+ mô hình về đúng 1 địa chỉ endpoint duy nhất: <http://localhost:20128/v1>.

- **Cơ chế định tuyến thông minh (Smart Routing & Auto-Fallback):** Nếu Claude 3.5 Sonnet bị nghẽn hoặc hết tiền, OmniRoute tự động chuyển sang DeepSeek-V3 hoặc Gemini 1.5 Pro trong tích tắc mà không làm đứt đoạn lệnh của Agent.
- **Công nghệ nén Token RTK + Caveman:** Tự động loại bỏ từ ngữ đệm thừa, rút gọn JSON rác và rút ngắn context trước khi gửi đi, giúp giảm 15% đến 95% chi phí API.
- **Đồng bộ Session trình duyệt:** Hỗ trợ kết nối tài khoản web ChatGPT / Claude / Gemini mà không bắt buộc phải có API key thương mại đắt đỏ.
- **Web Dashboard trực quan:** Theo dõi số token tiêu thụ, chi phí từng phút, thời gian phản hồi (TTFT) và biểu đồ sức khỏe của từng nhà cung cấp.

2. HƯỚNG DẪN CÀI ĐẶT QUA NPM / DOCKER

Cách 1: Cài đặt trực tiếp qua Node.js (Khuyến nghị cho Developer)

Cài đặt gói toàn cục và khởi động

```
# Cài đặt OmniRoute qua npm
npm install -g omniroute
```

```
# Khởi động gateway
omniroute

# Dashboard sẽ mở tại: http://localhost:20128/dashboard
```

Cách 2: Triển khai qua Docker Compose (Cho HomeLab / VPS)

File docker-compose.yml

```
services:
  omniroute:
    image: ghcr.io/diegosouzapw/omniroute:latest
    container_name: omniroute
    restart: unless-stopped
    ports:
      - "20128:20128"
    volumes:
      - ./config:/app/config
    environment:
      - PORT=20128
      - LOG_LEVEL=info
```

Lệnh khởi chạy container

```
docker compose up -d
```

3. KẾT NỐI VÀO CLAUDE CODE, CURSOR VÀ CLINE

Bất kỳ công cụ nào hỗ trợ định dạng OpenAI-compatible đều có thể chuyển hướng sang OmniRoute chỉ với 2 tham số:

Cấu hình cho Claude Code / Shell

```
export OPENAI_BASE_URL="http://localhost:20128/v1"
export OPENAI_API_KEY="omniroute-local-key"
```

Cấu hình trong Cursor / Windsurf Settings

```
- Override OpenAI Base URL: http://localhost:20128/v1
- API Key: omniroute-local-key
- Model Name: omniroute/smart-code-tier (tự động chọn model mạnh nhất còn quota)
```

4. KÍCH HOẠT CHẾ ĐỘ TIẾT KIỆM TOKEN (CAVEMAN COMPRESSION)

Trong Dashboard của OmniRoute (<http://localhost:20128/dashboard>), hãy chuyển sang tab 'Optimization':

- **Bật 'Caveman Mode':** Nén câu văn yêu cầu thành dạng súc tích, cắt giảm 20-30% token đầu vào.

- **Bật 'JSON Mini-Strip':** Lược bỏ khoảng trắng và thuộc tính null không cần thiết trong các tool call.
- **Đặt ngưỡng Fallback:** Thiết lập nếu Provider chính trả về mã lỗi 429 (Rate Limited) thì tự động chuyển tiếp sang Provider dự phòng trong 0.5s.

Tài khoản githot mới được tặng 5 xu để dùng thử công cụ AI Repo Explainer hoặc Repo Health Check.

Đăng ký: githot.io.vn/register · Tìm thêm hướng dẫn cài đặt: githot.io.vn

Chúc bạn tối ưu chi phí API và xây dựng hệ thống AI Agent siêu ổn định cùng OmniRoute!

© Kênh AI Mỗi Ngày | Hướng dẫn AI & Linux Tự Động Hóa

Nội dung kỹ thuật tổng hợp từ tài liệu chính thức của repo <https://github.com/diegosouzapw/OmniRoute>