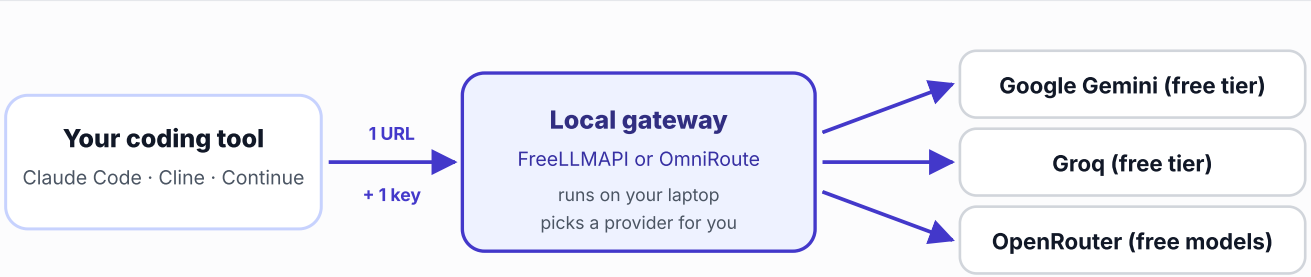
STUDENT SETUP GUIDE

Free LLM API Setup Guide

Hit your AI limit in the middle of a project? Combine the free tiers of several AI providers behind **one local API**, then use it inside your coding tool.



When one provider hits its limit, the gateway moves to the next one.

START HERE

FreeLLMAPI

Desktop app for Windows and macOS. No dashboard password. The easiest way to try this.

github.com/tashfeenahmed/freellmapi

ADVANCED

OmniRoute

More features and more providers, with more setup (npm or Docker, and a password to manage).

github.com/cheaper-inference/OmniRoute

	FreeLLMAPI	OmniRoute
Install with	Desktop app, or Docker	npm, or Docker
Needs	Nothing extra for the desktop app	Node.js 22.22.2+ (or 24+), or Docker
Local address	localhost:31415 (desktop) localhost:3001 (Docker)	localhost:20128
Free capacity **(the project's own claim)**	up to ~7.4 billion tokens/month, only if you connect *all* providers	~1.6 billion tokens/month (catalog estimate across providers)

Read this first: what you will and won't get

- The big token numbers are the **sum of many providers' free quotas**. You need your **own free key** from each provider. Three providers give far less than the headline.
- These are **free models** (like Gemini Flash, Llama and Qwen), not premium Claude, ChatGPT or Gemini plans.
- Limits don't disappear: they rotate. Quality can dip late in the day when top models hit their caps.
- Both projects say they are for **learning and experiments**, not production work.

Before you start

About 15 minutes. You only do this part once, and the same keys work with both tools.

What you need

Item	Details
A computer	Windows 10/11, macOS 12 or later, or Linux.
A coding tool	Claude Code, or the Cline or Continue extension in VS Code. Install it first and make sure it opens.
Node.js	Needed for the `npx` and `npm` commands. Check with `node -v`. FreeLLMAPI needs 20.18 or newer; OmniRoute needs 22.22.2 or newer. Get it from nodejs.org.
Docker (optional)	Only for the Docker install paths. Get it from docker.com.
A password manager or private note	You are about to create several API keys. Keep them private.

Get 3 free provider keys

These are the three providers we connect in this guide. Sign up, create an API key, and save it somewhere private.

Provider	Create your key here	Good to know
Google Gemini	aistudio.google.com/apikey	Sign in with a Google account. Choose "Create API key".
Groq	console.groq.com/keys	Fast open models such as Llama. Free tier has rate limits.
OpenRouter	openrouter.ai/keys	Free models have a `:free` ending in their name.

Fair-use rules

- Use **one account per provider**. Don't make extra accounts to get around limits.
- Free tiers change often. Check each provider's current limits and data policy on its own page.

Treat your keys like passwords

Never post them, put them in screenshots or commit them to GitHub. If one leaks, delete it in the provider's dashboard and create a new one.

Part 1: FreeLLMAPI START HERE

Repo: github.com/tashfeenahmed/freellmapi

> **Which port?**
> The desktop app runs on **port 31415**. Docker installs run on **port 3001**. The commands below show 3001. If you use the desktop app, swap in 31415.

1 Install it (pick ONE way)

A · EASIEST	B · FAST	C · TRANSPARENT
Desktop app	**Docker one-liner**	**Docker Compose**
Windows and macOS. No Docker, no password.	Needs Docker. One command does the setup.	Needs Docker and Git. You see every step.

A. Desktop app. Open the official Releases page.

- **Windows:** download the `.exe` installer and run it.
- **macOS:** download the `.dmg` (`arm64` for Apple Silicon, `x64` for Intel). Needs macOS 12 or later.
- Open the app. An icon appears in your menu bar or system tray. To open the dashboard: tray icon, then **Open Dashboard**.

B. Docker one-liner (Docker must be running). Read the script first at freellmapi.co/install.sh, then run:

```
curl -fsSL https://freellmapi.co/install.sh | bash
```

It sets up `~/freellmapi`, creates an encryption key, pulls the image and starts it. Open **http://localhost:3001** and create the first account (email and password, stored on your own computer).

C. Docker Compose (macOS, Linux or Windows with WSL):

```
git clone https://github.com/tashfeenahmed/freellmapi.git
cd freellmapi
ENCRYPTION_KEY="$(openssl rand -hex 32)"
printf "ENCRYPTION_KEY=%s\nPORT=3001\n" "$ENCRYPTION_KEY" > .env
docker compose up -d
```

Then open **http://localhost:3001** and create the first account.

2 Add your 3 provider keys

In the dashboard, open the **Keys** page. Add one key each for **Google AI Studio**, **Groq** and **OpenRouter** (the keys you created earlier). Optional: open the **Fallback** page to change the order in which providers are tried.

3 Copy your unified key

This is the one key your coding tool uses. It starts with `freellmapi-`.

- **Desktop app:** click the tray icon, then **Copy Key**.
- **Docker:** it is shown in the header of the **Keys** page.

4 Test it before using it

Easiest: open the **Playground** page in the dashboard and send "hi". If you get a reply, it works.

Or use a terminal (macOS, Linux, or Git Bash on Windows):

```
curl http://localhost:3001/v1/chat/completions \
  -H "Authorization: Bearer freellmapi-your-unified-key" \
  -H "Content-Type: application/json" \
  -d '{"model":"auto","messages":[{"role":"user","content":"hi"}]}'
```

`"model":"auto"` lets the gateway choose the provider. Try `auto:smart` for harder coding tasks or `auto:fast` for speed.

5 Connect your coding tool

Pick the tool you use and run its command in a terminal. Add `--dry-run` first to preview changes without writing anything. The dashboard's **Agents** page shows the same commands with your key already filled in.

CLAUDE CODE

```
npx freellmapi setup-claude --url http://localhost:3001 --api-key freellmapi-your-unified-key
```

Claude Code needs the plain address with **no** `/v1`. Restart Claude Code afterwards. The model behind it is a free model, not Claude.

CLINE (VS CODE)

```
npx freellmapi setup-cline --url http://localhost:3001 --api-key freellmapi-your-unified-key
```

Manual option: in Cline choose the **OpenAI Compatible** provider, Base URL `http://localhost:3001/v1`, your unified key, Model ID `auto`.

CONTINUE (VS CODE)

```
npx freellmapi setup-continue --url http://localhost:3001 --api-key freellmapi-your-unified-key
```

Manual option: add this under `models:` in your Continue config.

```
  - name: FreeLLMAPI Auto
    provider: openai
    model: auto
    apiBase: http://localhost:3001/v1
    apiKey: freellmapi-your-unified-key
```

Daily rhythm

Daily quotas reset at **UTC midnight, which is 5:30 AM in India**. If answers get weaker in the evening, top models have hit their caps. Come back after the reset, or try `auto:smart`.

Part 2: OmniRoute ADVANCED

Repo: github.com/cheaper-inference/OmniRoute · npm: npmjs.com/package/omniroute · omniroute.online

The project moved to the address above. The older address, github.com/diegosouzapw/OmniRoute, still redirects to it.

1 Install it (pick ONE way)

A. npm (recommended). First check `node -v`: you need 22.22.2 or newer.

```
npm install -g omniroute
omniroute
```

Using npm 11 or newer and npm blocks the install scripts? Run the install again with `--allow-scripts=` followed by the package names that npm lists in its warning. The "Quick Start" note in the OmniRoute README shows an example.

B. Docker (no Node needed). The `127.0.0.1` part keeps it private to your computer:

```
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
```

C. From source (only if you want to read or change the code):

```
git clone https://github.com/cheaper-inference/OmniRoute.git
cd OmniRoute
npm install
npm run dev
```

The dashboard opens at **http://localhost:20128**.

2 First login: set a real password

- **npm or source install:** the starting password is the public word `CHANGEME`. Log in, then change it straight away at **Dashboard, Settings, Security**.
- **Docker (no password variable set):** the dashboard asks you to create a password on first visit.

> **Never leave the default password on a gateway that holds your API keys.**
> Don't expose port 20128 to your Wi-Fi network or the internet.

3 Connect your 3 providers

Open **Providers** in the sidebar, click **Add Provider**, and add **Gemini (Google)**, **Groq** and **OpenRouter** with the keys you created. Use the test button on each connection to confirm it works.

4 Create your OmniRoute key

Open **API Keys** in the dashboard (`http://localhost:20128/dashboard/api-manager`) and create a new key. **Copy it now, it is shown only once.** This key is for your coding tools. It is not the same as your provider keys.

5 Test it

```
curl http://localhost:20128/v1/models -H "Authorization: Bearer YOUR_OMNIROUTE_KEY"
```

You should see a list of the models from your connected providers.

6 Connect your coding tool

Settings for any OpenAI-compatible tool: Base URL `http://localhost:20128/v1`, API key = your OmniRoute key, Model `auto`.

CLAUDE CODE

```
omniroute launch --api-key YOUR_OMNIROUTE_KEY
```

This starts Claude Code already pointed at OmniRoute. It writes no config files.

CLINE (VS CODE)

Choose the **OpenAI Compatible** provider in Cline and enter the three settings above.

CONTINUE (VS CODE)

Add this to `~/.continue/config.yaml`, then pick **OmniRoute - Auto** in the chat pane:

```
  - name: OmniRoute - Auto
    provider: openai
    model: auto
    apiBase: http://localhost:20128/v1
    apiKey: YOUR_OMNIROUTE_KEY
```

Check that it is really routing

In the dashboard, open **Monitoring, Logs** to see each request your tool sends and which provider answered.

Stay safe, fix problems

Free tools are great for learning. A few habits keep them from becoming a problem.

Safety checklist

- **Keep it local.** Don't open the port to your Wi-Fi, a public tunnel or the internet. FreeLLMAPI is single-user and protected only by your unified key.
- **Free is not private.** Some free providers log prompts or may use them for training. Never paste secrets, `.env` files, API keys, passwords, personal data or private or assignment code.
- **Protect your keys.** Hide them in screen recordings. Delete and recreate any key that leaks.
- **Read before you pipe.** Don't run `curl ... | bash` without reading the script first.
- **Stay updated.** Use the latest release of both tools. A critical flaw (CVE-2026-88062) was reported in older OmniRoute versions.
- **Know the limits.** Great for practice and college projects. Not for production, client work or anything critical.

Troubleshooting

Problem	What to try
Tool can't connect	Is the app running? Check the port: FreeLLMAPI desktop `31415`, FreeLLMAPI Docker `3001`, OmniRoute `20128`.
Claude Code can't connect or shows a URL error	Use the address **without** `/v1`. Cline, Continue and most other tools **need** the `/v1` ending.
`400 active chain is empty` (FreeLLMAPI)	Open the **Fallback** page and enable some models in the active chain.
Limit or 429 errors	That provider has hit its cap. Wait, or add another provider. Daily quotas reset at UTC midnight (5:30 AM IST).
Answers get worse in the evening	The best free models run out first. Try `auto:smart` or come back after the reset.
Not sure requests go through the gateway	FreeLLMAPI: run `npx freellmapi doctor`. OmniRoute: check **Monitoring, Logs** in the dashboard.
OmniRoute install fails	Check `node -v` (22.22.2 or newer). On npm 11 or newer, re-run the install with the allow-scripts option (Part 2, step 1).
Docker can't reach providers	Often caused by a proxy or an IPv6-only network. See the troubleshooting docs in the project repo.

Quick reference

	FreeLLMAPI	OmniRoute
Dashboard	localhost:31415 (desktop) localhost:3001 (Docker)	localhost:20128
Base URL for most tools	http://localhost:3001/v1	http://localhost:20128/v1
Base URL for Claude Code	http://localhost:3001 (no /v1)	use `omniroute launch`
Your key	unified key, starts with `freellmapi-`	key from the API Keys page
Model name	`auto` (or `auto:smart`)	`auto`

Links

- **FreeLLMAPI:** github.com/tashfeenahmed/freellmapi · Downloads (Releases)
- **OmniRoute:** github.com/cheaper-inference/OmniRoute · npm · Docker Hub
- **Free keys:** Gemini · Groq · OpenRouter

This is a community guide based on the projects' official documentation, checked on 27 October 2026. It is not affiliated with FreeLLMAPI, OmniRoute, Google, Groq, OpenRouter, Anthropic or OpenAI. Features, models and free-tier limits change often, so always confirm on the official pages.