BALANE

Der Preis des Denkens, nachgerechnet

Methodenbericht zur Reanalyse von Epoch AI, „The plunging price of thought“ (22.09.2026): Datengrundlage, exaktes Verfahren, alternative Spezifikationen, Robustheit und Grenzen.

Der Schnäppchenjäger
wechselt jedes Quartal zum billigsten Modell. So rechnet die Studie.

100 € heute, in einem Jahr:

8 €

Spanne 4 bis 11 €

Die normale Firma
bleibt bei OpenAI, Anthropic oder Google, übliches aktuelles Modell. Modellrechnung.

100 € heute, in einem Jahr:

96 €

Spanne 86 bis 119 €

Die Anspruchsvolle
nimmt jedes neue Modell in der höchsten Einstellung.

100 € heute, in einem Jahr:

334 €

Spanne 252 bis 458 €

Alle drei Werte stammen aus denselben Daten. Keiner ist falsch. Sie beantworten verschiedene Fragen. Die Überschrift „13-mal billiger“ beantwortet nur die erste.

VORBEMERKUNG

Dieser Bericht richtet sich nicht gegen Epoch AI. Man kann so rechnen, und Epoch nennt die Grenzen der eigenen Zahl im Bericht selbst. Die Zahl muss nur richtig eingeordnet werden, und wer sie liest, muss verstehen, was sie misst. Gegenstand dieses Berichts ist die Interpretation, nicht die Quelle.

Autor	Jonas Höttler, Balane
Version, Datum	1.0, Oktober 2026
Datengrundlage	Epoch AI, github.com/droodman/inference-cost, Commit 9163c17 vom 24.09.2026, abgerufen am 19.10.2026. Ergänzend Epochs Arbeitsmappe „Cost declines by tech.xlsx“, Eurostat, Fouquet (2011), Ember, FRED, getdeploying.com.
Begleitartikel	hoettler.com/de/blog/preis-des-denkens (mit interaktivem Rechner)
Reproduktion	Alle Skripte unter drafts/preis-des-denkens/analyse/ (Abschnitt 12)

KURZFASSUNG

Fragestellung, Vorgehen, Ergebnisse

Fragestellung. Epoch AI berichtet, dass der Preis für eine feste Leistung von KI-Modellen seit 2023 um rund 47 % pro Quartal fällt, hochgerechnet etwa 13-mal pro Jahr. In sozialen Medien wird daraus „dieselbe Leistung wird jedes Jahr 13-mal billiger, die Richtung ist eindeutig“. Wir prüfen, was die Zahl misst, wie stark sie von vertretbaren Annahmen abhängt und was sie für Kunden bedeutet, die nicht ständig das billigste Modell wählen.

Daten. Epochs öffentlicher Datensatz: 15.959 Messzeilen aus 1.298 Benchmark-Läufen (225 Modellbezeichnungen), nach Epochs Aufbereitung 11.240 Zeilen von 189 Modellen auf 11 Benchmarks. Preise sind Listenpreise zum Messzeitpunkt, zu 96 % von OpenRouter.

Methode. (1) Exakte Reproduktion von Epochs modellfreier Rechnung in Python. (2) Variation jeder Annahme einzeln und in allen 512 Kombinationen. (3) Drei Kundenprofile, Robustheitsprüfungen (Fensterlänge, Weglassen einzelner Modelle, Bootstrap). (4) Zerlegung des Rückgangs in Tokenmenge und Tokenpreis. (5) Fortschreibung der Vergleichsreihen (Strom, Rechenleistung) mit Epochs eigenen und öffentlichen Daten.

Ergebnisse. Die Reproduktion trifft Epochs Werte auf die zweite Nachkommastelle (47,14 / 42,96 / 53,11 / 46,97 / 44,01 % pro Quartal). Je nach Spezifikation reicht das Ergebnis von 26,4-mal billiger bis 1,3-mal teurer pro Jahr. 53 von 512 Kombinationen zeigen steigende Kosten. Ein üblicher Preis statt des Minimums ergibt 1,6-mal billiger pro Jahr, der übliche Preis der drei großen Anbieter praktisch keine Änderung (100 € werden 96 €, Spanne 86 bis 119 €). Wer jedes neue Modell in höchster Einstellung nutzt, zahlt pro Aufgabe 3,3-mal mehr pro Jahr (Spanne 2,5 bis 4,6). Drei Viertel des Rückgangs am Rekordpreis kommen über billigere Tokens. Epochs Stromkurve endet 1973. Danach ist Strom in den USA real nicht mehr billiger geworden (1973 bis 2025: ±0 %).

Schluss. Am billigsten Rand des Marktes fällt der Preis tatsächlich schnell. Für die meisten Kunden folgt daraus wenig. Die Daten stützen keine Aussage „wird sicher drastisch billiger“, sondern: Es kommt stark darauf an, welcher Preis, welche Anbieter, welche Aufgaben und welches Mittel gemeint sind.

Tabelle 0: Acht verbreitete Lesarten und was die Daten zeigen

Lesart im Feed	Befund dieses Berichts	Abschnitt
„Dieselbe Leistung wird 13-mal billiger.“	Gilt für den billigsten Anbieter bei sofortigem Wechsel. Übliches Modell: 1,6-mal. Übliches Modell der großen drei: praktisch unverändert.	5.1, 5.3
„Jedes Jahr 13-mal.“	Kein gleichmäßiger Trend: je nach Halbjahr 3,6- bis 132-mal. Am mittleren Messpunkt bewegt sich der Rekordpreis gar nicht.	5.6
„Schneller als jede Technologie zuvor.“	Im fairen Vergleich (schnellste drei Jahre) 1,3-mal so schnell wie DNA-Sequenzierung statt 4,2-mal.	10
„Leistung“ heißt Arbeit.	Gemessen an Mathematik, Schach, Prüfungsfragen. Programmieren und Faktenwissen: 5,3-mal.	5.4
„Etwas mehr zahlen, ein Vielfaches bekommen.“	Nicht gemessen. Jedes neue Modell in höchster Einstellung: 3,3-mal teurer pro Aufgabe und Jahr.	5.3
„Die Richtung ist eindeutig, nicht warten.“	Neue Spitzenleistung wird laut Epoch anfangs um rund zwei Drittel pro Quartal billiger. Das stützt ebenso das Abwarten.	11
„Das geht so weiter.“	Drei Viertel des Rückgangs hängen am Tokenpreis, also an Strom, Chips und Preispolitik.	7, 8
„Wenn es so gut läuft, ist es sicher.“	Schneller Wertverlust neuer Modelle ist für Anbieter und Investoren ein Risiko, keine Garantie.	11

Inhalt

1 Ausgangslage und Fragestellung
2 Daten
3 Epochs Verfahren, exakt
4 Alternative Spezifikationen
5 Ergebnisse
6 Robustheit
7 Zerlegung: Tokenmenge und Tokenpreis
8 Szenario-Rechner
9 Externe Reihen
10 Technologievergleich
11 Grenzen dieser Analyse
12 Reproduzierbarkeit
A Glossar
B Sechs Fragen an jede Preisstatistik
C Quellen

ABSCHNITT 1

Ausgangslage und Fragestellung

1.1 Epochs Aussage

Epoch AI untersucht, wie schnell der Preis sinkt, zu dem ein KI-Modell ein festes Leistungsniveau auf einem Benchmark erreicht. Das Ergebnis der modellfreien Rechnung: im Mittel über fünf Benchmarks 46,8 % billiger pro Quartal, entsprechend 12,5-mal billiger pro Jahr. Im Bericht wird das auf „etwa 13-mal“ gerundet und mit Strom, Batterien, Rechenleistung und DNA-Sequenzierung verglichen.

Der Bericht ordnet die Zahl selbst vorsichtig ein:

> *„Essentially no user stays permanently on the cost frontier.“*

> *„when comparing price drops for AI to drops for other technologies for which we have price series, we are comparing apples and oranges.“*

> *„They should not be read as exact.“*

> *„supra-normal profits in LLM service provision may be fleeting for any given model“*

Diese Vorbehalte gehen beim Weiterverbreiten verloren. Aus einer vorsichtig eingeordneten Messgröße wird „dieselbe Leistung wird jedes Jahr 13-mal billiger, die Richtung ist eindeutig“ und daraus „KI ist die sichere Wette“.

1.2 Fragestellung

1. Lässt sich Epochs Zahl aus den veröffentlichten Daten exakt reproduzieren?
2. Wie stark hängt sie von Annahmen ab, die man ebenso vertretbar anders treffen kann?
3. Was ergibt sich für Kunden, die nicht jedes Quartal zum billigsten Modell wechseln?
4. Woraus setzt sich der Rückgang zusammen, und was setzt seine Fortschreibung voraus?
5. Hält der Vergleich mit anderen Technologien, wenn man deren Reihen zu Ende zeichnet?

1.3 Einordnung

Kritisiert wird nicht Epochs Rechnung, sondern ihre Interpretation. Epochs Zahl ist korrekt berechnet (Abschnitt 3.6). Sie misst aber eine sehr bestimmte Größe: den niedrigsten Preis, den irgendein Kunde am gesamten Markt zahlen könnte, wenn er jedes Quartal zum günstigsten Modell in dessen sparsamster Einstellung wechselt. Dieser Bericht zeigt, wie weit andere, ebenso vertretbare Größen davon entfernt liegen.

IN EINFACHEN WORTEN

Stellen Sie sich vor, jemand sagt: „Flüge werden jedes Jahr 13-mal billiger.“ Im Kleingedruckten steht: gemeint ist der billigste Flug irgendeiner Airline, um fünf Uhr morgens, ohne Gepäck, gebucht von jemandem, der jedes Quartal alle Airlines durchsucht und sofort umbucht. Genau so ist die KI-Zahl gebaut.

ABSCHNITT 2

Daten

2.1 Quelle und Stand

Grundlage ist Epochs öffentliches Repositorium `github.com/droodman/inference-cost`, Stand Commit `9163c17` (24.09.2026), abgerufen am 19.10.2026. Verwendet werden:

- `data/cost_truncated_curves.csv`: eine Zeile je Modell, Benchmark, Einstellung (effort) und Token-Budget. Spalten unter anderem Trefferquote, mittlere Tokens pro Aufgabe, Kosten pro Aufgabe in US-Dollar, Trefferquote des vollständigen Laufs, Stichprobengröße, Preisdatum und Preisquelle.
- `data/Model versions-Grid view.csv`: Modellregister mit Release-Daten.
- `data/Cost declines by tech.xlsx`: die Reihen der Vergleichsgrafik (Blatt „Figure data“), Rohdaten zu Strom bis 2025 (Blatt „Electricity“) und eine Verbraucherpreistabelle (Blatt „CPI“).
- `src/prepare_data.R` und `src/envelope_frontier.R`: Datenaufbereitung und modellfreie Rate, die wir nach Python übertragen haben.

2.2 Umfang je Benchmark

Die Rohdatei enthält 15.959 Zeilen aus 1.298 Läufen mit 225 Modellbezeichnungen. Nach Epochs Aufbereitung (Abschnitt 2.4) bleiben 11.240 Zeilen von 189 Modellen.

Tabelle 1: Datenumfang nach Aufbereitung

Benchmark (Kürzel)	Inhalt	Zeilen	Modelle	Bestwert	Release-Daten
AIME (`aime`)	Mathematik-Olympiade	2.159	146	100 %	03/2023 bis 07/2026
Chess puzzles (`chess`)	Schachaufgaben	1.462	97	61,7 %	06/2023 bis 07/2026
FrontierMath Tier 1 bis 3 (`fm13`)	Forschungsmathematik	1.099	58	89,1 %	11/2023 bis 08/2026
GPQA (`gpqa`)	Naturwissenschaftliche Prüfungsfragen	2.164	166	93,6 %	03/2023 bis 07/2026
Mystery Game Puzzles (`mystery`)	Geheimes Spiel, gegen gezieltes Training gebaut	623	37	26,2 %	06/2023 bis 07/2026
FrontierMath 2025-02 privat	Forschungsmathematik	822	50	47,2 %	05/2024 bis 05/2026
FrontierMath Tier 4 (2025-07 privat)	Forschungsmathematik	570	32	31,2 %	06/2024 bis 05/2026
FrontierMath Tier 4 v2	Forschungsmathematik	499	38	87,8 %	01/2025 bis 08/2026
MATH Level 5	Schulmathematik, schwer	808	91	94,8 %	06/2023 bis 10/2025
SimpleQA (`simpleqa`)	Faktenwissen	449	42	77,3 %	10/2024 bis 06/2026
SWE-bench Verified (`swe_bench_verified`)	Programmieren	585	43	73,4 %	05/2024 bis 06/2026

Hervorgehoben: Epochs fünf Haupt-Benchmarks. Bestwert: höchste um Raten bereinigte Trefferquote (Abschnitt 2.5). Quelle: eigene Auswertung von `build_runs` (core.py).

2.3 Preise

Epoch misst Listenpreise zum jeweiligen Zeitpunkt, nominal in US-Dollar. Die Preisquellen der Rohzeilen:

Tabelle 2: Preisquellen (Anteil der Rohzeilen)

Quelle	Bedeutung	Anteil
`openrouter-wayback`	OpenRouter-Preis aus archivierten Seiten zum Messzeitpunkt	73,6 %
`openrouter-wayback, nearest-after`	dasselbe, nächster späterer Archivstand	21,9 %
`openrouter-live, nearest-after`	aktueller OpenRouter-Preis	0,7 %
`gpu-effective` (auch nearest-after)	selbst betriebene offene Modelle, umgerechnet über GPU-Mietpreise von 2026	3,4 %
`moonshot / alibaba-first-party`	Preis des Anbieters selbst	0,4 %

Zusammen rund 96 % OpenRouter. Strom, Chips, Subventionen und Margen werden nicht eigens erfasst. Sie wirken nur indirekt über diese Preise. Die KI-Preise sind nicht inflationsbereinigt. Über drei Jahre ist das ein kleiner Effekt, der reale Rückgang wäre minimal größer.

Nicht alle Preise stammen vom Release-Zeitpunkt. Bei einigen Modellen liegt das Preisdatum bis zu 2,7 Jahre später (zum Beispiel Llama-3.1-8B: 724 Tage, deepseek-llm-67b: 976 Tage, jeweils über GPU-Miete 2026). Das macht frühe Modelle eher zu billig und verzerrt den Rückgang daher eher nach unten.

2.4 Aufbereitung nach Epochs Code

1. **Namen bereinigen:** Endungen `-maas…` und Datumsendungen `-JJJJ-MM-TT…` entfernen.
2. **Zusammenfassen** je Modell, Benchmark, Einstellung und Budget: Mittelwert von Trefferquote, Tokens, Kosten und Trefferquote des vollen Laufs, Summe der Stichproben.
3. **Benchmarks** umbenennen (`fm_tiers1_3_v2` → `fm13`, `mystery_agent` → `mystery`), öffentliche Teilmengen (`*_public`) verwerfen.
4. **Release-Datum** = frühestes Datum aller Registereinträge, deren Kennung mit dem Modellnamen beginnt (Groß- und Kleinschreibung egal). Zu den Folgen siehe 4.7.
5. **Raterei-Bereinigung** (2.5), dann **Duplikate** über Benchmark, Modell, Einstellung, Trefferquote, Kosten und Release-Datum entfernen. Zeilen ohne gültige Kosten (ln c nicht endlich) oder ohne Datum fallen weg.
6. **Zeitachse**: $t = (\text{Release} - 1.1.2023) \,/\, 365{,}25$ in Jahren, je Benchmark zentriert: $t^c = t - \bar{t}_b$.

2.5 Raterei-Bereinigung

Bei Multiple-Choice-Tests erreicht man eine Trefferquote schon durch Raten (bei vier Antworten 25 %). Epoch bestimmt je Benchmark eine Ratequote f_b: den häufigsten Wert unter den Trefferquoten, die höchstens so groß sind wie der Median (bei Gleichstand der kleinste). Dann gilt

$$a' = \frac{\max\left(f_b, a\right) - f_b}{1 - f_b} \tag{1}$$

Ergebnis: GPQA $f = 0{,}25$, Mystery 0,092, AIME 0,001, Chess und FrontierMath 0. Alle weiteren Rechnungen verwenden a', im Folgenden kurz a.

2.6 Simulierte Budgetkürzung

Jeder Lauf liegt in 12 bis 13 Budgetstufen vor (64, 128, … Tokens). Jede Stufe gibt an, welche Trefferquote und welche Kosten sich ergeben hätten, wäre der Lauf bei diesem Token-Budget abgebrochen worden. Diese simulierten Kürzungen (in unseren Skripten „CAISI“) gehen bei Epoch als eigene Punkte in die Rechnung ein. Die Variante „nur echte Läufe“ behält je Lauf nur die Stufe mit dem höchsten Budget, also den vollständigen Lauf (4.6).

2.7 Weitere Datenquellen

Tabelle 3: Ergänzende Reihen

Quelle	Inhalt	Verwendung
Epoch, Cost declines by tech.xlsx	Vergleichsreihen, Strom USA nominal bis 2025 (Hirsh), CPI	Abschnitte 9.1, 10
Fouquet (2011), LSE Grantham Institute	Strompreis Großbritannien real, 1883 bis 2008	9.2
Eurostat `nrg_pc_204`, `prc_hicp_aind`	Haushaltsstrom, Verbraucherpreisindex HICP	9.2, 9.3
Ember	Großhandelsstrom Europa, Day-Ahead, monatlich und täglich	9.4
FRED	`APU000072610` (Haushaltsstrom USA), `DFF`, `ECBDFR`	9.5, 9.7
getdeploying.com	GPU-Preise, wöchentlich, Median über Anbieter	9.6
BTS, Deutsche Bahn, American Express	Flugpreise, Bahntarife	Exkurs in 10.5
Presse und Indizes (Monitoring Analytics, SemiAnalysis, TrendForce, WSJ, FT)	Strommarkt PJM, Chip-Mietpreise, Speicher, Geschäftszahlen	nur Orientierung der Szenario-Werte (8)

ABSCHNITT 3

Epochs Verfahren, exakt

Die Hauptzahl stammt aus Epochs modellfreier Rechnung (`pareto_decline_qtr` in `envelope_frontier.R`). Sie kommt ohne angepasste Funktionsform aus. Wir beschreiben sie für einen Benchmark b mit Läufen i, jeweils mit Trefferquote a_i, Kosten pro Aufgabe c_i und zentriertem Release-Zeitpunkt t_i.

3.1 Rekordpreis und Bestwert

Der Rekordpreis für Niveau a zum Zeitpunkt t ist der niedrigste Preis, zu dem irgendein bis dahin erschienenes Modell in irgendeiner Einstellung mindestens a erreicht:

$$C_b(a,t) = \min\{c_i : t_i \le t,\ a_i \ge a\} \tag{2}$$

$$S_b(t) = \max\{a_i : t_i \le t\} \tag{3}$$

$S_b(t)$ ist der Bestwert (State of the Art) zum Zeitpunkt t. Ein Modell bleibt in (2) ab seinem Release für immer zu seinem gemessenen Preis verfügbar.

3.2 Messgitter

Zeit: gleichmäßig verteilte Punkte t_k von $\min t_i$ bis $\max t_i$, Anzahl $n = \max(2, \text{round}(\text{Spanne} / h) + 1)$ mit $h = 1/28$ Jahr (rund 13 Tage). Verwendet werden nur Punkte mit vollem Horizont, $t_k + \Delta t \le \max t_i$.

Niveaus: ein festes Raster aus $N = 100$ Mittelpunkten über den erreichten Bereich, abgeschnitten am jeweiligen Bestwert:

$$a_j = a_{\min} + \left(a_{\max} - a_{\min}\right)\frac{j - \tfrac{1}{2}}{N}, \quad j = 1, \ldots, N, \quad \text{Knoten } (j,k) \text{ nur, wenn } a_j \le S_b(t_k) \tag{4}$$

$a_{\min}$ ist die kleinste positive, $a_{\max}$ die größte beobachtete Trefferquote des Benchmarks. Epoch nennt dieses Raster „fixed, clipped“. Weil das Raster am Bestwert endet, haben spätere Zeitpunkte mehr Knoten.

3.3 Änderung je Knoten und Quartalsrate

$$\delta_{jk} = \ln C_b(a_j, t_k + \Delta t) - \ln C_b(a_j, t_k), \quad \Delta t = 0{,}25 \tag{5}$$

$$r_b = 1 - \exp\left(\frac{0{,}25}{\Delta t} \cdot \bar{\delta}\right), \quad \bar{\delta} = \text{Mittelwert aller } \delta_{jk} \tag{6}$$

Das arithmetische Mittel der Log-Änderungen ist ein geometrisches Mittel der Preisverhältnisse. Knoten, an denen sich der Rekordpreis im Quartal nicht bewegt, gehen mit $\delta = 0$ (Verhältnis 1) ein. Epochs Code weist dazu den Anteil bewegter Knoten aus (`share_moved`).

3.4 Aggregation über Benchmarks und Jahresfaktor

$$\bar{r} = \frac{1}{5}\sum_b r_b, \quad F = (1 - \bar{r})^{-4}, \quad \text{Rechnung in einem Jahr} = \frac{100\,€}{F} \tag{7}$$

Mit Epochs Werten: $\bar{r} = 46{,}84\,\%$, $F = 12{,}52$, Rechnung 8,0 €. Dieselbe Aggregation (arithmetisches Mittel der Quartalsraten über die gewählten Benchmarks) verwenden alle Spezifikationen und der Rechner im Artikel.

3.5 Eine Eigenschaft der Konstruktion

In (2) wird das Minimum über eine Menge gebildet, die mit der Zeit nur wächst. $C_b(a,t)$ kann deshalb nicht steigen, $\delta_{jk} \le 0$ für jeden Knoten. Epochs Code vermerkt das selbst: „Non-decreasing in BOTH arguments by construction, so its curves can never cross and it can never run backwards.“ Dass der Preis fällt, ist in dieser Rechnung also kein Befund, sondern Konstruktion. Offen ist nur, wie schnell. Die Variante mit begrenzter Lebensdauer (4.2) prüft, ob das eine Rolle spielt. Sie tut es kaum.

IN EINFACHEN WORTEN

Der 100-Meter-Weltrekord kann nur besser werden, weil ein alter Rekord stehen bleibt, bis ihn jemand bricht. Daraus folgt nicht, dass Sie schneller laufen. Der Rekordpreis ist so ein Weltrekord.

3.6 Reproduktion

Tabelle 4: Epochs veröffentlichte Werte und unsere Reproduktion

Benchmark	Epoch, % pro Quartal	Reproduktion	bewegte Knoten	Knoten
AIME	47,1	47,14	47,7 %	4.159
Chess puzzles	43,0	42,96	41,1 %	3.031
FrontierMath Tier 1 bis 3	53,1	53,11	62,2 %	2.102
GPQA	47,0	46,97	46,4 %	5.039
Mystery Game Puzzles	44,0	44,01	24,6 %	3.518
Mittel (7)		46,84		12,5× pro Jahr

Quelle: core.py. Die Abweichung liegt unter 0,05 Prozentpunkten. Unterschiede in den folgenden Abschnitten gehen daher auf Annahmen zurück, nicht auf die Implementierung.

3.7 Was wir nicht nachgebaut haben

Epochs geglättete Regressionsmodelle (bevorzugt eine Box-Tidwell-Spezifikation, angepasst an die Rekordtreppe) haben wir nicht übertragen. Sie landen im Mittel auf derselben Zahl, 47,0 %. Für die Hauptzahl sind sie nicht nötig. Gebraucht werden sie für die Zusatzaussage, neue Spitzenleistung werde direkt nach Erscheinen um rund 66 % pro Quartal billiger. Diese Aussage stammt aus Heatmaps, auf denen ein großer Teil der Fläche keine Messwerte hat (Abschnitt 11).

ABSCHNITT 4

Alternative Spezifikationen

Jede Zahl in einer Überschrift beruht auf Entscheidungen. Wir variieren jede Entscheidung, für die eine andere Wahl ebenso vertretbar ist. Alle Varianten verwenden dieselbe Aufbereitung (Abschnitt 2) und dieselbe Aggregation (7). Implementiert in `precompute.py`.

Tabelle 5: Die Schalter

Achse	Epoch	Alternativen
Preis	`min`: Rekordpreis (2)	`min24`, `min12`: Modell nach 24 bzw. 12 Monaten nicht mehr verfügbar (8). `typ`: üblicher Preis (9)
Anbieter	alle	nur OpenAI, Anthropic, Google (4.3)
Niveau-Raster	fest, abgeschnitten, linear	skaliert (10), logit-verteilt (11), beides
Horizont Δt	ein Quartal	ein Jahr, umgerechnet über (6)
Läufe	mit simulierten Kürzungen	nur vollständige Läufe (4.6)
Release-Daten	wie im Code	korrigiert (Tabelle 6)
Mittel	Mittelwert der δ	Median der δ (4.8)
Benchmarks	fünf Haupt-Benchmarks	alle elf, oder alltagsnah: SWE-bench Verified und SimpleQA

4.1 Preis: begrenzte Lebensdauer

Ein Modell zählt nur während eines Fensters w nach seinem Release:

$$C_b^{(w)}\left(a,t\right) = \min\left\{c_i : t-w < t_i \le t,\ a_i \ge a\right\}, \quad w \in \left\{1,2\right\}\ \text{Jahre} \tag{8}$$

Damit kann der Rekordpreis wieder steigen, wenn das bisher billigste Modell aus dem Fenster fällt.

4.2 Preis: üblicher Preis („typisches Modell")

Statt des billigsten Modells am Markt nehmen wir das Modell in der Mitte. Für jedes Modell m sei $c_m(a)$ sein günstigster Preis, mit dem es Niveau a erreicht. $M(t)$ sind die Modelle, die in den zwölf Monaten vor t erschienen sind und a erreichen:

$$c_m\left(a\right) = \min_{i\,\in\, m,\ a_i \ge a} c_i, \qquad T_b\left(a,t\right) = \operatorname*{Median}_{m\,\in\, M(t)} c_m\left(a\right), \qquad M\left(t\right) = \left\{m : t-1 < t_m \le t\right\} \tag{9}$$

T_b ist nur definiert, wenn mindestens zwei Modelle einen Wert haben. Knoten, an denen T_b zu einem der beiden Zeitpunkte undefiniert ist, fallen weg. Eine Spezifikation braucht mindestens 20 gültige Knoten je Benchmark, sonst gilt sie dort als „zu wenig Daten". Weil T_b steigen kann, ist hier auch ein Preisanstieg messbar.

4.3 Anbieter

Filter auf den bereinigten Modellnamen mit dem regulären Ausdruck `^(gpt-(?!oss)|o1|o3|o4|claude-|gemini-)` : GPT-, o-Serie, Claude und Gemini, ohne die offenen gpt-oss-Modelle. Der Filter wirkt nach der Aufbereitung, die Ratequote (1) bleibt die des vollen Datensatzes. Auf den fünf Haupt-Benchmarks sind das 75 Modelle.

4.4 Niveau-Raster

Epochs Abbildung 7 zeigt vier Raster. Skaliert heißt: das Raster wird zu jedem Zeitpunkt über den bis dahin erreichten Bereich gelegt, jeder Zeitpunkt hat also gleich viele Knoten:

$$a_j\left(t\right) = a_{\min} + \left(S_b\left(t\right) - a_{\min}\right)\frac{j-\tfrac{1}{2}}{N} \tag{10}$$

Logit-verteilt heißt: gleiche Abstände auf der Logit-Skala, mit $\ell(x) = \ln\left(x\,/\,(1-x)\right)$ und Grenzen auf $[0{,}001;\,0{,}999]$ begrenzt:

$$a_j = \ell^{-1}\left(\ell\left(a_{\text{lo}}\right) + \left(\ell\left(a_{\text{hi}}\right) - \ell\left(a_{\text{lo}}\right)\right)\frac{j-\tfrac{1}{2}}{N}\right) \tag{11}$$

mit $a_{\text{hi}} = a_{\max}$ (fest) oder $S_b(t)$ (skaliert). Logit-Raster legen mehr Knoten an die Ränder, also an sehr niedrige und sehr hohe Niveaus.

4.5 Horizont

Mit $\Delta t = 1$ wird die Änderung über ein Jahr gemessen und über (6) in eine Quartalsrate umgerechnet. Ein längerer Horizont verdünnt weniger durch unbewegte Knoten, mittelt aber über mehr Geschichte und lässt weniger Zeitpunkte zu.

4.6 Läufe

Nur die Zeile mit dem höchsten Budget je Lauf (`budget = max(budget)` je `run_id`), danach dieselbe Aufbereitung. Damit fallen die simulierten Kürzungen (2.6) weg.

4.7 Release-Daten

Die Präfix-Zuordnung in 2.4 vergibt einigen Modellen das Datum eines früheren, ähnlich benannten Eintrags. Epochs Register enthält die richtigen Daten.

Tabelle 6: Korrigierte Release-Daten

Modell	im Code (Herkunft)	korrigiert
o3	31.01.2025 (o3-mini)	16.04.2025
o1	12.09.2024 (o1-mini)	17.12.2024
QwQ-32B	28.11.2024 (Preview)	05.03.2025
gemini-2.5-pro, gemini-2.5-flash	Preview-Datum	17.06.2025
gemini-1.5-pro-001	Preview-Datum	14.05.2024

Der o3-Fehler steht wörtlich im Bericht: „On January 31, 2025, OpenAI released … o3“. Ein freundlicher Hinweis, kein Argument: die Wirkung auf die Hauptzahl ist klein (Tabelle 7).

4.8 Mittelwert oder Median

Statt $\bar{\delta}$ in (6) der Median aller δ_{jk}. Der Median beschreibt den mittleren Knoten, also den Normalfall eines Niveaus in einem Quartal. Der Mittelwert beschreibt den Durchschnitt einschließlich weniger großer Sprünge.

4.9 Das Kombinationsgitter

Alle Achsen außer der Benchmark-Auswahl lassen sich frei kombinieren: 4 Preise × 2 Anbieter × 4 Raster × 2 Horizonte × 2 Lauf-Varianten × 2 Datierungen × 2 Mittel = **512 Rechenwege**. Für jeden wird (7) über die gewählten Benchmarks gebildet. Wie im Rechner gehen dabei die je Benchmark auf 0,1 Prozentpunkte gerundeten Raten ein.

> WICHTIG
>
> Die 512 Rechenwege sind keine Stichprobe. Sie sind weder unabhängig noch gleich plausibel. Wir zählen sie nur, um die Bandbreite zu zeigen, die vertretbare Entscheidungen erzeugen. Eine Wahrscheinlichkeit ist das nicht.

ABSCHNITT 5

Ergebnisse

5.1 Einzelne Spezifikationen

Jede Zeile ändert genau eine Annahme gegenüber Epoch (außer „üblich, große drei“: zwei).

Tabelle 7: Rückgang in % pro Quartal je Benchmark und Jahresfaktor

Spezifikation	AIME	Chess	FM 1–3	GPQA	Mystery	Mittel	pro Jahr
Epoch	47,1	43,0	53,1	47,0	44,0	46,8	12,5× billiger
Nur große drei	41,7	41,9	51,4	43,6	29,6	41,7	8,6× billiger
Üblicher Preis (9)	9,8	−3,9	15,7	14,0	18,1	10,8	1,6× billiger
Üblich, große drei	9,3	−1,2	−14,0	10,7	0,7	1,1	praktisch 1,0×
Median statt Mittel	0,0	0,0	28,9	0,0	0,0	5,8	1,3× billiger
Raster skaliert	60,8	35,2	67,8	58,0	50,4	54,4	23,2× billiger
Raster logit, fest	41,0	36,2	54,0	43,1	40,9	43,0	9,5× billiger
Raster logit, skaliert	53,7	34,2	63,8	49,0	44,5	49,0	14,8× billiger
Horizont ein Jahr	42,1	41,0	51,6	39,2	47,2	44,2	10,3× billiger
Nur vollständige Läufe	55,6	42,1	51,6	58,7	46,0	50,8	17,1× billiger
Release-Daten korrigiert	44,5	45,0	49,6	45,1	48,2	46,5	12,2× billiger
Lebensdauer 24 Monate	47,0	43,0	53,1	46,9	43,9	46,8	12,5× billiger
Lebensdauer 12 Monate	46,5	42,7	53,1	46,6	38,7	45,5	11,4× billiger

Negative Werte: teurer. Mittel und Jahresfaktor aus den ungerundeten Werten. Quelle: precompute.py. „Praktisch 1,0×“: 100 € werden 95,7 €, innerhalb von ±5 % pro Jahr, die der Rechner als „praktisch unverändert“ ausweist.

Harmlos sind Datierung, Horizont, Lebensdauer und die simulierten Läufe. Sie ändern wenig, teils machen sie die Zahl sogar größer. **Entscheidend** sind: welcher Preis, welche Anbieter, welches Mittel, und (5.4) welche Aufgaben.

5.2 Das Kombinationsgitter

Über alle 512 Rechenwege auf Epochs fünf Benchmarks reicht das Ergebnis von **26,4-mal billiger bis 1,3-mal teurer pro Jahr**. 53 Rechenwege zeigen steigende Kosten.

Was dieselben Daten alles hergeben

Jeder Punkt ist eine andere, vertretbare Art zu rechnen, 512 insgesamt. Links wird KI teurer, rechts billiger. Ergebnis: alles von 1,3× teurer pro Jahr bis 26,4× billiger pro Jahr. 53 Rechenwege zeigen steigende Kosten. Das ist keine Wahrscheinlichkeit, sondern die Bandbreite.



Abbildung 1: Jeder Punkt ist ein Rechenweg, logarithmische Achse. Links: teurer, rechts: billiger. Markiert: Epochs Wahl. Aus dem Rechner im Artikel.

Rechnet man mit ungerundeten Benchmark-Raten statt mit den auf 0,1 Prozentpunkte gerundeten Werten des Rechners, zeigen 54 statt 53 Rechenwege steigende Kosten. Die Spanne bleibt gleich.

5.3 Drei Kundenprofile

Statt einer Zahl drei Fragen: Was zahlt jemand in einem Jahr für eine Rechnung, die heute 100 € beträgt? Alle drei auf Epochs fünf Benchmarks.

Der Schnäppchenjäger

Epochs Spezifikation unverändert: $F = 12{,}52$, Rechnung 8,0 €. Spanne aus Epochs vier Rastern (4.4): 4,3 € (skaliert) bis 10,5 € (logit, fest).

Die normale Firma

Üblicher Preis (9) und nur große drei (4.3), sonst wie Epoch: $\bar{r} = 1{,}1\ \%$ pro Quartal, $F = 1{,}05$, Rechnung 95,7 €, gerundet 96 €. Spanne aus Abschnitt 6.1: 86 bis 119 €.

Die Anspruchsvolle

Nimmt jedes neue Modell in der stärksten Einstellung. Grundlage sind nur vollständige Läufe. Je Modell m der Lauf mit der höchsten Trefferquote, Kosten c_m^{best}. Je Benchmark eine Regression über die Modelle:

$$\ln c_m^{\text{best}} = \alpha + \beta_b t_m + \varepsilon_m, \qquad g_b = e^{\beta_b}, \qquad G = \exp\left(\frac{1}{5}\sum_b \beta_b\right) \tag{12}$$

Kleinste Quadrate. g_b: AIME 4,22, Chess 8,09, FrontierMath 1,66, GPQA 3,92, Mystery 1,88. $G = 3{,}34$, Rechnung 334 €. Die Tokens pro Aufgabe wachsen auf jedem der acht Benchmarks mit Daten, um das 1,5- bis 5,6-Fache pro Jahr (steigen.py). Unsicherheit: Abschnitt 6.2.

Tabelle 8: Kundenprofile

Profil	Definition	Faktor pro Jahr	100 € werden	Spanne
Schnäppchenjäger	Epoch: Rekordpreis, alle Anbieter	12,5× billiger	8 €	4 bis 11 €
Normale Firma	üblicher Preis, große drei	praktisch 1,0×	96 €	86 bis 119 €
Anspruchsvolle	stärkste Einstellung, Trend über Modelle	3,3× teurer	334 €	252 bis 458 €

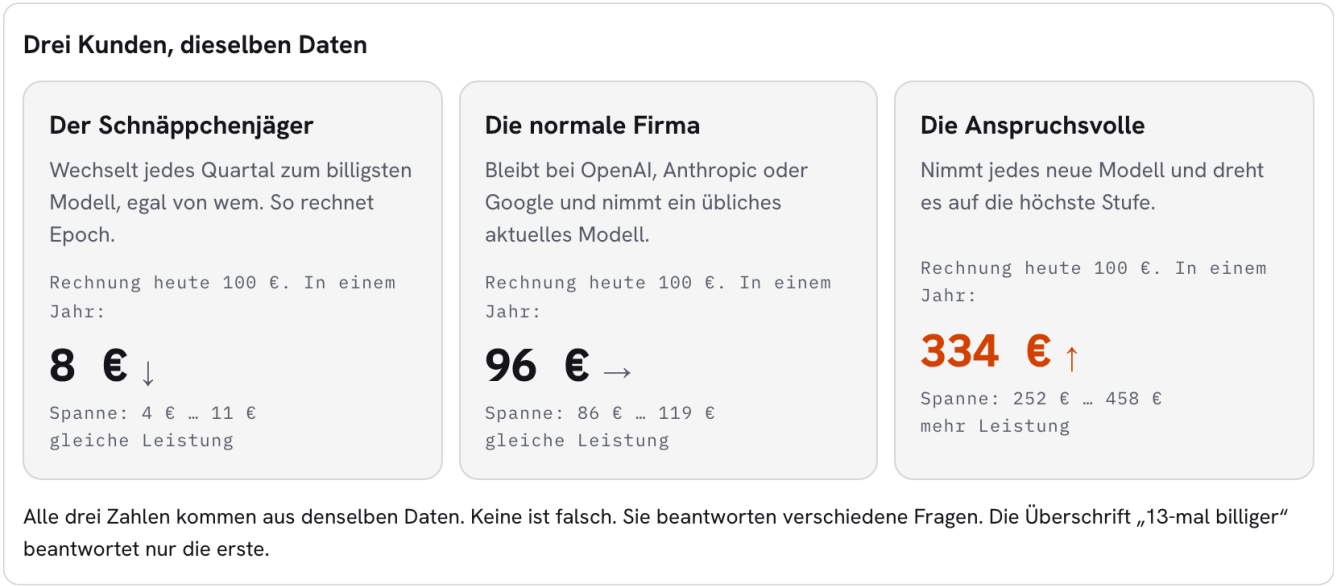

Abbildung 2: Die drei Kundenprofile im Rechner des Artikels.

Die Spannen zeigen: Die genauen Zahlen wackeln, die Richtung nicht. Der Schnäppchenjäger zahlt in jeder Rechnung deutlich weniger, die normale Firma in jeder Variante ungefähr gleich viel, die Anspruchsvolle in jeder Rechnung mehr. Die Anspruchsvolle bekommt dafür mehr Leistung. Ihre Zahl ist kein Preis für feste Leistung, sondern die Entwicklung ihrer Rechnung.

5.4 Alltagsnahe Aufgaben und alle Benchmarks

Tabelle 9: Andere Benchmark-Auswahl

Auswahl	Epoch-Spezifikation	pro Jahr	üblich, große drei	pro Jahr
Fünf Haupt-Benchmarks	46,8 %	12,5× billiger	1,1 %	praktisch 1,0×
Alltagsnah: SWE-bench Verified (27,5 %), SimpleQA (40,5 %)	34,0 %	5,3× billiger	−1,0 %	1,04× teurer
Alle elf Benchmarks	44,4 %	10,5× billiger	0,5 %	praktisch 1,0×
Nur FrontierMath 1 bis 3	53,1 %	20,7× billiger	−14,0 %	**1,7× teurer**

Rückgang in % pro Quartal. Quelle: precompute.py.

5.5 Leistungsniveaus

Epoch mittelt über alle Niveaus, von „kaum eine Aufgabe richtig“ bis zum Bestwert. Wir haben die Knoten aus (4) nach absolutem Niveau gruppiert und (6) je Gruppe gebildet (niveaus.py), dann über die Benchmarks gemittelt. Für den Vergleich der Kunden (niveaus_kunden.py) gilt ein Benchmark je Gruppe nur mit mindestens 20 Knoten, gemittelt wird geometrisch über die Jahresfaktoren.

Tabelle 10: Rückgang nach Leistungsniveau (um Raten bereinigt)

Niveau	Epoch, pro Jahr	Anteil der Knoten (AIME / Chess / FM / GPQA / Mystery)	Schnäppchenjäger	Normale Firma
unter 25 %	9,3×	32 / 67 / 55 / 43 / 100 %	9,6×	1,0×
25 bis 50 %	11,2×	28 / 29 / 34 / 27 / 0 %	11,3×	1,2×
50 bis 75 %	13,1×	27 / 4 / 10 / 23 / 0 %	20,2×	1,6×
ab 75 %	30,3×	13 / 0 / 2 / 7 / 0 %	42,2×	1,2×

Faktor: wie viel billiger pro Jahr. Spalte „Epoch“: arithmetisches Mittel der Quartalsraten über die Benchmarks mit Knoten im Band. Kundenspalten: geometrisches Mittel über Benchmarks mit mindestens 20 Knoten (ab 75 %: drei bzw. zwei Benchmarks).

Das spricht für Epoch: Hohe Leistung wurde am billigsten Rand sogar schneller billiger. Die Vermutung, die 13 komme von trivialen Niveaus, trifft nicht zu. Für die normale Firma wird hohe Leistung ab 75 % aber nur 1,2-mal billiger pro Jahr. Je höher die Leistung, desto größer der Abstand zwischen Überschrift und Rechnung. Einschränkungen: Hohe Niveaus sind erst kürzlich erreicht und dünn belegt. Gemessen wird also vor allem die Phase, in der ein neuer Bestwert seinen Aufpreis verliert. AIME ist ausgereizt (100 %). Auf dem Mystery-Benchmark, auf den niemand gezielt trainieren kann, erreichen die besten Modelle höchstens rund 30 % (roh 33 %, um Raten bereinigt 26 %).

5.6 Konzentration und Zeitverlauf

Auf vier von fünf Benchmarks bewegt sich der Rekordpreis an weniger als der Hälfte der Knoten (Tabelle 4). Der Median der Knoten ist 0, der Mittelwert besteht aus wenigen großen Sprüngen. Zuordnung (deep.py): Für jeden Knoten mit $\delta < 0$ wird der Rückgang dem neuen Rekordhalter zugerechnet. Einfluss: die acht wichtigsten Modelle je Benchmark einzeln weglassen und neu rechnen.

- **Mystery:** 37 Modelle im Datensatz, 9 tragen den Rückgang, die fünf wichtigsten 96 %.
- **Chess:** die drei wichtigsten 58 %. Ohne Gemini 3.5 Flash fällt die Rate von 43,0 auf 35,1 % pro Quartal, ohne GPT-4o steigt sie auf 50,6 %.
- **FrontierMath:** ohne GPT-5-mini 58,2 %, ohne GPT-5.6 Luna 49,9 %.

Tabelle 11: Rückgang nach Kalenderhalbjahr des Knotens (Mittel der fünf Benchmarks)

Halbjahr	1/2023	2/2023	1/2024	2/2024	1/2025	2/2025	1/2026
pro Jahr billiger	5,9×	7,4×	132×	28×	8,0×	3,6×	12,0×

Halbjahr nach Kalenderdatum des Ausgangszeitpunkts. Ein Mittelwert über Phasen, die um den Faktor 40 auseinanderliegen.

IN EINFACHEN WORTEN

Ein Millionär betritt eine Kneipe. Der Durchschnittslohn im Raum schießt nach oben. Keiner am Tresen verdient einen Cent mehr. Der Preis gleicht keinem Zug, der gleichmäßig fährt, sondern einem Aufzug, der meistens steht und manchmal fünf Stockwerke auf einmal fällt.

ABSCHNITT 6

Robustheit

6.1 Normale Firma: Fensterlänge und einzelne Modelle

Zwei Prüfungen (robust.py). Erstens die Fensterlänge in (9): 6, 12, 18 oder 24 Monate statt 12. Zweitens Leave-one-out: jedes der 75 Modelle der großen drei auf den fünf Benchmarks einmal weglassen und die Rate neu berechnen.

Tabelle 12: Rechnung der normalen Firma in einem Jahr (heute 100 €)

Prüfung	Ergebnis
Fenster 12, 18 oder 24 Monate	93 bis 96 €
Fenster 6 Monate	119 €
Leave-one-out über 75 Modelle	86 bis 109 €
Ausgewiesene Spanne (Minimum und Maximum aller Prüfungen)	86 bis 119 €

Die Aussage „praktisch kein Rückgang“ hält. Grenze: Das ist ein Modellergebnis, keine Zahl aus einer echten Buchhaltung (Abschnitt 11).

6.2 Anspruchsvolle: Bootstrap

Für (12): je Benchmark 1.000-mal die Modelle mit Zurücklegen ziehen, die Regression neu schätzen, G neu bilden (Zufallsstartwert 1). Ziehungen mit weniger als zwei verschiedenen Zeitpunkten fallen weg. Das 5. und 95. Perzentil ergeben das 90-%-Intervall: **2,52 bis 4,58**, also 252 bis 458 €. Dass es teurer wird, ist damit sicher.

Für „immer nur das jeweils stärkste Modell“ (nur Modelle, die beim Release einen neuen Bestwert setzen) gibt es zu wenige Daten: 2 bis 11 Modelle je Benchmark, Bootstrap 0,5 bis 2,9. Deshalb ist das Profil als „jedes neue Modell in höchster Einstellung“ definiert, nicht als „immer das Beste“.

6.3 Schnäppchenjäger

Die Spanne kommt aus Epochs eigenen vier Rastern (4.4): 4,3 bis 10,5 €. Sie ist keine statistische Unsicherheit, sondern zeigt, wie stark Epochs eigene Gestaltungsentscheidungen die Zahl bewegen.

6.4 Unempfindliche Annahmen

Tabelle 13: Weitere Einzelprüfungen (scenarios.py)

Prüfung	Mittel % pro Quartal	pro Jahr
Epoch	46,8	12,5×
Ohne Raterei-Bereinigung	46,2	11,9×
Ohne Preise aus GPU-Miete	46,8	12,5×
Nur Release-Daten o3, o1, QwQ korrigiert	46,6	12,3×
Horizont ein halbes Jahr	45,4	11,3×
Horizont anderthalb Jahre	44,7	10,7×
Ohne GPT-5.6-Familie	46,0	11,7×

6.5 Warum es keine Fehlerbalken für die Hauptzahl gibt

Echte Konfidenzintervalle für Epochs Hauptzahl gibt es nicht, auch bei Epoch nicht. Die Studie begründet das: Ein Rekordpreis ist ein Extremwert, den ein einziger günstiger Lauf festlegt, und das übliche Bootstrap-Verfahren funktioniert dafür nicht. Epoch schreibt: „We abandon the formal paradigm of statistical inference." Wir weisen deshalb Spannen aus Spezifikationen aus, keine Intervalle.

ABSCHNITT 7

Zerlegung: Tokenmenge und Tokenpreis

Am Rekordpreis gilt: Kosten einer Aufgabe = Tokens pro Aufgabe × Preis pro Token. Für jeden Knoten vergleichen wir den Rekordhalter jetzt und ein Quartal später und teilen die Log-Änderung auf (zerlegung.py; nur Läufe mit Tokenangabe größer 0):

$$\ln c = \ln n + \ln p \quad \Rightarrow \quad \delta_{jk} = \delta^{n}_{jk} + \delta^{p}_{jk}, \qquad s^{p}_{b} = \frac{\bar{\delta}^{p}}{\bar{\delta}} \tag{13}$$

n: Tokens pro Aufgabe, p: Preis pro Token (Kosten durch Tokens). s^{p}_{b}: Anteil des Rückgangs über den Tokenpreis, gemittelt über die fünf Benchmarks.

Tabelle 14: Zerlegung je Benchmark

Benchmark	Anteil Tokenpreis	Anteil Tokenmenge	nur weniger Tokens, pro Jahr
AIME	83 %	17 %	1,56×
Chess	40 %	60 %	3,83×
FrontierMath 1 bis 3	105 %	−5 %	0,85×
GPQA	89 %	11 %	1,32×
Mystery	60 %	40 %	2,53×
Fünf Haupt-Benchmarks	75 %	25 %	1,76×
SWE-bench Verified	117 %	−17 %	0,80×
SimpleQA	14 %	86 %	5,97×

Negativer Anteil Tokenmenge: die neuen Rekordhalter brauchen mehr Tokens als die alten (Programmieren, Forschungsmathematik). Letzte Spalte: $\exp\left(-4\bar{\delta}^{n}\right)$, Zeile „Fünf" als geometrisches Mittel.

Tokenpreis eingefroren: Bleibt nur der Mengeneffekt, wird der Schnäppchenjäger nicht 12,5-mal, sondern 1,76-mal billiger pro Jahr. Aus 100 € werden 57 € statt 8 €. Grenze: „Billigere Tokens" enthält auch den Wechsel zu kleineren, effizienteren Modellen, nicht nur Strom- und Chippreise.

ABSCHNITT 8

Szenario-Rechner

Der Rechner im Artikel (Komponente `ThoughtPriceLab`) schreibt die drei Profile unter veränderten äußeren Bedingungen fort. Das ist ausdrücklich keine Messung. Die Umrechnung von Input-Preisen auf die Rechnung verwendet grobe Kostenanteile eines KI-Rechenzentrums: Chips und Server 55 %, Strom 15 %. Eine saubere öffentliche Zahl dafür gibt es nicht.

$$W = \left(1 + 0{,}55 \cdot \chi + 0{,}15 \cdot \varepsilon\right) \cdot \left(1 + \sigma\right), \qquad \text{Rechnung}_{p} = 100\,€ \cdot m_{p} \cdot W \tag{14}$$

χ: Änderung der Chip- und Speicherpreise, ε: Änderung des Strompreises, σ: Preisaufschlag, wenn Subventionen der Investoren enden, jeweils als Anteil. m_{p}: Jahresmultiplikator des Profils, $1 / 12{,}52$, $1 / 1{,}05$ und $3{,}34$. Bei eingefrorenem Tokenpreis ersetzt $1 / 1{,}76$ den Multiplikator des Schnäppchenjägers (Abschnitt 7). Die Spannen werden mit demselben W multipliziert.

Tabelle 15: Voreingestellte Szenarien, Rechnung in einem Jahr (heute 100 €)

Szenario	W	Schnäppchenjäger	Normale Firma	Anspruchsvolle
Alles bleibt, wie es ist	1,000	8 €	96 €	334 €
Strom 50 % teurer ($\varepsilon = 0{,}5$)	1,075	9 €	103 €	359 €
Chips 40 % teurer ($\chi = 0{,}4$)	1,220	10 €	117 €	408 €
Subventionen enden, 50 % Aufschlag ($\sigma = 0{,}5$)	1,500	12 €	144 €	501 €
Tokenpreise fallen nicht mehr (gemessen)	1,000	57 €	96 €	334 €
Alles zusammen	1,943	**110 €**	**186 €**	**649 €**

Die Voreinstellungen orientieren sich an Entwicklungen 2025 und 2026: Großhandelsstrom im größten US-Strommarkt PJM +76 % von Anfang 2025 bis Anfang 2026 (Monitoring Analytics); SemiAnalysis-Index für H100-Jahresverträge +38 % von Oktober 2025 bis März 2026; Speicherchips Anfang 2026 fast doppelt so teuer wie ein Quartal zuvor (TrendForce). OpenAI setzte laut Wall Street Journal im zweiten Quartal 2026 6,7 Mrd. Dollar um und rutschte tiefer in die Verluste. Anthropic erreichte laut Financial Times im dritten Quartal 2026 einen bereinigten operativen Gewinn. Alle über Sekundärquellen.

EINORDNUNG

Ist „Alles zusammen“ wahrscheinlich? Vielleicht nicht. Unmöglich ist es nicht. Die Aussage „Die Richtung ist eindeutig“ setzt aber voraus, dass es unmöglich ist.

ABSCHNITT 9

Externe Reihen

Wer einen vergangenen Preistrend fortschreibt, nimmt an, dass die Bedingungen gleich bleiben. Die folgenden Reihen zeigen, wie sich diese Bedingungen tatsächlich bewegt haben. Für jede Reihe: Quelle, Verarbeitung, Ergebnis.

9.1 Strom USA nach 1973

Quelle: Epochs Arbeitsmappe, Blatt „Electricity“. Epochs Grafik nutzt eine kombinierte Realreihe (Hirsh bis 1922, EEI 1926 bis 1973), die 1973 endet. Dasselbe Blatt enthält Hirshs Nominalpreise bis 2025, das Blatt „CPI“ den Verbraucherpreisindex (CPIAUCSL, 1982 bis 1984 = 100). **Verarbeitung** (strom_nach_1973.py): Realpreis = Hirsh real, wo vorhanden, sonst Nominalpreis / CPI-Jahresmittel × 100. Ab 1974 wird diese Reihe so verkettet, dass sie 1973 genau auf Epochs Reihe trifft. Index 1892 = 100.

Tabelle 16: Haushaltsstrom USA, inflationsbereinigt

Zeitraum	Veränderung
1892 bis 1973 (Epochs Kurve)	−98 %
1973 bis 1984 (Ölkrise)	+29 %
1973 bis 2002 (Tiefpunkt)	−12 %
2002 bis 2025	+14 %
1973 bis 2025	±0 %

9.2 Großbritannien

Quelle: Fouquet (2011), Strompreis real in Pfund von 2000 je Tonne Öläquivalent, 1883 bis 2008, ab 2009 Eurostat (Haushalte, real, Abschnitt 9.3) verkettet auf 2008 (epoch_chart.py). **Ergebnis:** von 1900 bis 1973 zwanzigmal billiger, bis 1980 wieder 30 % teurer, 2008 real so teuer wie 1973, 2019 15 % teurer als 1973.

9.3 Westeuropa

Quelle: Eurostat `nrg_pc_204`, Haushalte mit 2.500 bis 4.999 kWh Jahresverbrauch (Band DC), alle Steuern und Abgaben, €/kWh; `prc_hicp_aind`, HICP 2015 = 100. **Verarbeitung** (eurostat.py, epoch_chart.py): Jahreswert = Mittel der beiden Halbjahre, nur Jahre mit beiden Halbjahren, real = nominal / HICP × 100. Vergleich erstes volles Jahr gegen 2025.

Tabelle 17: Haushaltsstrom seit 2007/08

Land	ab	nominal	real	real, Spitze
Frankreich	2008	×2,17	+61 %	+79 % (2024)
Spanien	2008	×1,81	+32 %	+75 % (2022)
Deutschland	2007	×1,87	+26 %	+39 % (2023)
EU-27	2007	×1,72	+13 %	+18 % (2023)
Italien	2008	×1,47	+7 %	+26 % (2022)
Niederlande	2007	×1,41	−8 %	+20 % (2023)

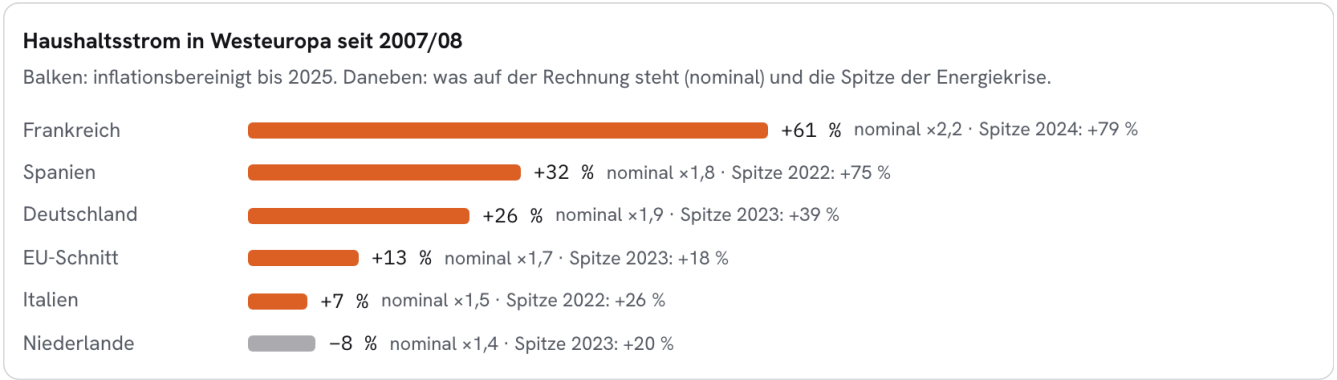

Abbildung 3: Haushaltsstrom in Westeuropa, inflationsbereinigt bis 2025, daneben nominal und die Spitze der Energiekrise. Eurostat, eigene Berechnung.

9.4 bis 9.7 Die letzten fünf Jahre

Reihen ab Oktober 2021, aufbereitet mit welt_reihen.py. Veränderung über zwölf Monate: letzter Wert gegen den letzten Wert, der mindestens ein Jahr zurückliegt (eine Woche Toleranz für Wochenreihen).

Tabelle 18: Äußere Bedingungen

Reihe	Quelle und Verarbeitung	Stand	12 Monate
9.4 Großhandelsstrom Deutschland	Ember, Day-Ahead, monatlich. Nur vollständige Monate: ein Monat, den die Tagesdaten noch nicht abgeschlossen haben, wird verworfen. Dazu die letzten 56 Tage täglich.	145,4 €/MWh (09/2026)	+72,5 %
9.5 Haushaltsstrom USA	FRED APU000072610 (BLS), $/kWh, monatlich	0,196 $/kWh (09/2026)	+4,3 %; seit 10/2021 +38 %
9.6 Miete KI-Chip	getdeploying.com, Nvidia H100, On-Demand, Median über 40 Anbieter, wöchentlich. Freie Wochendaten erst ab Oktober 2025.	3,81 $/h (19.10.2026)	+27,5 %
9.7 Leitzinsen	FRED DFF (Fed), ECBDFR (EZB-Einlagesatz), wöchentlich letzter Wert	Fed 3,88 %, EZB 2,50 %	10/2021: 0,08 % und −0,50 %

Großhandel Deutschland: Spitze 08/2022 mit 463,6 €/MWh, Tief 02/2024 mit 61,2 €/MWh, letzter Tageswert 07.10.2026. H100: +8,7 % in der letzten Woche. Leitzinsen über dem Niveau von 2021, aber unter der Spitze von 2023 (Fed 5,33 %, EZB 4,00 %).

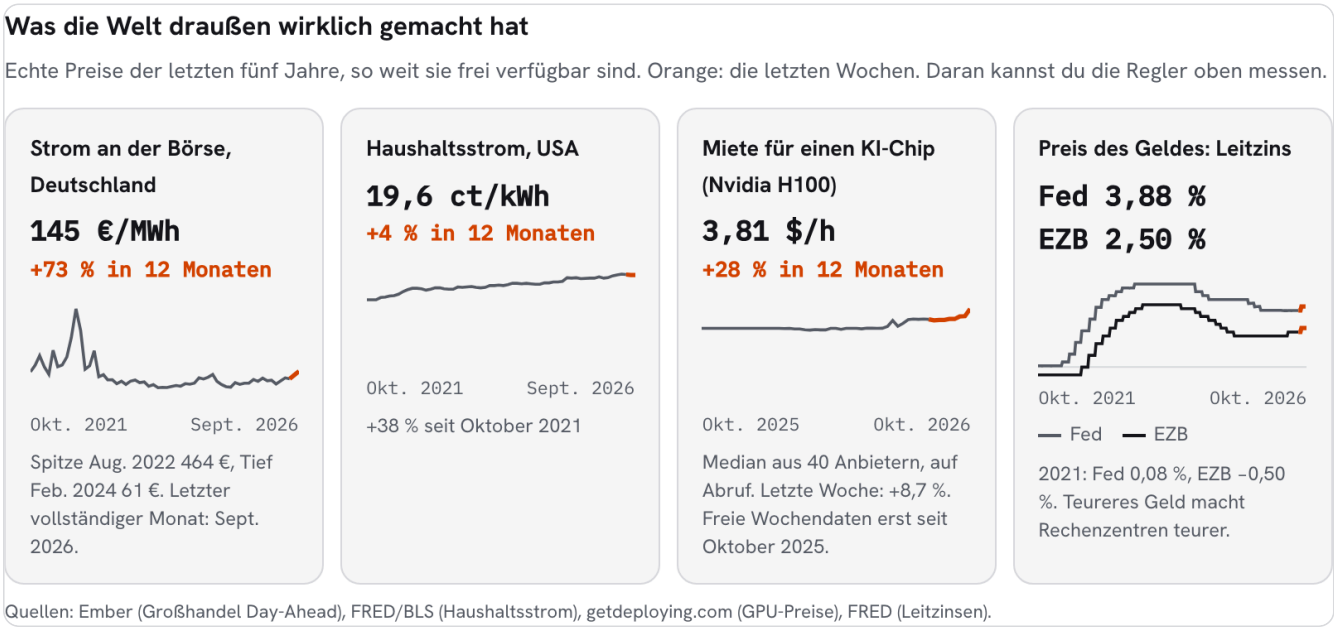

Abbildung 4: Die vier Reihen im Rechner des Artikels. Orange: die letzten Wochen. Achsen beginnen bei null.

ABSCHNITT 10

Technologievergleich

10.1 Die abgeschnittenen Kurven

Epochs Abbildung 1 zeigt, ab Beginn des jeweiligen Preisverfalls, KI neben Strom, Lithium-Batterien, Rechenleistung und DNA-Sequenzierung. Wir haben sie aus dem Blatt „Figure data“ nachgebaut und mit den Reihen aus Abschnitt 9 zu Ende gezeichnet. In unserem Nachbau reicht die Achse über den Startpreis hinaus, damit eine steigende Linie überhaupt sichtbar wird. Epochs Achse reicht über zwölf Zehnerpotenzen und endet nach oben am Startpreis. Auf so einer Skala ist ein Anstieg um 30 % wenige Pixel hoch.

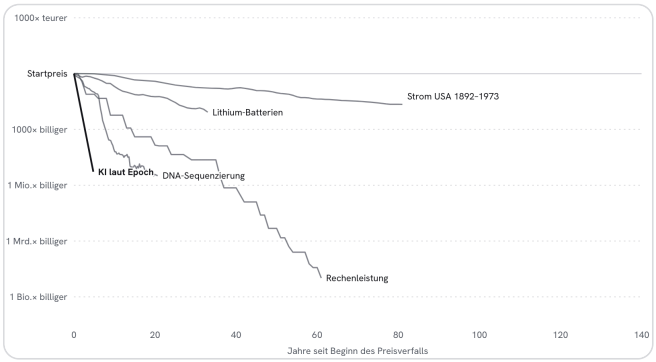

Abbildung 5a: Epochs Abbildung, nachgebaut aus Epochs Daten.

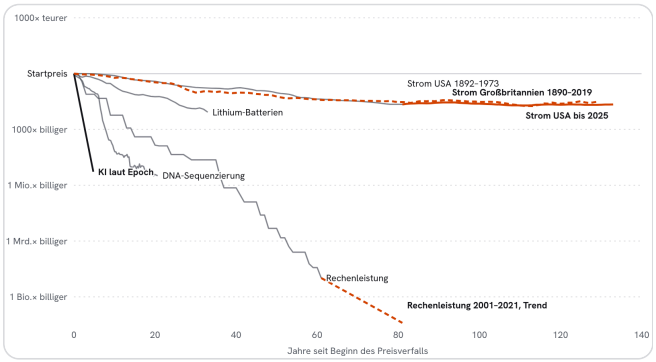

Abbildung 5b: Orange: was danach kam.

- **Strom** endet 1973. Der Rückgang stand da seit drei Jahren still, im Herbst kam die Ölkrise. Fortsetzung: Abschnitt 9.1 und 9.2.
- **Rechenleistung** endet 2001, laut Epoch „merely for lack of newer data“. Für danach verwenden wir Epochs eigene GPU-Studie (Hobbhahn und Besiroglu 2022): doppelte Rechenleistung pro Dollar etwa alle 2,5 Jahre (2006 bis 2021), also Faktor $2^{1/2{,}5} \approx 1{,}32$ pro Jahr, rund 24 % billiger. Eingezeichnet als Trend, nicht als Messreihe.
- **DNA** (bis 2022) und **Batterien** (bis 2024) enden mit den neuesten verfügbaren Daten.
- Die steile **KI-Linie** beginnt 2021. Gemessen ist sie erst ab 2023, davor mit 47 % pro Quartal zurückgerechnet.

10.2 Faires Zeitfenster

Drei Jahre KI werden mit 20 bis 80 Jahren anderer Technologien verglichen. Viele Technologien werden in einer frühen Phase besonders schnell billiger. Fair ist der Vergleich mit dem schnellsten Dreijahresfenster jeder Reihe. Epochs Arbeitsmappe enthält dafür eine Spalte „Fastest 3-year decline“, die im Bericht nicht vorkommt. Unsere Rechnung (techwin.py), mit Log-Raten pro Quartal:

$$\rho_{\text{ganz}} = -\frac{\ln p_{\text{Ende}} - \ln p_{\text{Start}}}{4\left(t_{\text{Ende}} - t_{\text{Start}}\right)}, \qquad \rho_{3\text{J}} = \max_{i} -\frac{\ln p_{j(i)} - \ln p_{i}}{4\left(t_{j(i)} - t_{i}\right)} \tag{15}$$

$j(i)$: erster Beobachtungspunkt mindestens drei Jahre nach i (höchstens vier). Jahresfaktor $e^{4\rho}$. „KI schneller“ = ρ_{KI} / ρ mit $\rho_{\text{KI}} = -\ln\left(1 - 0{,}47\right) = 0{,}635$.

Tabelle 19: Tempo der Technologien, ganzer Zeitraum und schnellste drei Jahre

Technologie	ganzer Zeitraum	schnellste 3 Jahre	KI schneller, Bericht	KI schneller, fair
DNA-Sequenzierung	1,84× pro Jahr	6,70×	4,2-fach	1,3-fach
Rechenleistung	1,52×	3,17×	6,0-fach	2,2-fach
Lithium-Batterien	1,15×	1,37×	17,9-fach	8,0-fach
Strom	1,05×	1,15×	50,9-fach	18,3-fach

Mit dem üblichen statt dem billigsten Modell (1,6-mal pro Jahr) liegt KI etwa beim langfristigen Tempo von Rechenleistung (1,52-mal).

10.3 Dieselben Daten, andere Darstellung

Tempo je Jahrzehnt (epoch_chart.py): mittlere jährliche Änderung innerhalb jedes Kalenderjahrzehnts, $100 \cdot \left(\left(v_z / v_a \right)^{1/(z-a)} - 1 \right)$, mit log-linearer Interpolation an den Jahrzehntgrenzen und nur für Jahrzehnte, die eine Reihe mindestens fünf Jahre abdeckt. Der Rechenleistungs-Trend nach 2001 bleibt außen vor, weil er keine Messreihe ist.

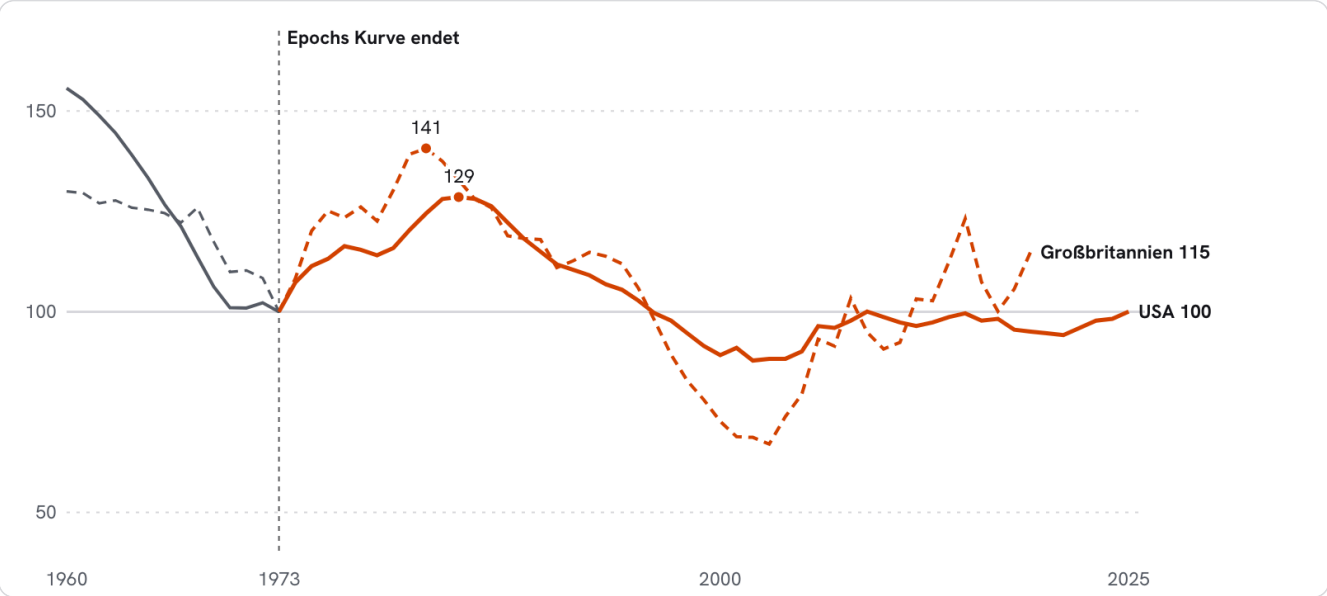

Abbildung 6: Strom ab 1960 auf normaler Skala, 1973 = 100. Grau: was Epoch zeigt. Orange: was danach kam. Durchgezogen USA, gestrichelt Großbritannien.

	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010	2020
Strom USA	−1,9	−7,0	−6,9	−3,1	−2,7	−8,0	−3,5	−4,2	**+1,4**	−0,4	−2,2	**+1,0**	−0,4	**+1,1**
Strom GB	−5,1	−5,4	−12	−0,5	−4,4	−3,5	−1,7	−1,7	**+1,7**	−1,6	−4,1	**+2,2**	**+2,7**	
Rechenleistung						−40	−31	−17	−29	−40	−38			
Batterien											−13	−15	−14	
DNA												−61	−37	
KI laut Epoch														−92
KI, normale Firma														−4,3
KI, voll aufgedreht														**+234**

Spalten: Jahrzehnte (1890 = 1890er). Prozent pro Jahr, Durchschnitt im Jahrzehnt. Orange: teurer geworden. KI in der Spalte 2020er: 2021 bis 2026, wie Epoch gerechnet.

Abbildung 7: Tempo je Jahrzehnt in % pro Jahr, billiger (−) oder teurer (+). Orange hinterlegt: teurer geworden. KI in der Spalte 2020er: 2021 bis 2026, wie Epoch gerechnet.

Jede dieser Technologien hatte schnelle Phasen, langsame und Rückschläge. Die KI-Zahl beschreibt genau eine Phase von drei Jahren. Welches Bild in einem Bericht steht, ist eine Entscheidung, keine Messung.

10.4 Marktpreis gegen Minimum

Bei Strom und Batterien zählen Marktpreise, also was Kunden im Schnitt bezahlt haben. Bei KI zählt das Minimum über alle Anbieter. Das sind genau die „apples and oranges“, von denen Epoch spricht.

EXKURS: BAHN UND FLUG

Die Deutsche Bahn wirbt mit ICE-Fahrten ab 8,99 €. Das Angebot gibt es, aber es ist begrenzt, und kaum jemand zahlt diesen Preis. Statistisch ist so ein Preis ein Ausreißer: ein kleines Kontingent am Rand. Wer wissen will, was Bahnfahren kostet, sortiert ihn aus. Die Studie misst dagegen nur den Rand. Zum Fahrplanwechsel im Dezember 2024 erhöhte die Bahn den Flexpreis im Schnitt um 5,9 %, die Super-Sparpreise blieben stabil. Als Grund nannte sie „deutlich gestiegene Kosten, insbesondere infolge der jüngsten Tarifabschlüsse“, zwei Jahre zuvor die Energiekosten. Im Dezember 2025 verzichtete sie im Fernverkehr auf eine Erhöhung. Das Muster bleibt: Das Lockangebot bleibt billig, der Preis, den die meisten zahlen, steigt.

Ein Inlandsflug in den USA ist inflationsbereinigt in dreißig Jahren um 37 % billiger geworden, im Schnitt von 618 auf 387 Dollar (BTS, Preise von 2025), gut anderthalb Prozent pro Jahr. Würde man den billigsten Aktionspreis verfolgen, käme eine spektakuläre Zahl heraus. Ein typischer Geschäftstarif kostete laut American Express schon Anfang der 2000er das Fünf- bis Sechsfache eines Urlaubertarifs. Mit der Methode „Aktionspreis gegen Durchschnittspreis“ wäre auch der Flug die schnellste Technologie aller Zeiten.

ABSCHNITT 11

Grenzen dieser Analyse

11.1 Feste Leistung

Alle Zahlen dieses Berichts halten die Leistung fest, Epochs und unsere Kundenprofile (mit Ausnahme der Anspruchsvollen). Für einen Preisvergleich ist das richtig. Problematisch wird es, wenn daraus eine Prognose wird, was KI ein Unternehmen künftig kostet. 2023 war ein brauchbarer Text beeindruckend, heute erwarten viele ganze Arbeitsabläufe.

Was man misst	Was passiert
Preis für gleichbleibende Leistung	sinkt stark
Anspruch an die Technik	steigt meistens
Tatsächliche Ausgaben	können steigen oder fallen

KERNSATZ

Wir messen den Preisverfall an den Anforderungen von gestern und verkaufen ihn als Kostenersparnis von morgen.

IN EINFACHEN WORTEN

Eine Postkutsche schaffte rund 10 Kilometer pro Stunde, der ICE fährt bis zu 300. Eine Kutschfahrt wäre heute spottbillig. Fährt nur keiner mehr. Nicht die Kutsche wurde teurer. Schnelleres wurde möglich, und weil es möglich wurde, haben es alle genutzt. Was möglich wird, wird genutzt. Und bezahlt.

11.2 Weitere Grenzen

- **Modellergebnisse, keine Buchhaltung.** Die normale Firma ist eine Rechnung über Epochs Daten. Eine echte Firma nutzt bestimmte Modelle für bestimmte Aufgaben. Wie oft welches Modell tatsächlich eingesetzt wird, steht in keinem Datensatz. Es gibt keine Nutzungsgewichte. Der übliche Preis ist eine zweite Lesart, keine bessere Wahrheit.
- **Modellauswahl.** „Üblicher Preis“ hängt davon ab, welche Modelle Epoch getestet hat.
- **Listenpreise, nominal.** Rabatte, Batch- und Cache-Preise, Verträge und Inflation sind nicht berücksichtigt. Preise stammen teils nicht vom Release-Zeitpunkt (2.3).
- **Benchmarks sind keine Arbeit.** Gemessen wird an Mathematik, Schach und Prüfungsfragen. Die beiden alltagsnäheren Tests zeigen den langsamsten Rückgang (5.4).
- **Dünne Daten an den Rändern.** Hohe Niveaus und Rekordhalter sind schwach belegt (5.5, 6.2).
- **Zerlegung.** Der Anteil 75 zu 25 schwankt stark je Benchmark (Tabelle 14).
- **Szenarien.** Kostenanteile und Voreinstellungen in Abschnitt 8 sind Annahmen. Einzelne Geschäftszahlen stammen aus Sekundärquellen.
- **Kombinationsgitter.** Die 512 Rechenwege sind keine Wahrscheinlichkeitsverteilung (4.9).
- **Epochs geglättete Modelle.** Die Aussage „66 % Rückgang pro Quartal direkt nach einem neuen Bestwert“ stammt aus Heatmaps (Epochs Abbildung 6). Dort liegt ein großer Teil der Fläche über der Rekordtreppe, also ohne Messwerte: AIME 49 %, Chess 57 %, FrontierMath 63 %, GPQA 40 % (heatarea.py). Dort rechnet das Modell nur hoch. Epochs eigene Tabelle 1 zeigt außerdem: Ein Modell über alle Modelle statt nur über den billigsten Rand ergibt 5,3 bis 10,2 % pro Quartal (1,2- bis 1,5-mal pro Jahr), mit erzwungener Hüllkurve 27 bis 38 %. Epoch verwirft beide Varianten mit Begründung. Es zeigt dennoch: Eine Formel ist keine Messung, sondern eine weitere Annahme.

11.3 Was die Daten nicht entscheiden

Ob man früh mit KI anfangen sollte. Das kann richtig sein, weil man lernt, welcher Prozess passt, wo die Daten liegen und wer das Ergebnis prüft. Das ist ein gutes Argument. Es braucht nur keine 13. Mit einer Zahl gestützt, die etwas anderes misst, wird es schwächer, nicht stärker. Dieselben Daten stützen ebenso das Abwarten: Neue Spitzenleistung wird laut Epoch anfangs um rund zwei Drittel pro Quartal billiger. Und für Anbieter beschreibt Epoch die Gewinne je Modell als „fleeting“. Ein Markt, in dem Leistung so schnell an Wert verliert, ist für Investoren ein Risiko, keine sichere Wette.

11.4 Andere Arbeiten

- **Toby Ord** (Oxford, Oktober 2025) zeigt, dass Fortschritt vor allem aus mehr Rechnen beim Antworten kommt: „So if it costs a factor of 10 or 100 times as much to answer each question, this really does affect their bottom line. And unlike training costs, these costs can't be made up in volume.“ Epoch dankt ihm im Bericht für den Einwand zum Technologievergleich und hat ihm öffentlich geantwortet. Redwood Research hält seine Methode für zu pessimistisch.
- **Gartner** (17.08.2026): Kosten pro agentischem Arbeitsablauf steigen bis 2028 auf mehr als das Fünffache, obwohl Tokens billiger werden („Inference Paradox“).
- **Gundlach et al.** (März 2026): ähnliche Methode, 5- bis 10-mal billiger pro Jahr statt 13. Epoch zitiert die Arbeit selbst.
- **Pasquale Pillitteri** (Blog, sinngemäß): Die Einschränkungen gehen beim Weiterteilen verloren. In der Buchhaltung einer Firma gibt es keine feste Leistung, man sollte mit Kosten pro Aufgabe planen.

> FAZIT
>
> Niemand weiß, wie sich KI-Preise in den nächsten Jahren entwickeln. Wir auch nicht. Am billigsten Rand fällt der Preis schnell, fast alle Rechenwege zeigen nach unten. Einen klaren Schluss auf „wird drastisch billiger“ geben die Daten aber nicht her. Sie geben her: Es kommt stark darauf an.

ABSCHNITT 12

Reproduzierbarkeit

Alle Zahlen lassen sich wiederholen. Die Skripte liegen im Projektordner unter `drafts/preis-des-denkens/analyse/`. Benötigt: Python mit `uv`. Aufruf jeweils `uv run --with pandas --with numpy python -I <skript>`, für Excel-Eingaben zusätzlich `--with openpyxl`. Pfade unten relativ zum Ordner `analyse/`, Epochs Repositorium daneben als `epoch-repo`.

```
git clone https://github.com/droodman/inference-cost epoch-repo   # Commit 9163c17
cd drafts/preis-des-denkens/analyse
D=../../../epoch-repo/data
uv run --with pandas --with numpy python -I core.py $D               # Tabelle 4
uv run --with pandas --with numpy python -I precompute.py $D lab.json # Tabellen 7, 9; 512 Rechenwege
uv run --with pandas --with numpy python -I scenarios.py $D          # Tabelle 13, scenarios.csv
uv run --with pandas --with numpy python -I mv2.py $D                # älteres 64er-Gitter, multiverse.csv
uv run --with pandas --with numpy python -I deep.py $D               # 5.6 Konzentration, Halbjahre
uv run --with pandas --with numpy python -I niveaus.py $D            # Tabelle 10, Epoch-Spalte
uv run --with pandas --with numpy python -I niveaus_kunden.py $D     # Tabelle 10, Kundenspalten
uv run --with pandas --with numpy python -I robust.py $D             # Tabelle 12, 6.2
uv run --with pandas --with numpy python -I steigen.py $D            # (12), steigen.json
uv run --with pandas --with numpy python -I zerlegung.py $D          # Tabelle 14, zerlegung.json
uv run --with pandas --with numpy python -I heatarea.py $D           # 11.2 Heatmap-Fläche
uv run --with pandas --with numpy python -I dates.py $D              # Tabelle 6, Datumsabgleich
X="$D/Cost declines by tech.xlsx"
uv run --with pandas --with numpy --with openpyxl python -I techwin.py "$X"           # Tabelle 19
uv run --with pandas --with numpy --with openpyxl python -I strom_nach_1973.py "$X" strom.json   # Tabelle 16
uv run --with pandas --with numpy python -I eurostat.py <nrg_pc_204.json> <prc_hicp_aind.json> eurostat_real.json
uv run --with pandas --with numpy --with openpyxl python -I epoch_chart.py "$X" epoch_chart.json  # Abb. 5 bis 7, Tabelle 17
uv run --with pandas --with numpy python -I welt_reihen.py <ember_monthly> <ember_daily> <gd_weekly> <fred_dir> welt.json
uv run --with pandas --with numpy python -I export_ts.py lab.json steigen.json zerlegung.json strom.json epoch_chart.json welt.json
data.ts
```

Tabelle 20: Skripte, Eingaben, Ausgaben

Skript	Was es tut	Ausgabe
`core.py`	Epochs Aufbereitung (`build_runs`) und modellfreie Rate (`decline`), Abschnitte 2 und 3	Rate, bewegte Knoten, Knoten je Benchmark
`precompute.py`	Alle Spezifikationen aus Abschnitt 4 für alle elf Benchmarks, Mittelwert und Median	`lab.json`
`export_ts.py`	Schreibt das Datenmodul des Rechners	`src/components/thought-price/data.ts`
`scenarios.py, mv2.py`	Einzelprüfungen, älteres Gitter aus 64 Kombinationen	`scenarios.csv`, `multiverse.csv`
`deep.py`	Zuordnung des Rückgangs zu Modellen, Weglassen einzelner Modelle, Halbjahre	Konsole
`niveaus.py`, `niveaus_kunden.py`	Rückgang je Leistungsniveau, Rekordpreis gegen üblichen Preis	Konsole
`robust.py`	Fensterlänge, Leave-one-out, Bootstrap der Anspruchsvollen	Konsole
`steigen.py`	Kosten- und Tokentrend der besten Einstellung je Modell	`steigen.json`
`zerlegung.py`	Tokenmenge gegen Tokenpreis	`zerlegung.json`
`heatarea.py, dates.py`	Datenlose Heatmap-Fläche, abweichende Release-Daten	Konsole
`techwin.py`	Ganzer Zeitraum und schnellstes Dreijahresfenster	Konsole
`strom_nach_1973.py`, `eurostat.py`	Strom USA nach 1973, Haushaltsstrom Europa real	`strom.json`, `eurostat_real.json`
`epoch_chart.py`	Kombinierte Grafik, Großbritannien, Europa, Tempo je Jahrzehnt	`epoch_chart.json`
`welt_reihen.py`	Reihen der letzten fünf Jahre	`welt.json`

Hinweis: scenarios.py verwendet für „große drei“ ein älteres, weiter gefasstes Namensmuster (einschließlich gpt-oss und chatgpt) und kommt dort auf 9,0× statt 8,6×. Maßgeblich für Bericht, Artikel und Rechner ist das Muster in precompute.py (4.3). Eingaben für eurostat.py und welt_reihen.py sind öffentliche Downloads (Eurostat-API, Ember, getdeploying.com, FRED) und nicht Teil des Repositoriums.

ANHANG A

Glossar

Benchmark
Ein Test für KI-Modelle: eine feste Sammlung von Aufgaben. Gemessen wird, wie viel Prozent ein Modell richtig löst.

Token
Ein Wortbaustein. Anbieter rechnen nach Tokens ab, ungefähr wie ein Taxi nach Kilometern. Denkt ein Modell länger, verbraucht es mehr Tokens.

Rekordpreis (Cost frontier)
Der billigste Preis, zu dem irgendein Modell ein bestimmtes Niveau schafft, Formel (2). Kann wie ein Weltrekord nur besser werden.

Bestwert (State of the Art)
Die höchste bis zu einem Zeitpunkt erreichte Trefferquote, Formel (3).

Knoten
Ein Punkt des Messgitters aus Leistungsniveau und Zeitpunkt, an dem die Änderung des Rekordpreises gemessen wird.

Geometrisches Mittel
Der Mittelwert der prozentualen Veränderungen, gebildet über Logarithmen. Wird von wenigen großen Sprüngen stark beeinflusst.

Median
Der Wert in der Mitte, wenn man alle der Größe nach sortiert. Ein Millionär in der Kneipe ändert ihn nicht.

Logit
Eine Umrechnung von Anteilen, die Werte nahe 0 und nahe 100 % auseinanderzieht.

Logarithmische Skala
Eine Achse, bei der jeder Schritt das Zehnfache ist. Gut für riesige Unterschiede, schlecht, um kleine Anstiege zu sehen.

Nominal und real
Nominal: was auf der Rechnung steht. Real: was es kostet, wenn man die allgemeine Teuerung herausrechnet.

Bootstrap
Man zieht aus den Daten viele neue Stichproben mit Zurücklegen, rechnet jedes Mal neu und schaut, wie stark das Ergebnis schwankt.

Leave-one-out
Man lässt nacheinander je ein Element weg, hier ein Modell, und schaut, wie stark sich das Ergebnis ändert.

Open Weights
Modelle, die man selbst betreiben kann. Oft billig, für viele Unternehmen wegen Betrieb, Datenschutz oder Verträgen aber keine echte Option.

ANHANG B

Sechs Fragen an jede Preisstatistik

1. **Wessen Preis?** Der billigste am Markt, ein üblicher oder Ihrer?
2. **Für welche Leistung?** Gemessen woran, und hat das mit Ihrer Arbeit zu tun?
3. **Durchschnitt oder Normalfall?** Oder sitzt ein Millionär in der Kneipe?
4. **Kann die Zahl überhaupt anders ausfallen?** Ein Weltrekord, der nur fallen kann, beweist nicht, dass etwas fällt.
5. **Womit verglichen?** Sprinter mit Sprinter oder Sprinter mit Marathonläufer? Und auf welcher Skala?
6. **Was muss gleich bleiben, damit es so weitergeht?** Strom, Chips, Investorengeld. Und wer garantiert das?

ANHANG C

Quellen

- Epoch AI (2026): The plunging price of thought. epoch.ai/publications/the-plunging-price-of-thought
- Epoch AI: Daten und Code. github.com/droodman/inference-cost, Commit 9163c17 (24.09.2026)
- Hobbhahn, M.; Besiroglu, T. (2022): Trends in GPU price-performance. Epoch AI. epoch.ai/blog/trends-in-gpu-price-performance
- Fouquet, R. (2011): Divergences in long run trends in the prices of energy and energy services. Review of Environmental Economics and Policy 5(2). Datensatz: LSE Grantham Institute.
- Eurostat: nrg_pc_204 (Haushaltsstrompreise), prc_hicp_aind (HICP).
- Ember: European wholesale electricity price data (monatlich, täglich). FRED: APU000072610, DFF, ECBDFR. getdeploying.com: GPU price dataset (weekly).
- Bureau of Transportation Statistics (2026): 2025 Annual Average Domestic Air Fare. bts.gov. Deutsche Bahn, Pressemitteilung „ICE-Fahren jetzt ab 8,99 Euro“; bahn.de/bahnbusiness (Flexpreis +5,9 % ab 15.12.2024); Nordkurier, Berliner Zeitung, eisenbahn.de (Begründungen, Verzicht 2025). American Express Business Travel Monitor (2001/2002), via Business Travel News und Travelmole.
- Ord, T. (2025): Evidence that Recent AI Gains are Mostly from Inference-Scaling. tobyord.com. Epoch AI (2026): How persistent is the inference cost burden? Redwood Research: AI's capability improvements haven't come from it getting less affordable. Gartner (17.08.2026): Gartner Predicts AI Inference Costs Per Agentic Workflow Will Increase More Than Fivefold Through 2028. Gundlach et al. (2026). Pillitteri, P. (2026), pasqualepillitteri.it.
- Monitoring Analytics via SOFX (PJM-Großhandelspreise), SemiAnalysis-Index via uvik.net (GPU-Mietpreise), TrendForce via globx.eu (Speicherpreise), Wall Street Journal via Luminix (OpenAI), Financial Times via ValueAdd VC (Anthropic).

Jonas Höttler, Balane · hoettler.com · Fragen und Widerspruch willkommen. Wer einen Rechenfehler findet: bitte melden. Genau dafür liegt alles offen.