

KI PRAKTISCH

Wenn die Nachfrage nach KI-Diensten explodiert, explodieren auch Ihre Kosten

Preisexplosionen bei KI-Diensten sind kein Zufall. GitHub Copilot erhöht Kosten um das 50-fache, Claude verschärft Limits, ChatGPT wandert hinter teurere Tiers.

AUTOR

Natascha Reiner

VERÖFFENTLICHT

5. Juni 2026

ONLINE LESEN

<https://wissen.strukturaflow.it.com/wenn-die-nachfrage-nach-ki-diensten-explodiert-explodieren-auch-ihre-kosten/>

Warum Preismodell-Änderungen bei KI-Diensten kein Ausnahmefall sind — und wie Sie als Unternehmen vorbereitet bleiben.

Das Geschäftsmodell dahinter — ein Fact zum Einstieg

KI ist kein Charity-Projekt. OpenAI, Anthropic, Microsoft, Google — sie alle subventionieren ihre Dienste in der Anfangsphase bewusst, um Marktanteile zu gewinnen. Günstige Preise, großzügige Free Tiers, scheinbar unbegrenzte Nutzung: Das ist Köder. Sobald die Abhängigkeit aufgebaut ist, beginnt das Lock-in. Ein Wechsel ist nicht in wenigen Minuten erledigt — Workflows, Integrationen, Mitarbeiter-Gewohnheiten hängen daran. Das ist kein Vorwurf, das ist Geschäftsmodell.

Das ist kein Einzelfall. Es ist ein Muster.

GitHub Copilot hat am 1. Juni 2026 von Flatrate auf Token-Verbrauch umgestellt. Power-User berichten von Kostensteigerungen zwischen dem 10- und 50-fachen — über Nacht. Die Begründung von GitHub: Die internen Infrastrukturkosten hatten sich seit Januar 2026 fast verdoppelt.

Das ist der aktuellste Fall. Aber bei weitem nicht der einzige.

Anthropic / Claude: Session-Limits für Claude Pro wurden im März 2026 still verschärft — Anthropic bestätigte offiziell, dass Limits während der Stoßzeiten schneller erschöpft werden. Rund 7 % der Pro-Nutzer treffen auf Limits, die es im Januar noch nicht gab. Wer Claude Code intensiv nutzt, braucht seit 2026 einen Max-Plan für 100 \$/Monat — fünfmal so teuer wie Pro. Zusätzlich führt ein neuer Tokenizer bei Opus 4.7 dazu, dass dieselben Eingaben bis zu 35 % mehr Tokens erzeugen — bei gleichem Preis pro Token.

OpenAI / ChatGPT: Was früher in ChatGPT Plus (20 \$/Monat) inklusive war, wandert sukzessive hinter teurere Tiers. Deep Research, Agent Mode, Codex, Sora: alles in Plus, aber mit engen Caps. Wer diese Features ohne Einschränkung nutzen will, landet bei Pro — 100 oder 200 \$/Monat. Einen detaillierten Copilot vs. ChatGPT Vergleich haben wir speziell für österreichische KMU zusammengestellt. OpenAIs Product-Chef Nick Turley hat öffentlich angekündigt, das Pricing werde sich „signifikant weiterentwickeln“ — und Unlimited-Pläne mit „unbegrenztem Strom“ verglichen: keine langfristige Option.

Google / Gemini: Google hat sein Gemini-Angebot mehrfach umstrukturiert. Der separate Gemini-Add-on für Workspace wurde eliminiert und in alle Workspace-Tiers zwangsbundled — inklusive Preiserhöhung. Gleichzeitig entstanden drei neue Abo-Stufen: AI Plus (9,99 \$), AI Pro (19,99 \$) und AI Ultra (249,99 \$). Features, die früher in einer Stufe verfügbar waren, verteilen sich heute über diese Tiers neu.

Das Muster ist immer dasselbe: günstig einsteigen, Abhängigkeit aufbauen, Preise anpassen.

Warnung: Das 500-Millionen-Dollar-Beispiel

Ein nicht namentlich genanntes Großunternehmen hat laut Axios-Recherche innerhalb eines einzigen Monats 500 Millionen US-Dollar für Anthropic Claude ausgegeben — weil keine Nutzungslimits für Mitarbeiter gesetzt wurden. Tausende Mitarbeiter liefen unkontrolliert, komplexe agentic Workflows fraßen Tokens in industriellem Maßstab. Resultat: finanzielle Schockstarre, erzwungene Notbremse.

Vergleichstabelle zeigt vier KI-Anbieter (GitHub Copilot, Anthropic Claude, OpenAI ChatGPT, Google Gemini) mit ihren früheren und aktuellen 2026-Preismodellen und Kostensprüngen.

Agentic KI vervielfacht das Kostenrisiko

Klassische KI-Nutzung: Eine Frage, eine Antwort. Überschaubar.

Agentic KI — also KI-Agenten, die selbstständig Aufgaben ausführen, Tools aufrufen, Code schreiben, Dateien durchsuchen, Ergebnisse prüfen und neu iterieren — ist eine andere Kategorie.

Goldman Sachs Research, Mai 2026

Agentic KI wird den globalen Token-Verbrauch bis 2030 um das 24-fache steigern — auf 120 Milliarden Tokens pro Monat. Bis 2040 könnten es das 55-fache sein. Der Treiber: Agenten arbeiten nicht auf eine Anfrage, sie laufen kontinuierlich, prüfen, iterieren, schreiben und korrigieren — mit deutlich höherem Token-Hunger als ein gewöhnlicher Chatbot.

Für Unternehmen bedeutet das: Wer agentic KI einführt ohne Governance-Struktur, riskiert Kostenexplosionen — nicht theoretisch, sondern nachweisbar.

Was Sie jetzt tun sollten

1. **Vertragstyp kennen:** Monatlich, jährlich oder usage-based? Was ist die Kündigungsfrist?
 2. **Budget-Caps setzen:** Jeder KI-Dienst, der es erlaubt — sofort aktivieren. Keine Ausnahmen.
 3. **Nutzung monitoren:** Wer verbraucht wie viel? Welche Workflows treiben die Kosten?
 4. **Alternativen kennen und getestet haben** — nicht erst, wenn die Rechnung kommt.
 5. **Nie 100 % Abhängigkeit von einem Anbieter.** Multi-Vendor oder zumindest Plan B.
-

Die wichtigste Erkenntnis

Nicht auf ein Pferd setzen. Nicht alle Ressourcen, alle Workflows, alle Mitarbeiter-Gewohnheiten an einen einzigen Dienst ketten. Das ist nicht Pessimismus — das ist unternehmerische Grundhygiene. Auch DSGVO-konforme KI-Nutzung gehört zu dieser Grundhygiene.

Wer heute nur einen KI-Dienst kennt, hat morgen ein Problem. Wer heute die Alternativen kennt, hat morgen die Wahl.

Wir besprechen die passende Strategie für Ihr Unternehmen. Verringern Sie die Abhängigkeit von einzelnen KI-Anbietern — strukturiert, pragmatisch, ohne monatelange Review-Schleifen.

NÄCHSTER SCHRITT

Mehr praktische KI-Anleitungen für KMU

Dieser Artikel ist Teil des KI-Hubs von Strukturaflow — einer deutschsprachigen Plattform für den praktischen KI-Einsatz in kleinen und mittleren Unternehmen.

<https://wissen.strukturaflow.it.com>