

THEMENBERICHT NR. 1/2026

Künstliche Intelligenz: von der Assistenz zur autonomen Forschung

Warum sinkende Kosten, agentische Systeme und maschinelle Forschung den Bedarf an Rechenleistung verändern.

Claudio Moffa, Portfoliomanager, moffa@hvp.li

KERNAUSSAGEN

KI verschiebt sich von Interaktion zu Delegation. Der Mensch gibt Ziele vor, Agenten arbeiten über Stunden oder Tage, und parallel.

KI erzeugt erste neue Forschungsergebnisse. Ein Modell von OpenAI hat zehn Resultate in Mathematik und Informatik geliefert, formal geprüft.

Sinkende Kosten erhöhen den Gesamtverbrauch. Je billiger Rechenarbeit wird, desto mehr Aufgaben lohnen sich. Die Nachfrage nach Rechenleistung steigt.

WORUM ES GEHT

Rechenarbeit wird je Aufgabe billiger. Trotzdem steigt der Bedarf an Rechenleistung, weil sich damit Aufgaben lohnen, die vorher zu teuer waren. Dieser Zusammenhang ist der Kern des Berichts.

Zwei Beobachtungen aus dem Sommer 2026 stützen ihn: KI-Systeme arbeiten über Tage selbständig an Aufgaben, und sie haben erste eigene Forschungsergebnisse geliefert. Was daran noch unbewiesen ist, steht in Kapitel 4, was es für Anleger bedeutet, in Kapitel 6.

3,1

Agenten-Arbeitstage je menschlichem Arbeitstag in der Forschung von OpenAI

10

neue Resultate in Mathematik und Informatik, erzeugt vom Modell Astra

600

USD Rechenkosten pro Tag beim mittleren Forscher, zu API-Preisen

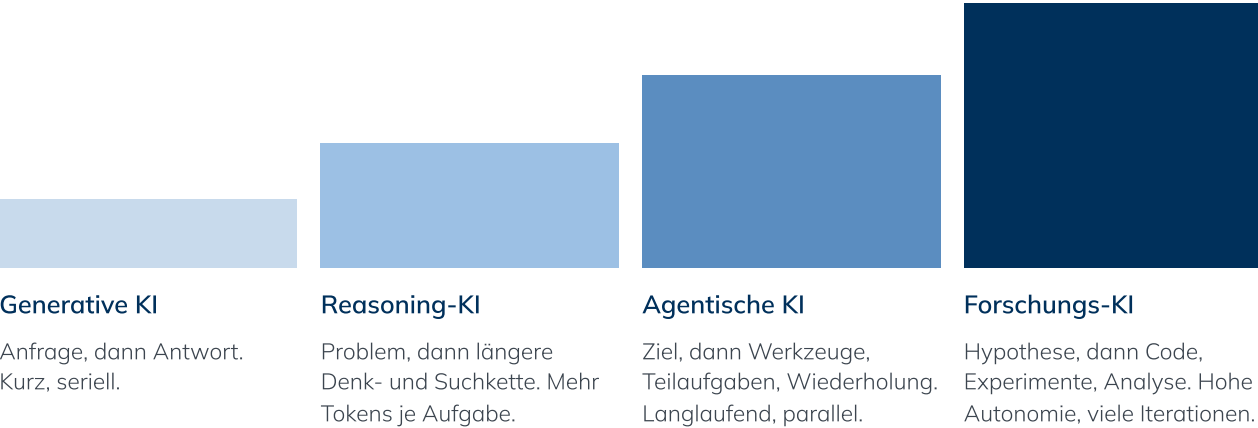
2000

USD Tokenkosten der zehn erfolgreichen Lösungswege, ohne Fehlversuche

Quelle: OpenAI, August und September 2026 [1], [2]. Von OpenAI selbst berichtet, nicht unabhängig geprüft.

1 Vom Chatbot zum Agentensystem

Bisher stellte ein Nutzer eine Frage und erhielt eine Antwort. Agentische Systeme zerlegen Aufgaben, benutzen Werkzeuge, führen Code aus, prüfen Resultate und setzen bei Fehlern neu an. Die Zahl der Nutzer verliert damit an Aussagekraft. Massgebend werden die Zahl gleichzeitig arbeitender Agenten und die Dauer ihrer Aufgaben.



Grafik 1: Vier Entwicklungsstufen der KI-Nutzung. Mit jeder Stufe steigen die Rechenarbeit je Aufgabe und die Dauer, über die ein System selbständig arbeitet. Die Balkenhöhe ist schematisch und beruht nicht auf Messdaten.

Was OpenAI in der eigenen Forschung misst

MESSGRÖSSE	BEOBACHTUNG	BEDEUTUNG
Agenten-Arbeitszeit	3,1 Agenten-Arbeitstage je menschlichem Arbeitstag	Maschinelle Arbeit übersteigt die menschliche Arbeitszeit
Rechenkosten	Über 600 USD pro Tag beim mittleren Forscher	Wissensarbeit wird rechenintensiv
Parallelität	Zunehmend vier und mehr Agenten gleichzeitig	Nutzung wächst in die Breite
Experimente	August 2026: Höchststand seit Januar 2025	Mehr Agenten, mehr Experimente

Tabelle 1: Agenten in der Forschungsorganisation von OpenAI. Quelle: OpenAI, 6. September 2026 [2].

OpenAI nennt den Stand einen «automated research intern»: Das System erledigt klar definierte Aufgaben von mehreren Tagen, unter menschlicher Leitung. Prioritäten, Bewertung und Entscheide bleiben beim Menschen.

2 Neue Mathematik aus dem Modell Astra

Am 1. August 2026 veröffentlichte OpenAI zehn Resultate in Mathematik und theoretischer Informatik, erzeugt von einer internen Version des Modells Astra [1]. Sie lösen langjährige offene Probleme oder bringen sie substanziell voran.

- 01

Die Beweise stammen vom System. Menschen bereiteten die Manuskripte auf.
-
- 02

Die Argumente wurden anschliessend in Lean formal geprüft. Das ersetzt keine unabhängige Replikation.
-
- 03

Tokenkosten für die zehn Lösungswege: rund 2000 USD. Ohne Fehlversuche, Training, Infrastruktur und menschliche Prüfung.
-
- 04

Ein Teil eines Beweises musste im August lokal korrigiert werden [4]. Solche Prüfungen gehören zum Übergang von der Demonstration zur belastbaren Wissenschaft.
-

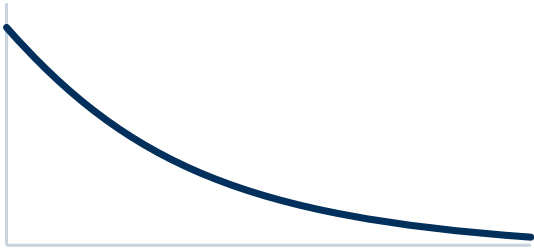
Damit erzeugt KI neues Wissen und automatisiert nicht mehr bloss bestehende Arbeit. Das zeigt sich zuerst in fünf Feldern.

Softwareentwicklung Code schreiben, testen, korrigieren	KI-Forschung Experimente bauen, ausführen, auswerten
Mathematik Offene Probleme als maschinelle Forschungsaufgaben	Cybersicherheit Schwachstellen suchen, Gegenmassnahmen entwickeln
Naturwissenschaften und Technik Simulation, Materialsuche, Wirkstoffentwicklung	

3 Ein Jevons-Effekt für Intelligenz

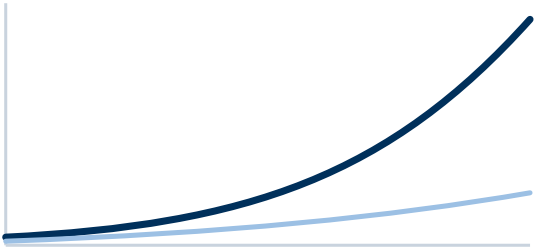
Effizienz gilt oft als Argument für sinkenden Ressourcenbedarf. Bei stark elastischer Nachfrage tritt das Gegenteil ein: Sinkende Kosten machen so viele neue Anwendungen wirtschaftlich, dass der Gesamtverbrauch steigt. Der Ökonom Jevons beobachtete das 1865 beim Kohleverbrauch effizienterer Dampfmaschinen.

KOSTEN JE TOKEN



Fallen mit jeder Modellgeneration.

GESAMTNUTZUNG VON RECHENLEISTUNG



— Tokens, Agentenstunden, Experimente
— Zahl der Nutzer

Grafik 2: Kosten je Einheit fallen, Gesamtverbrauch steigt. Schematische Darstellung ohne Messdaten.

WAS GÜNSTIGER ODER BESSER WIRD

WAS DADURCH ZUNIMMT

Günstigere Tokens

Mehr Aufgaben lassen sich automatisieren

Schnelleres Reasoning

Mehr Iterationen in derselben Zeit

Bessere Agenten

Längere Aufgaben werden delegiert

Mehr Parallelität

Mehrere Lösungswege gleichzeitig

Forschungsautomation

Mehr Experimente, kürzere Innovationszyklen

Tabelle 2: Wie Effizienzgewinne die Nachfrage erhöhen. Eigene Darstellung.

OpenAI selbst erwartet, dass Rechenleistung zum wichtigsten Engpass wird, sobald andere Engpässe der Forschung abnehmen [2]. Betroffen sind alle fünf Ebenen der Infrastruktur.

Beschleuniger, GPU und TPU	mehr Training, Reasoning, Inferenz
Speicher, HBM	grössere Modelle, mehr Datenbewegung
Netzwerk und Optik	grössere Cluster, mehr Parallelität
Rechenzentren	Agenten arbeiten rund um die Uhr
Strom und Kühlung	höhere Leistungsdichte, höhere Auslastung

Grafik 3: Fünf Ebenen der KI-Infrastruktur. Jede Ebene wird relevant, wenn Agenten dauerhaft und parallel arbeiten. Die Darstellung ist gestalterisch, nicht quantitativ.

4 Was noch nicht bewiesen ist

- 01

Keine «Intelligence Explosion» und keine autonome Selbstverbesserung. OpenAI sagt das ausdrücklich [2].
- 02

Die 2000 USD sind keine Vollkosten. Training, Fehlversuche, Infrastruktur und Menschen fehlen darin.
- 03

Lean prüft die Logik, nicht die Bedeutung, Neuheit oder Relevanz eines Resultats.
- 04

Alle Messdaten stammen von OpenAI. Eine unabhängige Bestätigung steht aus.
- 05

Die Leiden Declaration [3] zeigt, dass Mathematiker Fragen zu Urheberschaft und Verantwortung ernst nehmen. Sie lehnt KI nicht ab.

5 Woran wir es erkennen

Sechs Indikatoren zeigen in den nächsten 12 bis 24 Monaten, ob sich die These bestätigt. Die Spalte «Stand» schreiben wir in jeder Ausgabe fort.

INDIKATOR	STAND SEPTEMBER 2026	FRAGE FÜR DIE NÄCHSTE AUSGABE
Agenten-Arbeitszeit je Wissensarbeiter	3,1 Agenten-Tage je Arbeitstag bei OpenAI [2]	Steigt der Wert, auch ausserhalb von OpenAI?
Kosten je Aufgabe gegenüber Gesamtnutzung	600 USD pro Tag und Forscher [2], 2000 USD für zehn Beweise [1]	Fallen die Kosten je Aufgabe, steigt die Gesamtnutzung?
Autonom durchgeführte Experimente	Höchststand seit Januar 2025 [2]	Nimmt die Aktivität weiter zu?
Aufgaben von mehreren Stunden oder Tagen	Mehrtägige Aufgaben unter menschlicher Leitung [2]	Verlängert sich der Zeithorizont?
Unabhängige Replikation der Resultate	Eine lokale Korrektur [4], keine vollständige Verifikation	Bestätigen oder widerlegen Fachleute die Resultate?
Engpässe bei Rechenzentren, Netzwerk, Speicher, Energie	OpenAI nennt Rechenleistung als künftigen Engpass [2]	Verlängern sich Lieferzeiten, steigen Preise und Auslastung?

Tabelle 3: Beobachtungsrahmen. Quellen in eckigen Klammern siehe Quellenverzeichnis.

6 Was das für Anleger bedeutet

Einordnung, keine Empfehlung für ein bestimmtes Finanzinstrument.

Wo Nachfrage entsteht

Entlang der fünf Infrastrukturebenen aus Grafik 3, von Beschleunigern bis zu Strom und Kühlung.

Bei den Anbietern der Modelle.

In einem zweiten Schritt bei Unternehmen, die Aufgaben an Agenten delegieren und so ihre Kosten je Arbeitsergebnis senken.

Was dagegen spricht

Unbewiesene These. Bleibt die Delegation hinter den Erwartungen zurück, ist ein Teil der Infrastruktur nicht ausgelastet.

Konzentration. Die Nachfrage hängt von den Investitionsentscheiden weniger grosser Unternehmen ab.

Kapitalintensität. Rechenzentren und Energie binden über Jahre Kapital, bevor sich die Nachfrage zeigt.

Energie und Regulierung. Strom ist die unterste Ebene des Stapels und kann das Wachstum bremsen.

Wie wir damit umgehen

Wir betrachten KI als Faktor, der Unternehmen in vielen Branchen unterschiedlich trifft. Ein Unternehmen kann von steigender Rechenarbeit profitieren, sie als Kosten tragen oder durch Agenten produktiver werden. Diese drei Fragen stellen wir in der Anlageauswahl.

Glossar

Agent	KI-System, das ein Ziel in Teilschritte zerlegt, Werkzeuge nutzt und über längere Zeit selbständig arbeitet.
AGI	Allgemeine künstliche Intelligenz: erledigt die meisten geistigen Aufgaben mindestens so gut wie ein Mensch. Es gibt keine anerkannte Definition.
API-Preise	Listenpreise für die Nutzung eines Modells über die Programmierschnittstelle, meist je Million Tokens. «Sol» bezeichnet das OpenAI-Modell, dessen Preise zugrunde liegen.
Inferenz	Der Betrieb eines fertig trainierten Modells, also das Erzeugen von Antworten. Gegenstück zum Training.
Jevons-Effekt	Wird eine Ressource effizienter genutzt, kann ihr Gesamtverbrauch steigen, weil mehr Anwendungen wirtschaftlich werden. Jevons, 1865.
Lean	Beweisassistent, mit dem ein Computer jeden Schritt eines mathematischen Beweises formal prüfen kann.

Reasoning	Ein Modell durchläuft vor der Antwort eine längere Kette von Zwischenschritten. Verbraucht mehr Tokens je Aufgabe.
Rechenleistung	Englisch «Compute»: Prozessoren, Speicher, Netzwerk und Rechenzentren für Training und Betrieb von KI-Modellen.
Token	Kleinste Verarbeitungseinheit eines Sprachmodells, etwa ein Wortteil. Nutzung und Preise werden in Tokens gemessen.

Quellen

- [1] OpenAI, «Ten advances in mathematics and theoretical computer science», 1. August 2026, openai.com/index/ten-advances-in-mathematics
- [2] OpenAI, «Research acceleration: The view inside OpenAI», 6. September 2026, openai.com/index/research-acceleration-view-inside-openai
- [3] Leiden Declaration on Artificial Intelligence and Mathematics, 2. Juni 2026, leidendeclaration.ai
- [4] Sienicki, M. und Sienicki, K., «Auditing an AI-Generated Mathematical Proof: A Correction to a Greedy Conditioning Lemma in Quantum Parallel Repetition», arXiv:2608.14673, August 2026

Hinweis. Dieser Bericht dient der Information und Einordnung technologischer Entwicklungen. Er stellt keine Anlageberatung, keine Empfehlung für ein bestimmtes Finanzinstrument und keine Prognose zukünftiger Renditen dar. Aussagen zu noch nicht unabhängig replizierten Forschungsergebnissen sind als vorläufig zu verstehen. Die Grafiken 1 bis 3 sind schematische Darstellungen ohne Messdaten.

HighValue Partners AG

Drescheweg 1a, 9490 Vaduz, Liechtenstein
www.hvp.li

Autor

Claudio Moffa, Portfoliomanager
moffa@hvp.li, +423 384 51 04