8 Tipps gegen Claude-Limits

Wie du mit dem gleichen Abo bis zu 70% mehr aus Claude rausholst. Konkrete Anleitungen — von offensichtlich bis „das wusste ich nicht“.

Warum du ständig ans Limit kommst

Claude rechnet nicht in Nachrichten — sondern in **Tokens**. Jedes Wort das du schickst, jedes Wort das Claude antwortet, jede Tool-Verbindung, jeder Datei-Inhalt — alles wird in Tokens umgerechnet. Ein langes Gespräch mit vielen Tools kann in 30 Minuten so viele Tokens verbrauchen wie 4 Stunden kurze Chats.

Das Limit ist also nicht fix — **es hängt davon ab, wie du Claude nutzt**. Und genau da setzen diese 8 Tipps an.

01

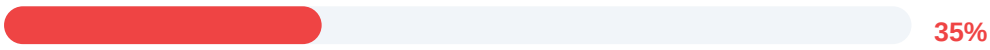
Extended Thinking ausschalten

Extended Thinking zwingt Claude bei **jeder Nachricht** in einen tiefen Denkprozess — auch wenn du nur fragst „fasse das in 2 Sätzen zusammen“. Das kann den Token-Verbrauch pro Nachricht verdoppeln bis verdreifachen. Bei vielen Nutzern ist es standardmäßig aktiviert, ohne dass sie es wissen.

So setzt du es um:

Gehe in die Claude-Einstellungen und deaktiviere Extended Thinking. Schalte es nur gezielt ein, wenn du komplexe Analysen, mehrstufiges Reasoning oder schwierige Code-Probleme lösen willst. Für 80% deiner Aufgaben brauchst du es nicht.

Einsparpotenzial:



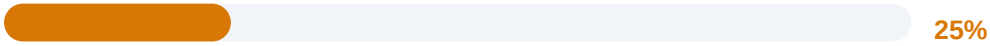
02

Unnötige MCP-Connectoren deaktivieren

Jeder verbundene Connector (Notion, Slack, Google Drive etc.) wird bei **jeder einzelnen Nachricht** komplett in den Kontext geladen — inklusive Tool-Definitionen, Authentifizierung und Schema. Auch wenn du den Connector in dieser Session gar nicht brauchst. 5 Tools verbunden = 5x zusätzlicher Token-Overhead pro Nachricht.

So setzt du es um:
Gehe in deine MCP-Einstellungen und setze alle Connectoren auf **„on demand“**, die du nicht in jeder Session brauchst. Aktiviere nur die 1–2 Tools, die du gerade wirklich nutzt. Bei Cowork: Entferne Plugins temporär, die du diese Woche nicht brauchst.

Einsparpotenzial:



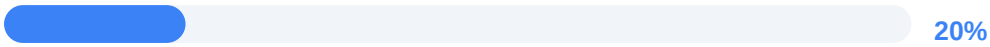
03

Neue Conversations statt Endlos-Threads

In einem langen Thread muss Claude bei jeder Nachricht die **komplette bisherige Conversation** erneut verarbeiten. Nach 30 Nachrichten liest Claude also 30x den gesamten Verlauf — und jedes Mal zählt das gegen dein Limit. Ein frischer Thread startet bei null.

So setzt du es um:
Starte eine neue Conversation sobald du das Thema wechselst oder ein Ergebnis abgeschlossen ist. **Faustregel: Maximal 10–15 Nachrichten pro Thread.** Danach neuen Thread öffnen und nur den relevanten Kontext übertragen.

Einsparpotenzial:



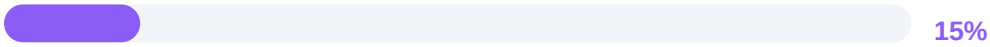
04

Peak-Zeiten meiden — das versteckte Throttling

Das wissen die wenigsten: Anthropic **drosselt dein 5-Stunden-Limit an Werktagen zwischen 14:00 und 20:00 Uhr** (deutsche Zeit / 5am–11am PT). Gleicher Plan, gleicher Preis — aber weniger Output pro Zeitfenster. Dein Wochen-Limit bleibt gleich, aber du erreichst es schneller.

So setzt du es um:
Plane deine intensivsten Claude-Sessions auf den **Morgen (7–12 Uhr)** oder **späten Abend (nach 21 Uhr)**. Am Wochenende gibt es kein Throttling. Wenn du unter der Woche nachmittags arbeitest, nutze diese Zeit für leichtere Aufgaben und spare die schweren für Off-Peak.

Einsparpotenzial:



05

Modell-Routing: Nicht alles braucht Opus

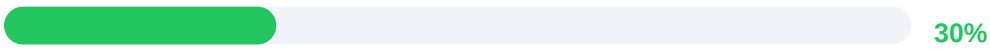
Opus 4.6 ist das teuerste Modell — und zieht am meisten vom Limit ab. Für einfache Aufgaben wie Zusammenfassungen, Formatierung oder kurze Fragen reicht **Sonnet oder Haiku** völlig aus. Die verbrauchen einen Bruchteil.

So setzt du es um:
Wechsle bewusst zwischen Modellen:

Opus 4.6 — Nur für: Komplexe Analysen, Coding, lange Dokumente
Sonnet 4.6 — Für: Skripte, Recherche, Alltagsaufgaben
Haiku 4.5 — Für: Formatierung, Zusammenfassungen, schnelle Fragen

Du kannst das Modell in Claude jederzeit oben im Chat wechseln.

Einsparpotenzial:



06

Tasks in Phasen aufteilen

Wenn du sagst „recherchiere das Thema, analysiere es, schreib ein Skript, erstelle eine PDF und fass alles zusammen“ — verbrennt Claude in einer einzigen Antwort das Budget von 5 normalen Nachrichten. Jeder Schritt erzeugt Output-Tokens, Tool-Calls und Reasoning.

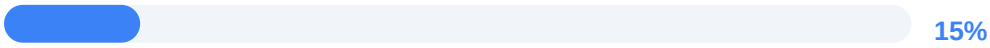
So setzt du es um:

Teile komplexe Aufgaben in **einzelne Schritte** auf:

Nachricht 1: „Recherchiere X“
Nachricht 2: „Schreib daraus ein Skript“
Nachricht 3: „Erstelle die PDF“

So hast du nach jedem Schritt Kontrolle und verschwendest keine Tokens für Ergebnisse, die du gar nicht brauchst.

Einsparpotenzial:



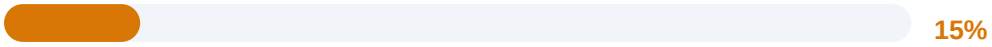
07

Context Files statt sich wiederholen

Wenn du Claude in jeder Session neu erklärst wer du bist, was dein Business macht und wie du arbeitest, verbrennst du jedes Mal hunderte Tokens für Kontext, den Claude eigentlich schon kennen sollte.

So setzt du es um:
Nutze **CLAUDE.md** und Memory-Dateien. Speichere dort deinen Brand-Kontext, Schreibstil, Zielgruppe und wiederkehrende Anweisungen. Claude liest diese Dateien automatisch — du musst sie nie wieder eintippen. Bei Cowork: Die Auto-Memory Funktion macht das teilweise automatisch.

Einsparpotenzial:



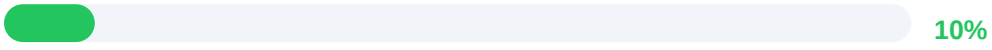
08

Extra Usage aktivieren (Team/Enterprise)

Wenn du auf einem Team- oder Enterprise-Plan bist, gibt es eine Option die viele übersehen: **Extra Usage**. Statt hart gekappt zu werden, kannst du nach Verbrauch weitermachen — Pay-as-you-go.

So setzt du es um:
Gehe in **Admin Settings** → **Billing** → **Extra Usage** und aktiviere die Option. Danach werden Überschreitungen zum Standard-Token-Preis abgerechnet. Keine Unterbrechung, kein Warten auf Reset. **Hinweis:** Nur verfügbar für Team und seat-based Enterprise, nicht für Pro.

Einsparpotenzial:



Cheatsheet: Alle 8 Tipps auf einen Blick

#	Tipp	Einsparung	Aufwand
1	Extended Thinking aus	**~35%**	5 Sek.
2	MCP-Connectoren reduzieren	**~25%**	2 Min.
3	Neue Threads ab 15 Nachrichten	**~20%**	0 Sek.
4	Peak-Zeiten meiden (14–20 Uhr)	**~15%**	0 Sek.
5	Modell-Routing (Haiku/Sonnet)	**~30%**	5 Sek.
6	Tasks in Phasen aufteilen	**~15%**	0 Sek.
7	Context Files nutzen	**~15%**	10 Min.
8	Extra Usage aktivieren	**Unbegrenzt**	2 Min.

Das Fazit

Du musst nicht auf Max upgraden oder weniger mit Claude arbeiten. Die meisten Nutzer verschwenden 50–70% ihres Limits durch Dinge, die sie in 5 Minuten ändern können. **Extended Thinking aus + Connectoren reduzieren + Modell-Routing** — allein diese drei Tipps machen den größten Unterschied.

Quellen: Claude Help Center, J.D. Hodges, LaoZhang AI Blog, PCWorld, TechRadar, The Register

Short Media — @tim_huck_ auf Instagram