

GPT-6 Astra hält Einzug in Microsoft Copilot

Executive Summary

Anfang September 2026 hat Microsoft OpenAIs neues Spitzenmodell GPT-6 Astra in die Copilot-Welt geschoben – und zwar nicht als Preview mit Feigenblatt, sondern als allgemeine Verfügbarkeit in Microsoft Foundry, in Copilot Cowork, in Copilot Studio und in GitHub Copilot. Das ist eine andere Hausnummer als die üblichen Modellwechsel. Astra ist nicht für Chat gebaut, sondern für das Abarbeiten langer Aufgabenketten quer über Anwendungen hinweg, inklusive Computer Use: Das Modell liest Bildschirmhalte und bedient freigegebene Oberflächen. Das eröffnet endlich einen Weg zu den Altanwendungen ohne API, bei denen seit Jahren Menschen Felder abtippen.

Die Zahlen dahinter sind ordentlich. Auf Mind2Web erledigt Astra Aufgaben rund 1,9-mal schneller als der Vorgänger GPT-5.6 Sol – 72,6 Prozent gelöste Aufgaben in etwa 40 Minuten gegenüber 65,7 Prozent in rund 75 Minuten. Das Kontextfenster liegt bei rund 1,05 Millionen Token, und der Recall bleibt selbst zwischen 512K und 1M bei 96,3 Prozent, während Sol dort auf 73,8 Prozent abrutscht. Die Halluzinationsrate ist bei maximaler Denkstufe von 92 auf 51 Prozent gefallen, bei gleichzeitig vier Punkten mehr Genauigkeit.

Und die Rechnung? Astra kostet je Token das 2,5-Fache von Sol, braucht aber im Schnitt nur ein Drittel der Ausgabe-Token für dasselbe Ergebnis. Unter dem Strich landest du bei rund 32 Prozent geringeren Kosten pro erledigter Aufgabe. Rechne also nie in Token-Preisen, sondern in Kosten pro fertigem Arbeitsergebnis. Wer das verwechselt, führt entweder die falsche Sparsamkeitsdebatte oder unterschreibt ahnungslos die falsche Rechnung.

Der eigentliche Handlungsdruck liegt nicht bei der Qualität, sondern bei der Governance. Astra ist in Foundry hinter einem Limited Access Program eingesperrt, es gibt zum Start keine EU Data Zone, und in GitHub Copilot ist das Modell in vielen Tenants schlicht automatisch angekommen, weil neue Modelle standardmäßig aktiviert werden. Kurz: Das Modell ist schneller in deinem Tenant als deine Freigabeschleife.

FAKTEN · Die harten Eckdaten in einem Absatz

Allgemein verfügbar seit dem 3./4. September 2026 in Microsoft Foundry, Copilot Cowork, Copilot Studio und GitHub Copilot.

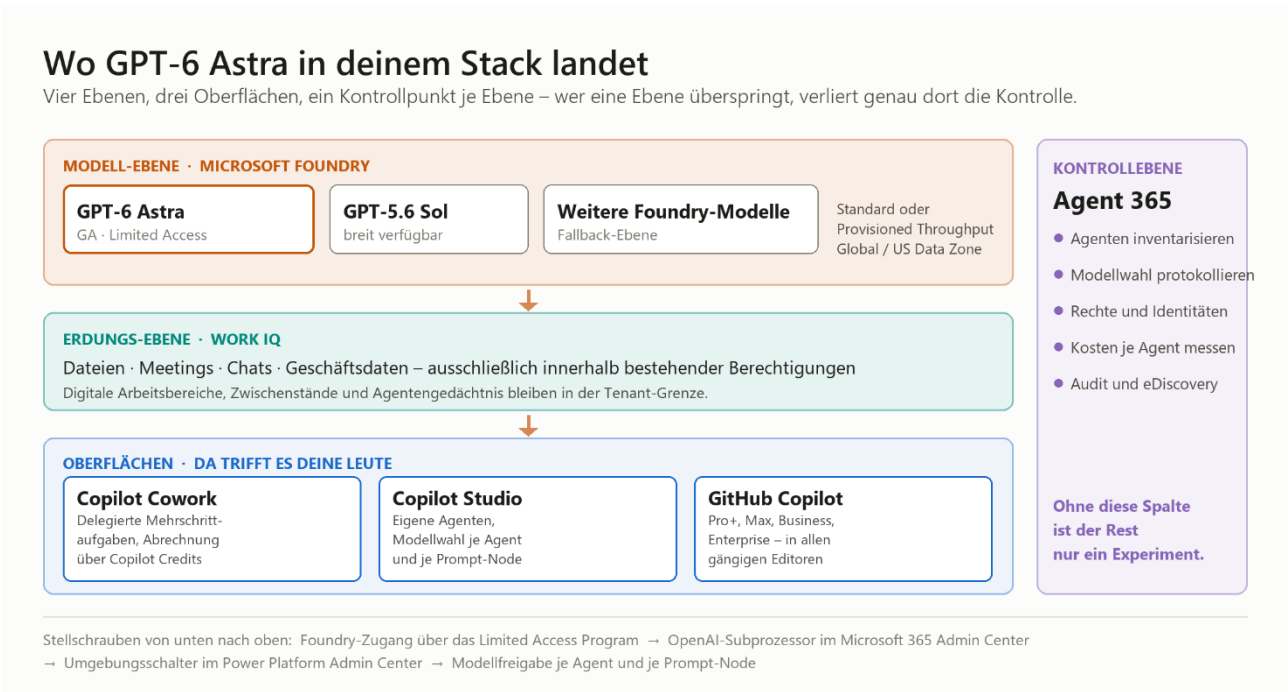
Kontextfenster rund 1,05 Millionen Token, maximale Ausgabe 128.000 Token, Denkstufen von low bis max.

Foundry Standard Global: 10 US-Dollar Eingabe und 50 US-Dollar Ausgabe je Million Token; US Data Zone rund 10 Prozent Aufschlag.

Erste Referenzkunden auf Azure sind unter anderem Albertsons und Replit.

Worum geht es im Detail?

Um zu verstehen, warum dieser Modellwechsel anders ist als die letzten fünf, musst du dir anschauen, wo Astra überhaupt einsteigt. Es gibt nicht den einen Schalter. Es gibt vier Ebenen, und jede hat ihren eigenen Kontrollpunkt in einem anderen Admin-Portal – was Microsoft vermutlich für Modularität hält und dein Betriebsteam für einen Running Gag.



Vier Ebenen vom Modell bis zur Oberfläche – und rechts die Kontrollebene, die in der Praxis am häufigsten fehlt.

Ganz unten sitzt Microsoft Foundry als Modell-Ebene. Dort wählst du zwischen Standard, also nutzungsabhängiger Abrechnung, und Provisioned Throughput, also reservierter Kapazität mit vorhersagbarer Latenz. Genau diese Wahl ist der Latenzhebel, den viele übersehen: Wenn ein Agent im Kundenservice hängt, weil die Standard-Kapazität gerade woanders gebraucht wird, hilft dir kein besseres Modell.

Darüber liegt Work IQ. Das ist die Schicht, die Astra in deinen Dateien, Meetings, Chats und Geschäftsdaten erdet – und zwar ausdrücklich innerhalb der bestehenden Berechtigungen. Wer seine SharePoint-Rechte nie aufgeräumt hat, bekommt jetzt einen unbestechlichen Prüfer geschenkt, der in Sekunden findet, was zehn Jahre lang niemand gesucht hat. Das ist kein Copilot-Problem, das ist ein Berechtigungsproblem mit neuer Beleuchtung.

Darüber kommen die Oberflächen. Copilot Cowork nimmt delegierte Mehrschrittaufgaben entgegen und rechnet über Copilot Credits ab. Copilot Studio erlaubt die Modellwahl je Agent und sogar je Prompt-Node. GitHub Copilot bringt Astra in Pro+, Max, Business und Enterprise, quer über VS Code, Visual Studio, JetBrains, Xcode, Eclipse, CLI, github.com und die Mobil-Apps. Und ganz rechts steht Agent 365 als Kontrollebene, mit der du Agenten beobachtest, steuerst und absicherst.

Oberfläche	Was Astra dort tut	Dein Kontrollhebel
Copilot Cowork	Delegierte Mehrschrittaufgaben über OneDrive, SharePoint, Mail und Kalender	Ausgabenrichtlinie plus Freigabe von OpenAI als Subprozessor im

		Microsoft 365 Admin Center
Copilot Studio	Modellwahl je Agent und je Prompt-Node, eigene Werkzeuge und Konnektoren	Umgebungsschalter im Power Platform Admin Center unter Einstellungen, Produkt, Features
GitHub Copilot	Lange Programmieraufgaben, Planung, Selbstprüfung vor der Fertigmeldung	Copilot-Einstellungen der Organisation – Achtung, neue Modelle sind per Vorgabe aktiv
Microsoft Foundry	Eigene agentische Pipelines, Computer Use, Werkzeugaufrufe	Limited Access Program, Deployment-Typ und gewählte Datenregion

WARNUNG · Das Modell war schneller da als dein Change Advisory Board

In GitHub Copilot werden neue Modelle automatisch aktiviert, solange kein Administrator den globalen Standard abgeschaltet oder das Modell explizit gesperrt hat.

Der Schalter für OpenAI als Microsoft-Subprozessor ist in vielen Tenants bereits im Juli automatisch aktiviert worden.

Wenn du also glaubst, Astra sei noch nicht im Haus: prüfe es, statt es zu glauben. Ein Kunde aus dem Maschinenbau hat im ersten Review vierzehn produktive Agenten gefunden, von denen laut Dokumentation genau drei existierten.

Ein Wort zum Hintergrund, denn ohne den versteht man die Preisgestaltung nicht: OpenAI hat Astra nicht auf Gesprächsqualität optimiert, sondern auf Token-Effizienz bei langen Aufgabenketten. Bei maximaler Denkstufe braucht das Modell etwa ein Drittel der Ausgabe-Token vergleichbarer Spitzenmodelle bei gleichem Ergebnis – 27.000 statt 78.000 Token in einer unabhängigen Messung. Genau deshalb darf der Token teurer sein. Der Regler dafür heißt Denkstufe und kennt die Werte low, medium, high, xhigh und max. Jede Stufe nach oben kauft dir Gründlichkeit und bezahlt sie mit Denk-Token, die als Ausgabe abgerechnet werden.

Der zweite Hintergrund ist unangenehmer und erklärt das Limited Access Program: Astra erreicht in der Kategorie Cyberfähigkeiten die höchste Risikostufe des OpenAI-Sicherheitsrahmens. Das Modell ist schlicht gut genug im Finden und Ausnutzen von Schwachstellen, dass es nicht voraussetzungslos an jeden ausgegeben wird. Die Gegenmaßnahme sind deutlich schärfere Ablehnungen – rund 91,5 Prozent bei unzulässigen Anfragen gegenüber 59 Prozent bei den Vorgängern. Für dich heißt das zweierlei: Erstens ist dieser Zugang ein Privileg, das entzogen werden kann. Zweitens wirst du in legitimen Sicherheitsszenarien häufiger auf Verweigerungen stoßen, als du es gewohnt bist.

Inhaltlich lohnt der Blick auf die Benchmarks, aber bitte mit gesundem Misstrauen. Astra ist kein pauschal besseres Modell, sondern ein deutlich besserer Handwerker bei langen, werkzeugintensiven Aufgaben. Bei reinem Weltwissen gibt es sogar Rückschritte – auf Humanity's Last Exam kommt Astra auf 57,2 Prozent und liegt damit hinter dem eigenen Vorgänger mit 65,0 Prozent. Das ist kein Skandal, sondern eine Prioritätenentscheidung: Ausführung schlägt Quizshow.

Prüfung	GPT-6 Astra	GPT-5.6 Sol	Was dir das sagt
---------	-------------	-------------	------------------

Mind2Web	72,6 % in ca. 40 Min.	65,7 % in ca. 75 Min.	Besser und fast doppelt so schnell bei Web-Aufgaben
Terminal-Bench v4.0	56 %	37 %	Der größte Sprung überhaupt – Werkzeugnutzung
DeepSWE v1.1	74,1 %	72,7 %	Solider, aber kein dramatischer Vorsprung
Langkontext-Recall 512K-1M	96,3 %	73,8 %	Große Dossiers sind endlich wirklich nutzbar
Halluzination bei max. Denkstufe	51 %	92 %	Halbiert, bei vier Punkten mehr Genauigkeit
Humanity's Last Exam	57,2 %	65,0 %	Kein Allwissender – Erdung bleibt Pflicht

Was sind Chancen? Was sind Risiken?

Die Chance ist unangenehm konkret: Aufgaben, die du bisher in mundgerechte Häppchen zerlegt und Schritt für Schritt begleitet hast, kannst du am Stück abgeben. Der Monatsbericht, der aus sieben Quellen zusammengeklaut wird. Der Fehlerbericht, der reproduziert, gefixt und als Pull Request eingereicht wird. Das Power-BI-Dashboard, das jemand baut, der eigentlich Controlling macht. Deine Rolle verschiebt sich vom Bedienen zum Prüfen – und das ist die eigentliche Umstellung, nicht die Technik.

Ein Praxisbeispiel aus einem Versicherungsprojekt im Sommer: Die Schadenabteilung hat ein Bestandssystem aus den Neunzigern, ohne API, mit einer Oberfläche, die nur ein Terminalemulator liebt. Seit Jahren tippen dort vier Leute Werte aus PDF-Dateien ab. Computer Use löst das. Und genau hier fängt das Risiko an, denn was ein Agent auf dem Bildschirm liest, kann ihn auch steuern.

WARNUNG · Prompt Injection hat jetzt eine Benutzeroberfläche

Sobald ein Agent Bildschirminhalte liest, um zu handeln, können dargestellte Inhalte sein Verhalten beeinflussen. Tool Poisoning zieht damit in die grafische Oberfläche um.

Ein manipuliertes PDF, ein präparierter Ticket-Kommentar, ein freundlicher Satz in einem Formularfeld – es braucht keinen Exploit, nur eine Zeile Text an der richtigen Stelle.

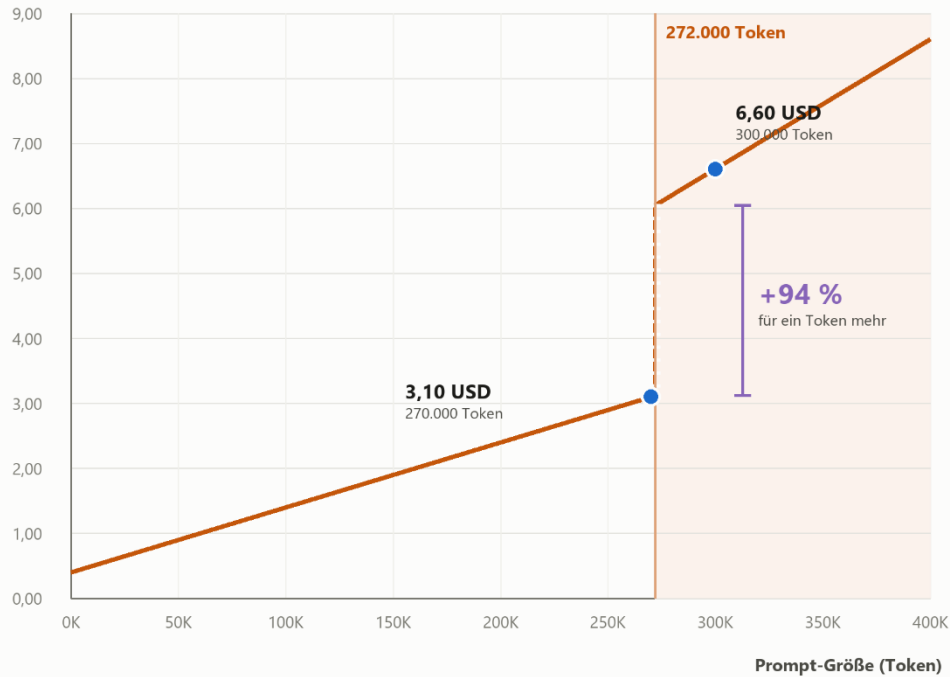
Konsequenz: Vor jedem folgenreichen System gehört ein menschlicher Freigabepunkt. Nicht als Papiertiger, sondern als technischer Stopp.

Das zweite Risiko ist banaler und trifft dich trotzdem härter: die Rechnung. Astra bepreist Requests oberhalb von 272.000 Token komplett neu – nicht nur den Überhang. Eingabe verdoppelt sich, Ausgabe steigt um die Hälfte, und zwar rückwirkend für den ganzen Aufruf. Ein Request mit 270.000 Token kostet rund 3,10 US-Dollar, derselbe Request mit 300.000 Token rund 6,60 US-Dollar.

Die 272K-Klippe: was ein einziges Token extra kostet

Kosten eines Foundry-Requests auf Standard Global, gerechnet mit 8.000 Ausgabe-Token · Angaben in US-Dollar

Kosten je Request (USD)



WARUM DER SPRUNG?

Überschreitet ein Request die Grenze von 272.000 Token, wird nicht nur der Überhang teurer, sondern der komplette Request neu bepreist:

Eingabe 10 → 20 USD/Mio
Ausgabe 50 → 75 USD/Mio

Chunking ist hier kein Schönheitswettbewerb, sondern Budgetschutz.

ZUM VERGLEICH

GPT-5.6 Sol liegt bei 4 USD Eingabe und 20 USD Ausgabe je Million Token.

Astra kostet je Token rund das 2,5-Fache – braucht dafür aber deutlich weniger davon.

Der Sprung an der 272K-Grenze trifft den kompletten Request, nicht nur den Überhang – Chunking ist hier reiner Budgetschutz.

Preisposition (Standard Global)	Bis 272K Token	Über 272K Token
Eingabe je Mio. Token	10,00 USD	20,00 USD
Zwischengespeicherte Eingabe	1,00 USD	2,00 USD
Cache-Schreibvorgang	12,50 USD	25,00 USD
Ausgabe je Mio. Token	50,00 USD	75,00 USD
US Data Zone	rund 10 % Aufschlag	rund 10 % Aufschlag

Dazu kommt in Copilot Cowork eine zweite Kostenlogik. Dort wird nicht nur die Modellnutzung abgerechnet, sondern auch die Kontextbeschaffung, die Werkzeugaufrufe und die Laufzeit – umgerechnet in Copilot Credits zu einem US-Cent je Credit. Die Denkstufe von Light bis Max ist damit kein Qualitätsregler, sondern ein Kostenregler. Setze sie nicht global auf Max, nur weil es sich gründlicher anfühlt.

Wie das schiefgeht, hat mir im Juli ein Mittelständler vorgeführt. Eine gut gemeinte Automatisierung hat jeden Morgen für vierzig Projekte einen Statusbericht erzeugt, jeweils mit maximaler Denkstufe und mit dem kompletten Projektordner im Kontext, weil Auswählen ja Arbeit gewesen wäre. Nach elf Tagen kam die Frage, warum die Copilot-Credits schon aufgebraucht seien. Die Antwort: Nicht das Modell war teuer, sondern die Großzügigkeit. Dieselbe Automatisierung mit zugeschnittenem Kontext und mittlerer Denkstufe lief anschließend für einen Bruchteil – und die Berichte waren sogar besser lesbar, weil weniger Rauschen drin war.

Dass das Ganze trotzdem trägt, zeigen die ersten Referenzen auf Azure. Albertsons und Replit stehen dort als frühe Anwender, und die zitierte Begründung ist bemerkenswert unsexy: Es gehe um das Gleichgewicht aus Geschwindigkeit und Kontrolle, um gleichbleibende Sicherheits- und Betriebsregeln beim Hochskalieren. Das ist keine Marketingpoesie, sondern genau der Punkt. Wer Astra als Produktivitätsbeschleuniger einkauft und die Kontrollseite weglässt, kauft sich vor allem schnelleren Ärger.

WICHTIG · Kein EU-Datenraum zum Start – das ist kein Detail

Astra startet ausschließlich in Global Standard und in der US Data Zone. Global routet weltweit, die US-Zone bleibt in den USA. Beides erfüllt keine EU-Datenresidenzanforderung.

Für Workloads mit harter EU-Anforderung bleibt vorerst GPT-5.6 Sol in der EU Data Zone die Option. Für die spätere EU-Zone wird ein Aufschlag von rund 20 Prozent erwartet, nicht die bisherigen 10.

In Copilot Cowork müssen Tenants im EU-Datenraum die Funktion ohnehin manuell aktivieren. Dieser Klick ist eine Entscheidung, kein Häkchen.

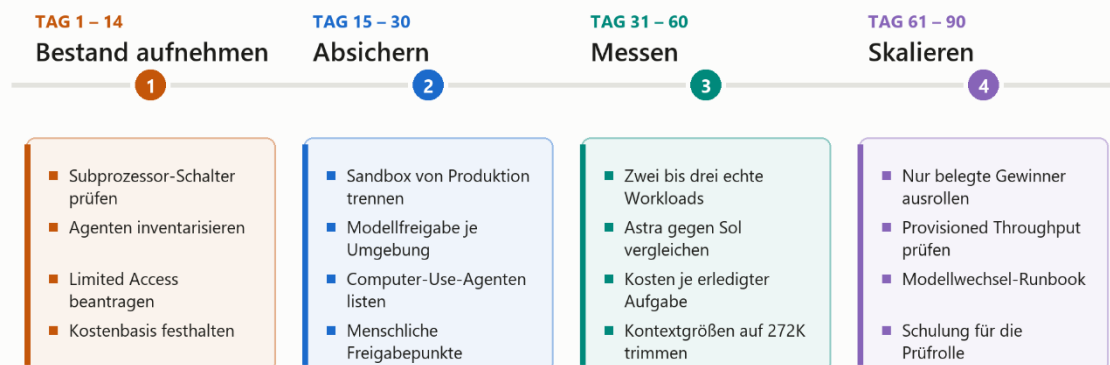
Und dann ist da noch das Tor davor. Astra liegt in Foundry hinter einem Limited Access Program, weil das Modell in der Cyber-Kategorie die höchste Risikostufe erreicht. Die Prüfung dauert fünf bis zehn Werktage, und Anträge von privaten Mailadressen werden abgelehnt. Das ist kein Self-Service. Plane den Antrag also nicht für den Tag, an dem der Proof of Concept starten soll – sonst wird es ein Proof of Waiting.

Was müssen wir jetzt schon vorbereiten?

Die unbequeme Wahrheit lautet: Modellversionen erreichen produktive Tenants inzwischen schneller, als Governance-Zyklen laufen. Du kannst das beklagen oder du kannst deine Zyklen umbauen. Der folgende Fahrplan ist bewusst auf 90 Tage angelegt – lang genug für belastbare Messungen, kurz genug, dass dir die Fachbereiche nicht in der Zwischenzeit ihre eigene Schatten-KI bauen.

90 Tage bis zur verantwortbaren Astra-Nutzung

Nicht schneller als nötig, aber keinen Tag später – die Modelle sind schon im Tenant, während die Freigabe noch im Umlauf ist.



Abbruchkriterium für jede Phase: Wenn du nicht benennen kannst, welcher Agent mit welchem Modell was tun darf, geht es nicht in die nächste Phase.

Vier Phasen, jede mit einem Abbruchkriterium – Tempo entsteht durch Reihenfolge, nicht durch Weglassen.

In den ersten beiden Wochen geht es ausschließlich um Bestandsaufnahme. Prüfe den Subprozessor-Schalter, statt ihn vorauszusetzen. Erfasse alle Agenten inklusive der Frage, welches Modell an welchem Prompt-Node hängt, denn genau das bestimmt Kosten, Verhalten und Risikoprofil. Und stelle den Antrag für das Limited Access Program, damit die Wartezeit parallel läuft.

In Phase zwei trennst du Sandbox und Produktion auf Umgebungsebene. Das ist der einzige Ort, an dem du Modellwahl wirklich durchsetzen kannst, ohne jede Nacht zu prüfen, ob Microsoft dir über Nacht ein neues Standardmodell untergeschoben hat. Parallel listest du alle Agenten mit Computer Use auf und definierst, vor welchem System ein Mensch zustimmen muss.

TIPP · Miss in erledigten Aufgaben, nicht in Token

Nimm zwei bis drei echte Arbeitsabläufe, keine Demos. Lasse Astra und Sol dieselbe Aufgabe machen und miss drei Dinge: Kosten, Durchlaufzeit und Nacharbeit durch Menschen.

Die Nacharbeit ist der Posten, den fast alle vergessen – und der den 2,5-fachen Tokenpreis regelmäßig wieder einspielt.

Halte deine Kontexte unter 272.000 Token. Ein gut geschnittenes Dossier ist billiger als ein großzügiges, und meist auch besser.

Ab Tag 61 rollst du nur aus, was du belegen kannst. Prüfe dabei, ob Provisioned Throughput für die Abläufe sinnvoll ist, bei denen eine schwankende Antwortzeit fachlich weh tut. Und baue dir ein Runbook für den Modellwechsel: Ein Limited-Access-Zugang kann entzogen werden, Modelle werden abgekündigt, Preise ändern sich. Halte deine Anwendungen deshalb hinter einer Abstraktionsschicht, damit ein Modellwechsel eine Konfigurationsänderung bleibt und kein Projekt wird.

Der am häufigsten unterschätzte Punkt steht ganz am Ende der Liste: Schulung. Nicht Prompt-Schulung, davon gibt es genug. Sondern Schulung für die Prüfrolle. Wenn ein Agent dir ein fertiges Ergebnis auf den Tisch legt, brauchst du Leute, die wissen, wie man ein plausibel aussehendes, aber falsches Ergebnis erkennt. Das ist eine andere Kompetenz als die, die deine Belegschaft in den letzten zwanzig Jahren trainiert hat, und sie entsteht nicht von allein.

Häufig gestellte Fragen

Ist GPT-6 Astra automatisch in unserem Tenant aktiv?

Teilweise ja. In GitHub Copilot werden neue Modelle standardmäßig aktiviert, solange kein Administrator den globalen Standard abgeschaltet oder das Modell gesperrt hat, und der Schalter für OpenAI als Subprozessor wurde in vielen Microsoft-365-Tenants bereits automatisch gesetzt. In Microsoft Foundry dagegen brauchst du eine Freigabe über das Limited Access Program.

Erfüllt GPT-6 Astra die Anforderungen an EU-Datenresidenz?

Zum Start nicht. Das Modell ist nur in Global Standard und in der US Data Zone verfügbar; eine EU Data Zone existiert noch nicht. Für Workloads mit harter EU-Anforderung bleibt vorerst GPT-5.6 Sol in der EU Data Zone die tragfähige Wahl.

Lohnt sich der Wechsel trotz des 2,5-fachen Token-Preises?

Für lange, werkzeugintensive Aufgaben in der Regel ja, weil Astra bei vergleichbarem Ergebnis deutlich weniger Token verbraucht und damit rund ein Drittel weniger je erledigter Aufgabe kostet. Für kurze Zusammenfassungen, Klassifikation oder einfache Frage-Antwort-Szenarien bleibt das kleinere Modell wirtschaftlicher.

Was ist die 272K-Grenze und warum kostet sie so viel?

Überschreitet ein einzelner Request 272.000 Eingabe-Token, wird der gesamte Aufruf neu bepreist: Eingabe zum doppelten Satz, Ausgabe mit einem Aufschlag von 50 Prozent. Dadurch kann derselbe Arbeitsschritt bei 300.000 Token fast doppelt so teuer sein wie bei 270.000, weshalb sauberes Zuschneiden des Kontexts direkt auf die Rechnung durchschlägt.

Welches neue Sicherheitsrisiko bringt Computer Use mit?

Wenn ein Agent Bildschirminhalte liest, um Aktionen auszuführen, können dargestellte Inhalte sein Verhalten beeinflussen – Prompt Injection wandert damit in die Benutzeroberfläche. Deshalb gehört vor jedes System mit folgenreichen Aktionen ein technisch erzwungener menschlicher Freigabepunkt.