Eigenflow

GUIDE · 7 NOVEMBRE 2026

Claude Code sans abonnement

FreeLLMAPI pas à pas, limites comprises

PAR

Abyane Nabeebaccus

INSTAGRAM

@abyane_nb

ÉDITION

novembre 2026

Tu as commenté TOKENS, voici le guide. Le dépôt, github.com/tashfeenahmed/freellmapi, est sous licence MIT. Il comptait 48 679 étoiles le 7 novembre 2026 à 15 h 55, le jour de sa version v0.16.21.

FreeLLMAPI installe sur ton ordinateur un petit serveur qui parle le même format que l'API de Claude. Claude Code lui envoie ses requêtes comme à Anthropic, et le serveur les fait suivre aux offres gratuites de 34 fournisseurs. Ce sont donc d'autres modèles que Claude qui te répondent. On y voit un bon moyen d'essayer des modèles gratuits dans un outil que tu connais, à tenir loin du travail client.

On n'a pas lancé l'outil. Ce guide suit la documentation du dépôt, le code de sa commande npx et les pages officielles d'Anthropic et des fournisseurs, tous relus le 7 novembre 2026. Eigenflow n'a aucun lien avec l'auteur du dépôt ni avec Anthropic.

À savoir avant de lire

- Les réponses viennent de modèles gratuits choisis par le routeur (Gemini, GPT-OSS, DeepSeek, Qwen et d'autres). Anthropic ne prend pas en charge Claude Code branché sur d'autres modèles que Claude.
- Le dépôt réserve l'outil à un usage personnel, *not production* selon son README.
- Les 7,4 milliards de tokens par mois additionnent des quotas plafonnés à la minute et à la journée. Les meilleurs modèles s'épuisent dans la journée et repartent à minuit UTC, soit 1 h du matin à Paris en hiver.
- Google et Mistral peuvent se servir de tes requêtes gratuites pour améliorer leurs modèles. N'y mets ni code ni données client.
- La commande `setup-claude` lancée sans option modifie ton Claude Code habituel. `launch` et `--profile` le laissent intact.

1. Ce que fait FreeLLMAPI

Le serveur tourne sur ta machine et imite l'API Messages d'Anthropic (`/v1/messages`), celle qu'utilise Claude Code. Pour l'y brancher, on change l'adresse où Claude Code envoie ses requêtes (la variable `ANTHROPIC_BASE_URL`) et on lui donne la clé du routeur.

Pour chaque requête, le routeur prend le premier modèle de ta liste qui a une clé valide et du quota. Si le fournisseur répond « trop de requêtes » (code 429) ou tombe en erreur, il passe au suivant, jusqu'à 20 essais. Une conversation reste 30 minutes sur le même modèle. Les noms opus, sonnet et haiku mènent au choix automatique du routeur, ou au modèle que tu fixes dans l'onglet Anthropic de la page Keys.

Ce que veulent dire les 7,4 milliards de tokens

Le README annonce environ 7,4 milliards de tokens gratuits par mois, sur 34 fournisseurs et 635 points d'accès. C'est la somme des quotas gratuits de son catalogue. Chaque fournisseur plafonne son offre en requêtes par minute et par jour, parfois en tokens. Chez Groq, le modèle gpt-oss-120b s'arrête à 30 requêtes par minute, 1 000 requêtes et 200 000 tokens par jour, soit environ 6 millions de tokens par mois.

Le routeur ne compte que sur ce que tu branches, c'est-à-dire tes clés et quelques fournisseurs sans clé comme OVH (2 requêtes par minute) ou AI Horde (file d'attente). Avec trois clés, tu n'as qu'une petite partie de ce total.

Le routeur reste gratuit, sous licence MIT. Son auteur vend une option, Premium (19 $ par an ou 49 $ une fois), qui met à jour le catalogue des modèles le jour même. Une installation gratuite reçoit chaque nouveau modèle 30 jours plus tard.

---

2. Les limites, et ce que deviennent tes données

Le README l'écrit noir sur blanc : *This project is for personal experimentation and learning, not production.* Avant de livrer quoi que ce soit de réel, il conseille une API payante. Ses autres limites :

- Pas d'accès durable aux modèles de pointe. GPT-5.x, Grok 4.x, Kimi K3, DeepSeek V4 Pro ou Gemini 3.x figurent sur certaines offres gratuites, avec les plus petits quotas par jour.
- L'intelligence baisse dans la journée. Les modèles de tête (souvent Gemini 3.6 Flash, DeepSeek V4 et Kimi K2.6) atteignent leur plafond, et le routeur descend vers de plus petits jusqu'à minuit UTC.
- La latence varie. Groq répond très vite, d'autres beaucoup moins, et tu prends celui qui est libre.
- Aucune garantie de service, et des offres gratuites qui changent sans prévenir.

Ce que les fournisseurs font de tes requêtes

Pages officielles relues le 7 novembre 2026 :

- Google (AI Studio et API Gemini). Sur les services non payants, ses conditions du 23 mars 2026 lui permettent d'utiliser tes requêtes et les réponses pour améliorer ses produits et ses technologies d'apprentissage automatique. Des relecteurs humains peuvent les lire, d'où la consigne : *Do not submit sensitive, confidential, or personal information to the Unpaid Services.* Une appli ouverte à des utilisateurs de l'Espace économique européen, de Suisse ou du Royaume-Uni doit passer par les services payants.
- Mistral. En offre gratuite, sa page d'aide dit pouvoir entraîner ses modèles sur tes requêtes et ses réponses. Tu peux refuser dans la console d'administration, menu Privacy, option Anonymous improvement data (article du 3 septembre 2026).
- OpenRouter. D'après sa FAQ, rien n'est enregistré par défaut, et les fournisseurs qui gardent les données, ou dont la politique n'est pas confirmée, sont écartés tant que le réglage d'entraînement reste coupé.

- Groq. Pas de conservation par défaut, sauf des journaux jusqu'à 30 jours en cas de panne ou d'abus (Your Data in GroqCloud). La page ne parle pas d'entraînement.
- Cerebras. Sa politique de confidentialité du 27 août 2024 dit ne pas garder les requêtes ni les réponses de son service d'inférence. Elle ne parle pas d'entraînement non plus.

D'où notre règle. Pas de code ni de données client dans une offre gratuite, jamais. Garde ce montage pour des projets personnels, avec des données publiques ou inventées.

---

3. Installer FreeLLMAPI

L'appli de bureau

Sur la page Releases (v0.16.21 le 7 novembre), prends le .dmg arm64 pour un Mac Apple Silicon, le x64 pour un Mac Intel (macOS 12 minimum) ou l'installeur .exe pour Windows. D'après le dépôt, les versions Mac sont signées et notarisées.

L'appli ne demande ni compte ni mot de passe, et le tableau de bord s'ouvre depuis son icône dans la barre des menus (Open Dashboard). Le serveur de l'appli écoute sur 127.0.0.1:31415 par défaut, et son menu copie cette adresse et ta clé unifiée. Dans la suite, remplace http://localhost:3001 par cette adresse.

Docker, en une commande

Il faut Docker (ou Podman) installé et démarré. La commande du README :

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

Elle exécute un script téléchargé à la volée, et le dépôt propose lui-même de le lire avant. Télécharge-le, ouvre-le (ici dans TextEdit, sur Mac), puis lance-le :

```
À COLLER

curl -fsSL https://freellmapi.co/install.sh -o install.sh
open -e install.sh
bash install.sh
```

Le 7 novembre, ce script faisait 161 lignes, que nous avons lues sans les exécuter. Il vérifie Docker, crée ~/freellmapi avec une clé de chiffrement et un fichier docker-compose, puis démarre l'image ghcr.io/tashfeenahmed/freellmapi:latest sur 127.0.0.1:3001, une adresse joignable depuis ton seul ordinateur.

Ouvre ensuite http://localhost:3001 et crée le compte du tableau de bord (e-mail et mot de passe). Ta clé unifiée, qui commence par freellmapi-, s'affiche en haut de la page Keys. Claude Code ne verra que celle-là.

---

4. Créer tes clés gratuites

Le routeur se sert des clés que tu ajoutes sur la page Keys (bouton Add key). Trois offres simples, vérifiées sur leurs pages officielles le 7 novembre 2026 :

Groq

- Clé : console.groq.com/keys.
- Carte : la FAQ de facturation ne demande un moyen de paiement que pour passer de l'offre Free à l'offre Developer.
- Limite gratuite (page des limites) : pour gpt-oss-120b, 30 requêtes par minute, 1 000 par jour, 8 000 tokens par minute et 200 000 par jour, pour toute ton organisation.

OpenRouter

- Clé : openrouter.ai/keys.
- Carte : on ne l'a pas trouvé. La FAQ parle seulement d'une petite allocation gratuite à l'inscription.
- Limite gratuite (page Limits) : sur les modèles dont le nom finit par :free, 20 requêtes par minute et 50 par jour, puis 1 000 par jour une fois 10 crédits achetés.

Google AI Studio

- Clé : la page API Keys d'AI Studio, bouton Create API key.
- Carte : le niveau Free demande un projet actif, sans compte de facturation ; il en faut un pour passer au niveau 1 (page des limites).
- Limite gratuite : on ne l'a pas trouvée sur la page publique, qui renvoie à chaque projet dans AI Studio. Les quotas du jour repartent à minuit, heure du Pacifique.

Cerebras, que le README range encore parmi les gratuits, a changé ses règles. À la question d'une offre gratuite permanente, sa page des limites répond « No ». Un nouveau compte reçoit 5 $ de crédits valables 30 jours, après avoir ajouté un moyen de paiement vérifié.

---

5. Brancher Claude Code sans toucher à ton Claude habituel

La commande npx freellmapi se lance sans installation. Son paquet npm (version 0.8.2 le 7 novembre) demande Node.js 20.18 à 24. Sans l'option `--api-key`, elle te demande ta clé unifiée sans l'afficher, et la clé ne reste pas dans l'historique du terminal. Trois façons de faire, de la plus prudente à la plus envahissante. Commence par la première. Une session suffit pour juger les réponses.

Pour une session : launch

> npx freellmapi launch --url http://localhost:3001

Claude Code démarre branché sur le routeur, pour cette session seulement. Le README de la commande précise qu'aucun identifiant n'est écrit sur le disque, et son code retire de la session tes éventuelles clés Anthropic pour qu'elles ne partent pas vers le routeur. Ta configuration habituelle ne bouge pas.

Un profil séparé : setup-claude --profile

Pour garder le réglage, crée un profil à part. Avec `--dry-run`, la commande montre ce qu'elle écrira, sans rien modifier :

> npx freellmapi setup-claude --profile gratuit --url http://localhost:3001 --dry-run

Relance-la sans `--dry-run` pour appliquer. D'après son code, elle écrit `~/.claude/profiles/gratuit/settings.json` avec l'adresse du routeur et ta clé, lisible par toi seul. Ce profil a son propre dossier de réglages et d'historique. Pour l'ouvrir :

> env CLAUDE_CONFIG_DIR="$HOME/.claude/profiles/gratuit" claude

À éviter : setup-claude sans profil

Sans `--profile`, la commande écrit l'adresse du routeur, ta clé et un même modèle gratuit (à la place d'Opus, de Sonnet et de Haiku) dans `~/.claude/settings.json`, le fichier de ton Claude Code habituel. Elle copie d'abord l'ancien fichier (la ligne Backup donne le chemin de la copie). Toutes tes sessions passent ensuite par FreeLLMAPI, et d'après Anthropic, la valeur du fichier l'emporte sur une variable posée dans le terminal. Pour revenir en arrière, remets la copie en place.

Ce que dit Anthropic

La documentation de Claude Code appelle ce montage une passerelle LLM. Sa page Other LLM gateways précise : *Anthropic doesn't endorse, maintain, or audit third-party gateway products, and doesn't support routing Claude Code to non-Claude models through any gateway.* Anthropic ne cautionne ni ne vérifie ces passerelles, et ne prend pas en charge Claude Code branché sur d'autres modèles que Claude.

La même page ajoute que, tant qu'une clé de passerelle est active, Claude Code garde ta connexion claude.ai sans l'envoyer, et que les limites de ton abonnement ne s'appliquent pas à ces requêtes. Remote Control et la dictée vocale sont alors coupés (page de connexion). Pour vérifier où partent tes requêtes, tape cette commande dans Claude Code. La ligne Anthropic base URL doit afficher l'adresse du routeur.

> /status

---

6. Savoir quel modèle a répondu

Le modèle change au fil des quotas. Le routeur l'indique dans chaque réponse avec l'en-tête X-Routed-Via, au format fournisseur/modèle. Cette requête de test, adaptée de la documentation, affiche la ligne (remplace TA_CLE par ta clé unifiée) :

À COLLER

```
curl -si http://localhost:3001/v1/messages \
  -H "x-api-key: TA_CLE" \
  -H "anthropic-version: 2023-06-01" \
  -H "content-type: application/json" \
  -d '{"model": "claude-sonnet-4-5", "max_tokens": 64, "messages": [{"role":
"user", "content": "Bonjour"}]}' \
  | grep -i x-routed
```

Sans terminal, le Playground du tableau de bord affiche sous chaque message qui a répondu, et en combien de temps. La page Models montre l'ordre de passage du moment.

---

7. Les conditions d'utilisation, selon la revue du dépôt

En mai 2026, le dépôt a relu les conditions de chaque fournisseur pour un serveur personnel, à un seul utilisateur. Ni lui ni ce guide ne donnent d'avis juridique, alors vérifie les conditions de chacun.

- « Probablement OK » : Groq, Cerebras (qui interdit de vendre ou de céder ses clés), Mistral et OpenRouter (qui interdit la revente et les services concurrents).
- « Prudence » : Google Gemini, dont les conditions de mars 2026 visent un usage professionnel, *not for consumer use* ; NVIDIA NIM, offre d'essai *evaluation only, not production* ; Z.ai, pour une clause contre la redirection de trafic.
- « Ambigu » : Cloudflare Workers AI.
- « À éviter » : Cohere, dont les conditions excluent l'usage *personal, family or household*.

Ses règles de bon sens : un compte par fournisseur, pas de revente, un serveur pour toi seul, et aucune offre gratuite traitée comme un service de production.

---

8. Désinstaller

Sur Mac, d'après la FAQ du dépôt, coupe Start at login si tu l'avais activé, quitte l'appli (Quit FreeLLMAPI) et mets FreeLLMAPI.app à la Corbeille. Supprime ensuite le dossier `~/Library/Application Support/FreeLLMAPI/`, où restent tes clés et tes réglages, car la désinstallation n'y touche jamais. Sur Windows, passe par Paramètres puis Applications, et supprime le dossier FreeLLMAPI dans %APPDATA%.

Pour Docker, ces commandes arrêtent le conteneur, effacent sa base (option -v), puis retirent l'image et le dossier :

À COLLER

```
cd ~/freellmapi
docker compose down -v
docker image rm ghcr.io/tashfeenahmed/freellmapi:latest
cd ~ && rm -rf ~/freellmapi
```

Côté Claude Code, supprime `~/.claude/profiles/gratuit`, ou remets la copie de `settings.json` si tu avais lancé `setup-claude` sans profil. Supprime enfin tes clés sur les consoles de Groq, d'OpenRouter et de Google, car une clé effacée du routeur reste valide chez le fournisseur.

Sources

Pages relues le 7 novembre 2026.

- FreeLLMAPI : dépôt et README, installation, clients et agents, API, limites et revue des conditions, appli de bureau, commande npx, script d'installation, catalogue.
- Anthropic : Other LLM gateways, Connect Claude Code to an LLM gateway, variables d'environnement.
- Google : conditions de l'API Gemini, limites, clés d'API.
- Groq : limites, facturation, données.
- OpenRouter : limites, FAQ, collecte des données.
- Cerebras : limites, politique de confidentialité.
- Mistral : entraînement sur tes données, refuser l'entraînement.