



OmniRoute. Dein Modell-Backup.

Richte einen gemeinsamen Zugang für mehrere KI-Modelle ein. Wenn ein Anbieter ausfällt oder sein Kontingent erschöpft ist, kann eine verfügbare Alternative übernehmen.

01**Claude Code**

Du bleibst in deiner Arbeitssitzung.

02**OmniRoute**

Die Regel wählt einen verfügbaren Zugang.

03**Modell A / B**

Hauptmodell oder geeigneter Ersatz.

WAS DU AM ENDE EINGERICHTET HAST

Ein lokales Gateway, zwei Anbieterzugänge und eine Modellgruppe namens mein-backup. Dazu ein kleines Website-Beispiel und ein Test, der zeigt, welcher Anbieter antwortet.

Katalog: Das Projekt nennt 352 Anbieter-Einträge und 1.312 rohe Chat-Modell-IDs. Das sind keine 352 gleichwertigen Claude-Ersatzmodelle.

Kosten: Die Software ist frei. Modellaufrufe unterliegen den Preisen und Kontingenten deiner Anbieter.

Grenzen: Sind alle geeigneten Zugänge erschöpft, stoppt auch der Router. Unterbrechungsfreiheit ist nicht garantiert.

Gilt für Claude Code am Computer. Der normale Chat auf claude.ai bekommt dadurch keine zusätzlichen Claude-Nachrichten. Beim Wechsel kann ein anderes Modell antworten.

[OmniRoute auf GitHub](#) • Quellen: R • P • C



Installieren & starten.

VORBEREITUNG

Installiere [Node.js 24 LTS](#) und [Claude Code](#). Öffne danach ein neues Terminal: unter Windows PowerShell, auf dem Mac die App Terminal. Der reguläre Claude-Code-Zugang benötigt einen unterstützten Tarif oder API-Zugang.

1 • PRÜFEN • JEDE ZEILE EINZELN AUSFÜHREN

```
node --version  
npm --version  
claude --version
```

Alle drei Befehle sollen eine Versionsnummer ausgeben. Lege einen leeren Ordner omni-test an und öffne dort das Terminal.

2 • OMNIROUTE INSTALLIEREN UND STARTEN

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

Lass dieses Terminal geöffnet. Öffne im Browser <http://localhost:20128>. Folge beim ersten Start dem Setup und richte den angebotenen Administratorzugang ein.

Dein Kontrollpunkt: Das OmniRoute-Dashboard lädt. Das erste Terminal betreibt den Router; ein zweites Terminal ist später für Claude Code da.

Ein GitHub-ZIP allein installiert die Anwendung nicht. Der npm-Weg oben ist der dokumentierte Schnellstart. Du brauchst für dieses lokale Beispiel keinen eigenen Server.



Zwei Zugänge verbinden.



Für dieses Beispiel nutzen wir Groq und die Gemini API. Beide bieten begrenzte kostenlose Nutzung. Prüfe beim Anbieter, welches Modell für dein Konto verfügbar ist.

1 • GROQ

Öffne [Groq: API Keys](#), melde dich an und erstelle einen Schlüssel. Öffne in OmniRoute Providers, suche Groq und füge den Schlüssel als Verbindung hinzu.

2 • GEMINI

Öffne [Google AI Studio: API Keys](#). Erstelle einen aktuellen Gemini-API-Schlüssel für dein Projekt. Verbinde ihn in OmniRoute unter Providers → Gemini. Prüfe im Google-Projekt den Tarifstatus.

3 • JE EIN MODELL AUSWÄHLEN

Öffne bei beiden Anbietern Available Models. Wähle je ein für dein Konto nutzbares Textmodell mit Werkzeugaufrufen. Übernimm die Modell-ID exakt; wir nennen die Auswahl hier A = Groq und B = Gemini.

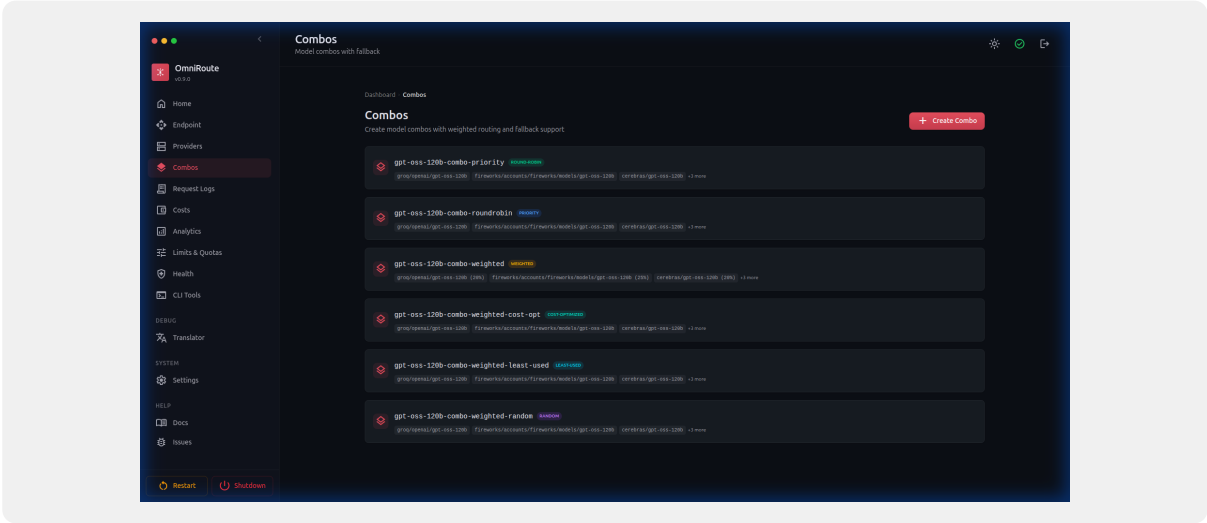
Für den Gratis-Test: Verwende nur Modelle mit verfügbarer Free-Tier-Quote. Aktiviere keine kostenpflichtige Abrechnung für dieses Beispiel.

Für jeden Zugang die angebotene Verbindungsprüfung ausführen. Ein gespeicherter Schlüssel allein beweist noch keinen erfolgreichen Modellaufruf. Aktuelle Limits: [Groq](#) und [Gemini](#). Schlüssel nur in den Verbindungsfeldern hinterlegen, nicht in Prompts.

[Groq-Einstieg](#) • [Gemini-Schlüssel](#) • Quellen: G • AI • GL • AP • P



Deine Ersatzreihenfolge.



Offizielle ältere Ansicht (v0.9.0): Modellgruppen unter „Compos“. Beispielnamen und Modell-IDs im Bild nicht übernehmen.

Öffne Compos → Create / Create New. Eine Combo ist eine benannte Gruppe von Modellen, die der Router nach einer Regel verwendet.

Name	mein-backup
Strategy	Priority: bevorzugtes Modell zuerst, Ersatz bei einem Fehler.
Models	1. Dein Modell A von Groq 2. Dein Modell B von Gemini

Speichere die Combo. Lass die Komprimierung für den ersten Test ausgeschaltet. Kontrolliere, dass beide Einträge tatsächlich auf deine verbundenen Modelle zeigen.

Claude als Hauptmodell? Dieses Gratis-Beispiel startet mit Groq. Ein Claude-Zugang muss separat eingebunden und an erste Stelle gesetzt werden. Dein Claude-Abo wird hier nicht automatisch übernommen.

[Compos nachlesen](#) • Quellen: U • C



Claude Code verbinden.

Lass OmniRoute laufen. Öffne ein zweites Terminal im Ordner omni-test. Starte Claude Code über den dokumentierten Launcher und wähle dabei deine Combo:

```
omniroute run claude --model mein-backup
```

Der Launcher setzt die Gateway-Verbindung für diesen Prozess. Dein Zielmodell heißt nun mein-backup; die Combo bestimmt den tatsächlich antwortenden Anbieter.

PRAXISAUFTRAG • IN DIE GESTARTETE CLAUDE-SITZUNG

```
Erstelle demo.html als lokale, responsive Ressourcenliste für KI-Einsteiger. Fiktive Beispieldaten: „E-Mails klarer schreiben“ (Text), „Website prüfen“ (Web), „Meeting-Notizen ordnen“ (Text). Baue Filter für Alle, Text und Web. Verwende HTML, CSS und JavaScript ohne externe Pakete. Speichere die Datei und erkläre, wie ich sie im Browser öffne.
```

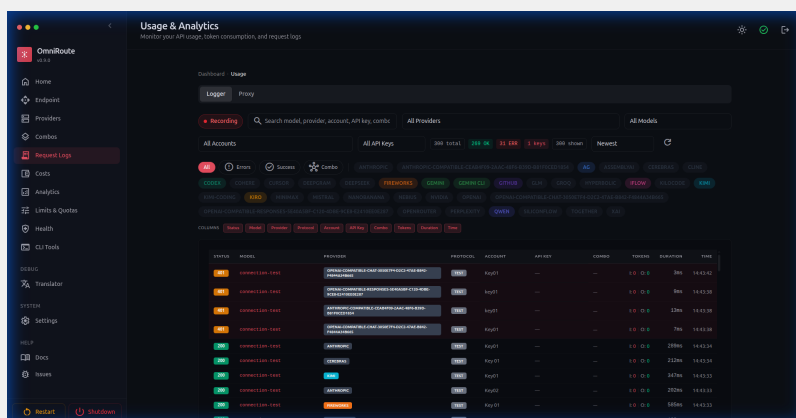
PRÜFEN

- ☐ demo.html existiert und öffnet sich im Browser.
- ☐ „Alle“ zeigt drei Einträge, „Text“ zwei und „Web“ einen.
- ☐ Im OmniRoute-Anfragelog siehst du die Anfrage, die Combo und das tatsächlich verwendete Modell.

Der Modellname in einer KI-Antwort ist kein zuverlässiger Nachweis. Nutze den Logeintrag. Scheitert schon dieser Test, behebe zuerst die Verbindung oder wähle ein kompatibleres Modell.



Den Wechsel prüfen.



Offizielle ältere Logansicht (v0.9.0). Status, Modell und Anbieter helfen bei der Prüfung; die Beispielaufträge sind keine eigenen Testergebnisse.

KONTROLLIERTER WECHSELTEST

1. Notiere den Anbieter des ersten Aufrufs. Deaktiviere dessen Verbindung unter Providers vorübergehend. Ersatzverbindung und Claude-Sitzung bleiben aktiv.
2. Sende in derselben Sitzung den folgenden Auftrag:

Ergänze in der bestehenden `demo.html` ein Suchfeld, das nach Titeln filtert. Behalte die drei Einträge und die Kategorien. Prüfe, dass die Suche „Website“ genau einen Eintrag findet.

3. Prüfe im Anfragelog, ob jetzt der Ersatzanbieter antwortet. Öffne die Datei erneut und teste die Suche. Aktiviere danach die erste Verbindung wieder.

Bestanden: Der Ersatzanbieter bearbeitet dieselbe Datei und die Suche funktioniert. Das bestätigt die Ersatzroute bei deaktiviertem Hauptzugang.

Dies ist kein künstlicher 429-Fehlertest. Ob ein echtes Anbieterlimit automatisch abgefangen wird, zeigt erst der entsprechende Fehler- und Folgeaufruf im Log. Wiederholungen können warten; Kontext und Werkzeugfunktionen können sich beim Wechsel unterscheiden.

[Routing & Anfragelog](#) • Quellen: U • C • GL



Weniger Tokens.

Komprimierung verkürzt geeigneten Kontext vor dem Modellaufruf, zum Beispiel lange, wiederholte Tool-Ausgaben. Ein kurzer, bereits präziser Prompt bietet viel weniger Einsparpotenzial.

Bis zu 90 % weniger Tokens ist keine pauschale 90-%-Ersparnis auf der Rechnung. Die Projektwerte beziehen sich auf geeignete Inhalte und bestimmte Komprimierungsverfahren.

SO VERGLEICHST DU SINNVOLL

Nutze nach dem erfolgreichen Routing-Test zwei Kopien desselben Projektstands. Führe denselben Auftrag mit demselben Modell einmal ohne Komprimierung und einmal mit Lite aus. Vergleiche Tokenwerte, Fehler und Ergebnisqualität.

Den Modus findest du laut Dokumentation unter Context & Cache → Caveman; pro Routing-Combo ist auch ein Override möglich. Prüfe vor dem Test, welche Einstellung für mein-backup wirksam ist.

RECHENBEISPIEL • FIKTIV, KEIN BENCHMARK

Ohne / mit	10.000 / 7.000 Input-Tokens
-------------------	-----------------------------

Ersparnis	$(10.000 - 7.000) \div 10.000 = 30 \%$
------------------	--

Die oft genannte Zahl „1,6 Millionen kostenlose Tokens pro Monat“ ist kein persönliches OmniRoute-Guthaben. Das Projekt aggregiert wechselnde Anbieter-Kontingente. Entscheidend ist die verfügbare Quote deiner eigenen Zugänge.

[Komprimierung erklärt](#) • [Gratis-Budgets erklärt](#) • Quellen: K • F



Wenn es noch hakt.

401 / 403

Prüfe im Log, welcher Zugang ablehnt. Anbieter-Schlüssel und OmniRoute-Zugang sind verschiedene Schlüssel.

429 / Quote voll

Prüfe den betroffenen Anbieter und die verfügbare Quote des Backups. Sind beide erschöpft, hilft nur ein weiterer geeigneter Zugang oder Warten.

Modell fehlt

Combo-Namen und aktuelle Modell-IDs prüfen. Ein Eintrag im Katalog garantiert keinen Zugriff mit deinem Konto.

Tools schlagen fehl

Teste ein anderes Modell mit Werkzeugunterstützung. Bei langen Sitzungen zusätzlich das Kontextlimit prüfen.

ZUR NORMALEN NUTZUNG ZURÜCK

Beende die über OmniRoute gestartete Claude-Sitzung mit /exit. Öffne ein frisches Terminal und starte claude wie zuvor. Manuell gesetzte Gateway-Umgebungsvariablen müsstest du zusätzlich entfernen.

Dein Setup steht: Anbieter verbunden, Combo gespeichert, ersten Aufruf und Ersatzroute geprüft. Starte die nächste echte Aufgabe über dieselbe Combo.

Recherche: 09.09.2026. Offizielle Dokumentation und Links geprüft; diese Anleitung wurde hier nicht mit echten Anbieterzugängen durchlaufen. Mehr verständliche KI-Praxis: [@willi_neubauer.ai](https://willi-neubauer.ai).