

# Copilot Chat: Kostenpflichtige Modelle über die Modellauswahl, Credits verteilt der Admin

*Der Modellwähler in Copilot Chat bekommt eine Kasse. Wer das für ein kleines UI-Update hält, hat noch keine Azure-Rechnung mit einem unerklärlichen Copilot-Posten gesehen.*

## Executive Summary

Microsoft schiebt das nächste Stück nutzungsbasierte Abrechnung direkt vor die Augen deiner Anwender. Laut Roadmap-Eintrag **571400** können Nutzer in Copilot Chat über die Modellauswahl künftig auf Modelle zugreifen, die nach Verbrauch abgerechnet werden. Das Ganze kommt für Web und Windows-Desktop, Zielmonat für die allgemeine Verfügbarkeit ist Oktober 2026, also jetzt. Der Haken steht gleich mit im Text: Damit jemand diese Modelle nutzen kann, muss der Tenant-Admin Credits zuweisen. Wer sein Kontingent leer geredet hat, darf beim Admin neue anfordern.

Klingt nach Komfortfunktion, ist aber ein Governance-Thema mit Preisschild. Ein Copilot Credit kostet im Pay-as-you-go-Modell **einen US-Cent**. Gesteuert wird das über Spending Policies im Cost-Management-Dashboard des Microsoft 365 Admin Centers, bezahlt erst aus Capacity Packs, dann aus einem Pre-Purchase-Plan und am Ende über ein verknüpftes Azure-Abo. Du musst also entscheiden, wer überhaupt Credits bekommt, wie viele, und wer Nachschub genehmigt, bevor der erste Vertriebsleiter mit dem teuersten Modell seine Weihnachtskarten textet.

### 1 Fakten: Das Wichtigste in drei Sätzen

**Was:** Nutzungsbasierte Modelle direkt im Modellwähler von Copilot Chat (Roadmap 571400, Web und Windows, GA-Ziel Oktober 2026).

**Wie:** Zugriff nur mit freigegebenen Credits; aufgebraucht heißt: beim Admin nachfordern.

**Womit:** Spending Policies unter Copilot > Cost Management, 1 US-Cent je Credit, Limits setzen am Monatsersten zurück.

## Worum geht es im Detail?

### Der Hintergrund: Warum Microsoft jetzt Maut erhebt

Frontier-Modelle sind teuer. Nicht ein bisschen teurer, sondern um Größenordnungen teurer als die Standardmodelle, die hinter »Auto« arbeiten. Microsoft hat lange versucht, das mit einer Pauschallizenz zu überdecken, und lernt gerade dieselbe Lektion wie jedes All-you-can-eat-Bufferet: Irgendwann kommt der Gast, der zehn Teller Hummer isst. Die Antwort heißt seit diesem Jahr Usage-based Billing. Am 21. September 2026 kündigte Microsoft mit **MC1476200** an, die nutzungsbasierte Abrechnung zu vereinfachen; Cowork ist seit Juni allgemein verfügbar und läuft

bereits über Credits. Der Modellwähler in Copilot Chat ist die logische nächste Station: Die Standardmodelle bleiben in der Lizenz, die Edelmodelle kosten extra.

Nach dem aktuellen Stand, den Marktbeobachter aus Microsofts Kostenmodell ableiten, zerfällt die Modellwahl in drei Klassen. Microsoft selbst hat die konkreten Grenzen noch nicht vollständig veröffentlicht, die Tabelle ist also eine Momentaufnahme und kein Vertrag.

| Modellklasse                  | Beispiele (Stand Oktober 2026)            | Abrechnung                                   | Was der Nutzer merkt                               |
|-------------------------------|-------------------------------------------|----------------------------------------------|----------------------------------------------------|
| Auto                          | Microsoft wählt das Modell                | in der Lizenz                                | nichts, außer gelegentlich mittelmäßiger Antworten |
| Inklusive mit Fair Use        | GPT-5.6, Sonnet 5                         | in der Lizenz                                | kaum etwas, solange niemand übertreibt             |
| Inklusive mit Nutzungsgrenzen | Opus 5                                    | in der Lizenz bis zur Grenze, danach Credits | Warnhinweis, dann Wahl: Auto oder zahlen           |
| Nur gegen Credits             | Fable, Astra und weitere Frontier-Modelle | nutzungsbasiert, 1 US-Cent je Credit         | ohne zugewiesene Credits: kein Zugriff             |

## Die Mechanik: Policy statt Sparschwein

Die Steuerung sitzt im Microsoft 365 Admin Center unter **Copilot > Cost Management**. Dort legst du Spending Policies an: Name, Geltungsbereich (alle Nutzer oder bestimmte Sicherheitsgruppen), welche Dienste Credits verbrauchen dürfen, ein monatliches Gesamtlimit, optional ein Limit pro Nutzer, Schwellenwerte für Warnmails und die Abrechnungsmethode. Microsoft verlangt dabei mindestens **2.000 Credits pro Nutzer und Monat**, also 20 US-Dollar. Darunter lässt sich die Policy schlicht nicht speichern. Wer am Limit ankommt, verliert den Zugriff auf die betroffenen Dienste bis zum Monatsersten.

Jetzt der Teil, über den in Workshops gestritten wird. Die Roadmap spricht davon, dass Admins Credits **zuweisen**. Die Microsoft-Dokumentation sagt ausdrücklich, dass Spending Policies Credits weder reservieren noch verteilen, sondern **Grenzen** setzen. Übersetzt: Du füllst keinem Mitarbeiter ein Sparschwein, du gibst ihm einen Dispo-Rahmen. Beahlt wird erst, was verbraucht wird. Das ist gut für die Liquidität und schlecht für alle, die gern mit »Wir haben doch Budget reserviert« argumentieren.

## Vom Klick zur Rechnung

Copilot Chat · nutzungsbasierte Modelle

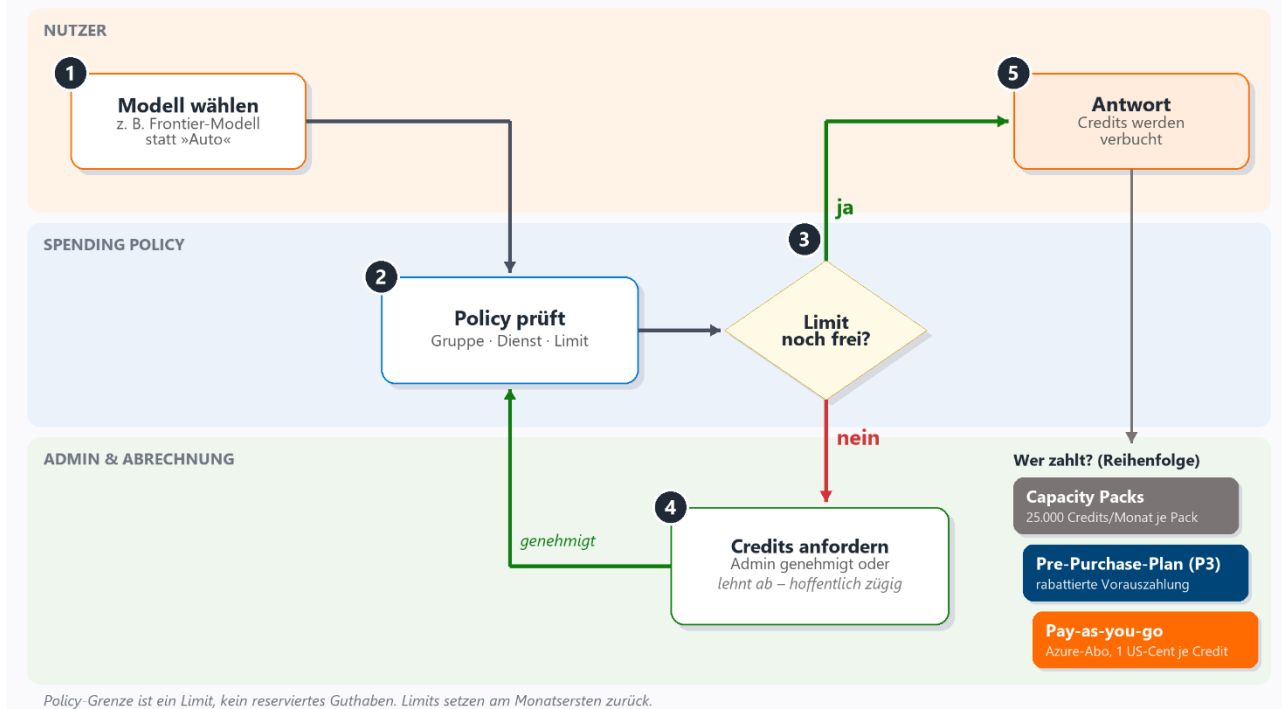


Abbildung 1: Vom Klick im Modellwähler bis zur Abrechnung, inklusive Umweg über den Admin.

Bezahlt wird in fester Reihenfolge: zuerst Capacity Packs (je 25.000 Credits pro Monat), dann ein Copilot Credit Pre-Purchase-Plan (P3) mit Rabatt, zuletzt Pay-as-you-go über das verknüpfte Azure-Abo. Beim Rest der Policy-Logik lohnt sich ein genauer Blick, denn sie verhält sich anders, als viele intuitiv erwarten:

| Situation                          | Was Microsoft tut                                                                                     | Was das in der Praxis heißt                                                            |
|------------------------------------|-------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------|
| Nutzer steckt in mehreren Policies | Es gewinnt die Policy mit dem höchsten Pro-Kopf-Limit, dann das höchste Gesamtlimit, dann die jüngste | Limits addieren sich nicht. Wer sich zwei Gruppen erschleicht, bekommt nur die bessere |
| Limit erreicht                     | Nutzer kann mehr Credits anfordern, fällt aber nicht in eine andere Policy                            | Ohne schnellen Genehmigungsweg steht der Nutzer bis zum Monatsersten                   |
| Wechsel der Entra-ID-Gruppe        | Neue Policy gilt sofort, der bisherige Verbrauch wandert mit                                          | Abteilungs-Hopping zum Limit-Reset funktioniert nicht                                  |
| Neue Dienste von Microsoft         | »Auto-apply new services« ist standardmäßig an                                                        | Neue Dinge hängen automatisch an deinem Budget                                         |
| Policy gelöscht                    | Verbrauch bleibt im Reporting, nichts wird erstattet                                                  | Löschen ist kein Rückgängig-Knopf                                                      |

### ★ Wichtig: Wer darf hier eigentlich was?

Global Admins und Billing Admins wählen und ändern die Abrechnungsmethode. AI Admins und License Admins dürfen Policies bearbeiten, Limits und Alarmer anpassen, aber keine neuen Policies anlegen. Reader-Rollen wie AI Reader oder Global Reader sehen nur den Verbrauch. Gib dem Controlling die Reader-Rolle und nicht den Global Admin, auch wenn es höflich fragt.

## Was sind Chancen? Was sind Risiken?

Fangen wir mit der guten Nachricht an, sie ist kürzer. Bisher hatten Unternehmen bei Premium-Modellen zwei Optionen: allen alles erlauben oder allen alles verbieten. Mit Credits im Modellwähler bekommst du endlich die Mittelstufe. Die Rechtsabteilung, die für Vertragsanalysen das stärkste Modell braucht, bekommt es. Der Kollege, der damit Out-of-Office-Nachrichten reimt, bekommt »Auto«. Und weil nur bezahlt wird, was verbraucht wird, kannst du teure Modelle in einer Pilotgruppe testen, ohne gleich eine Lizenzstufe für alle zu kaufen.

| Chancen                                                             | Risiken                                                                                 |
|---------------------------------------------------------------------|-----------------------------------------------------------------------------------------|
| Zugang zu Frontier-Modellen genau dort, wo der Mehrwert messbar ist | Unklare Kosten, weil niemand vorher weiß, was ein Prompt verbraucht                     |
| Kosten pro Gruppe sichtbar, Showback und Chargeback werden möglich  | Genehmigungsstau beim Admin, Nutzer weichen auf private KI-Konten aus                   |
| Pilotieren ohne Lizenzsprung für die ganze Belegschaft              | »Auto-apply new services« hängt Neues still an bestehende Budgets                       |
| Pro-Kopf-Limits bremsen Ausreißer automatisch                       | Limits greifen erst nach Verbuchung, das Reporting hinkt bis zu zwei Stunden            |
| Abrechnung über vorhandene Azure-Verträge und Rabatte (P3)          | Externe Modellanbieter arbeiten teils als eigenständige Verarbeiter, Datenschutz prüfen |

Das größte Risiko ist nicht die Technik, sondern die Mathematik. Rechne einmal mit 500 Nutzern durch, was die Obergrenzen bedeuten. Microsoft setzt bei neuen Käufen standardmäßig 4.000 Credits pro Nutzer und Monat an. Das sind 40 US-Dollar pro Kopf, also mehr als die Copilot-Lizenz selbst. Die Grafik zeigt, warum der Satz »Machen wir einfach ohne Limit, die Leute sind ja vernünftig« in die Kategorie berühmter letzter Worte gehört.

Ein Beispiel aus einem Projekt, anonymisiert, aber nicht erfunden: Ein Mittelständler hatte Cowork für eine Pilotgruppe freigeschaltet, ohne Pro-Kopf-Limit. Ein besonders begeisterter Produktmanager ließ an einem Freitagnachmittag ein Dutzend große Rechercheaufgaben parallel laufen, jeweils mit dem stärksten Modell. Unabhängige Tests messen für eine einzige große Cowork-Aufgabe zwischen 700 und gut 1.000 Credits, also sieben bis zehn Dollar. Das klingt harmlos, bis es zwölf Aufgaben pro Tag sind, zwanzig Arbeitstage lang. Am Montag kannte der CFO den Namen des Produktmanagers. Das war nicht als Auszeichnung gemeint.

## Was 500 Nutzer im Monat maximal kosten dürfen

Obergrenze pro Monat bei 1 US-Cent je Credit, im Vergleich zur Copilot-Lizenz (30 USD je Nutzer)



Rechenbeispiel: Obergrenzen, keine Prognose. Prepaid Credits werden zuerst verbraucht.

Abbildung 2: Maximales monatliches Kostenrisiko je Limitstufe für 500 Nutzer im Vergleich zur Lizenz.

### ⚠ Warnung: Der Dispo kennt keine Gnade

Ohne Pro-Kopf-Limit kann ein einzelner Nutzer das komplette Policy-Budget in wenigen Tagen verbrennen, und der Rest der Gruppe steht bis zum Monatsersten ohne Premium-Modelle da. Die Alarm-Mails kommen erst, wenn die Schwelle schon erreicht ist, und dann täglich. Das ist kein Frühwarnsystem, das ist ein Nachruf in Fortsetzungen.

Das zweite Risiko ist menschlich. Wer am Limit hängt, fordert beim Admin Credits an. Wenn diese Anfrage in einem Ticket-Queue landet, das im Schnitt neun Tage braucht, hat der Nutzer längst sein privates KI-Abo gezückt und kopiert Kundendaten hinein. Herzlichen Glückwunsch, du hast gerade Schatten-KI mit Steuergeld subventioniert. Ein Genehmigungsprozess, der langsamer ist als das Problem, ist kein Governance-Instrument, sondern ein Exit-Schild.

## Was müssen wir jetzt schon vorbereiten?

Der Zeitplan ist eng. Seit dem 19. Oktober ist die nutzungsbasierte Abrechnung bei neuen Direktkäufen von Microsoft 365 Copilot Business standardmäßig aktiv, ab dem 1. Dezember auch bei neuen Käufen über CSP-Partner. Bestehende Verträge ändert das nicht, aber spätestens bei der nächsten Lizenzaufstockung bist du drin.

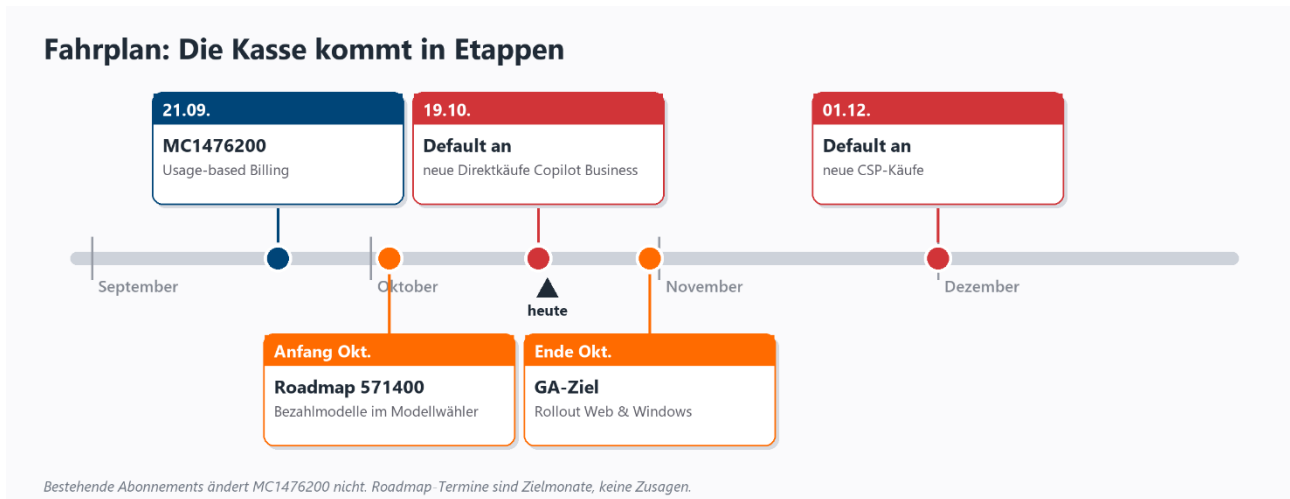


Abbildung 3: Die relevanten Termine rund um Credits und Modellwähler im Herbst 2026.

Damit du nicht vom Rollout überrascht wirst, arbeite diese Liste in der genannten Reihenfolge ab:

1. **Bestandsaufnahme:** Prüfe unter Copilot > Cost Management, ob bereits eine Default-Policy aktiv ist und welches Azure-Abo dranhängt.
2. **Modellanbieter klären:** Unter Copilot > Settings steuerst du die KI-Anbieter, die als eigenständige Verarbeiter arbeiten. Datenschutz und Betriebsrat vorher ins Boot holen, nicht hinterher.
3. **Gruppenmodell definieren:** Wenige Sicherheitsgruppen mit klaren Limits, keine Einzelfallzoos.
4. **Pro-Kopf-Limit setzen:** Immer. Ohne Ausnahme. Auch für den Vorstand, gerade für den Vorstand.
5. **Auto-apply prüfen:** In regulierten Umgebungen zunächst ausschalten und neue Dienste bewusst aufnehmen.
6. **Genehmigungsweg festlegen:** Wer entscheidet über Nachforderungen, in welcher Zeit, nach welchen Kriterien?
7. **Alarmer verteilen:** IT bei 50 und 80 Prozent, Controlling bei 80 und 100 Prozent.
8. **Kommunikation:** Den Nutzern erklären, warum ein Modell im Wähler plötzlich etwas kostet und was »Auto« leistet.

Ein bewährtes Startmodell sieht so aus. Die Werte sind Ausgangspunkte für die Diskussion, keine Naturgesetze:

| Gruppe           | Wer                             | Limit pro Nutzer und Monat | Max. Kosten pro Kopf | Nachforderung                   |
|------------------|---------------------------------|----------------------------|----------------------|---------------------------------|
| Standard         | alle Lizenznutzer               | 2.000 Credits (Minimum)    | 20 USD               | Selbstbedienung nein, Ticket ja |
| Fachspezialisten | Recht, Controlling, Entwicklung | 6.000 Credits              | 60 USD               | Teamleitung genehmigt           |
| Pilot Frontier   | max. 25 benannte Personen       | 15.000 Credits             | 150 USD              | IT und Fachbereich gemeinsam    |

### ✓ **Tipp: Genehmigung mit Service-Level**

Lege fest, dass Nachforderungen innerhalb eines Arbeitstages entschieden werden, und hinterlege drei einfache Kriterien: konkreter Anwendungsfall, erwarteter Nutzen, Ablaufdatum. Wer zweimal im Quartal nachfordert, gehört in die nächsthöhere Gruppe. Das spart dir mehr Tickets als jede Automatisierung.

### ★ **Wichtig: Reporting einplanen**

Richte einen festen monatlichen Termin ein, an dem IT und Controlling gemeinsam das Consumption-Dashboard durchgehen. Die Reader-Rollen erlauben das ohne Änderungsrechte. Wer erst bei der Azure-Rechnung hinschaut, betreibt keine Kostenkontrolle, sondern Archäologie.

## Häufig gestellte Fragen

### **Was kostet ein Copilot Credit in Microsoft 365 Copilot Chat?**

Im Pay-as-you-go-Modell kostet ein Copilot Credit einen US-Cent, 100 Credits entsprechen also einem Dollar. Günstiger wird es über Capacity Packs mit 25.000 Credits pro Monat oder über einen Pre-Purchase-Plan, die vor Pay-as-you-go verbraucht werden.

### **Brauche ich für kostenpflichtige Modelle in Copilot Chat eine zusätzliche Lizenz?**

Nein, eine eigene Lizenz pro Modell gibt es nicht; die Premium-Modelle werden zusätzlich über Credits abgerechnet. Voraussetzung ist, dass der Admin die nutzungsbasierte Abrechnung aktiviert und eine Spending Policy für den Nutzer oder seine Gruppe eingerichtet hat.

### **Was passiert, wenn mein Credit-Limit in Copilot Chat aufgebraucht ist?**

Du verlierst den Zugriff auf die kostenpflichtigen Modelle bis zum nächsten Monatsersten, an dem die Limits zurückgesetzt werden. Bis dahin kannst du beim Admin zusätzliche Credits anfordern oder mit »Auto« und den inkludierten Modellen weiterarbeiten.

### **Wo legt der Administrator Credit-Limits für Copilot fest?**

Im Microsoft 365 Admin Center unter Copilot > Cost Management auf der Registerkarte Configuration. Dort entstehen Spending Policies mit Geltungsbereich, Gesamtlimit, Pro-Kopf-Limit, Warnschwellen und Abrechnungsmethode.

### **Kann ein Nutzer sein Credit-Limit durch einen Gruppenwechsel zurücksetzen?**

Nein. Wechselt ein Nutzer im laufenden Monat die Entra-ID-Gruppe, gilt zwar sofort die neue Policy, sein bisheriger Verbrauch wird aber mitgenommen und auf das neue Limit angerechnet.