



REVENUE OPERATIONS & KI

Was KI mit den Umsatzkosten gemacht hat, steht nicht in der Zeile, die alle zitieren.

Zwei Lager streiten über KI und Softwaremargen – beide mit Umfragedaten. Die testierten Abschlüsse geben keinem von beiden recht.

Sinan Isoglu, MBA (Quantic)

Führungskraft für Umsatzwachstum, Dozent und Doktorand

VERÖFFENTLICHT

10. August 2026

AKTUALISIERT

20. August 2026

LESEZEIT

12 Min.

WÖRTER

2.511

<https://isoglu.com/de/insights/journal/was-ki-mit-den-umsatzkosten-gemacht-hat/>

Management Summary	2
Ein deutscher Leser hat diese Zeile oft gar nicht	3
Alle argumentieren mit derselben Art von Beleg – und es sind keine Abschlüsse	3
Was die Abschlüsse sagen	4
Der ehrliche Teil: Der Anstieg ist zum Stillstand gekommen	5
Warum die Kennzahl das nicht klären kann – mit den Zahlen des Kompressionslagers	6
Das Unternehmen, das der These am ähnlichsten sieht, ist kein Beleg für sie	6
Ich wollte ein Messinstrument liefern. Es ist zweimal gescheitert – und das ist der nützliche Teil	7
Wie sollten Finance und RevOps die tatsächlichen KI-Bereitstellungskosten messen?	8
Grenze	9
Quellen	10

MANAGEMENT SUMMARY

Die These, KI mache die Umsatzkosten von Software variabel und die 80-Prozent-Bruttomarge sei vorbei, stützt sich ausschließlich auf Benchmark-Umfragen und Anbieterangaben. Aus testierten Abschlüssen hatte das niemand gerechnet. Für die 64 Werte des BVP Nasdaq Emerging Cloud Index – 53 davon mit lückenloser Quartalsreihe – stieg die Bruttomarge zwischen dem ersten Quartal 2022 und dem ersten Quartal 2026 von 73,9 auf 75,4 Prozent, und 40 von 53 Unternehmen verbesserten sich. Entscheidend ist aber etwas anderes: Diese Kennzahl kann den behaupteten Effekt gar nicht messen. Nach der Kostenzahl des Kompressionslagers selbst müssten 65 Prozent des gesamten Umsatzes aus KI stammen, damit die Marge um 15 Punkte fällt.

Schlagwörter: Umsatzkosten · Künstliche Intelligenz · Maschinelles Lernen · Bruttomarge · KI-Inferenzkosten · Stückkostenrechnung · Revenue Operations

Im nächsten Budgetgespräch wird jemand sagen, KI habe die Bruttomarge von Software zerstört: Inferenz sei ein echter variabler Kostenblock, die 80-Prozent-Ära vorbei, der neue Korridor liege bei sechzig und etwas. Diese Person zitiert eine Zahl, und diese Zahl stammt aus einer Umfrage. In den testierten Abschlüssen ist es nicht passiert: Bei 53 börsennotierten Cloud-Unternehmen stieg die Bruttomarge zwischen dem ersten Quartal 2022 und dem ersten Quartal 2026 von 73,9 auf 75,4 Prozent, und vierzig der dreiundfünfzig verbesserten sich.

Das ist die kleinere Hälfte des Befundes. Die größere: Die Bruttomarge hätte die Kompression gar nicht zeigen können. Beide Lager streiten über eine Kennzahl, die ihre Frage nicht beantworten kann.

Ein deutscher Leser hat diese Zeile oft gar nicht

Bevor es um die Zahlen geht, eine Eigenheit, die im angelsächsischen Diskurs niemand erwähnt: Nach § 275 HGB existiert die Bruttomarge in der deutschen Standard-Gliederung überhaupt nicht. Absatz 2 – das Gesamtkostenverfahren – gliedert nach Kostenarten: Materialaufwand, Personalaufwand, Abschreibungen. Eine Zwischenzeile „Bruttoergebnis“ kommt darin nicht vor. Erst Absatz 3, das Umsatzkostenverfahren, weist die „Herstellungskosten der zur Erzielung der Umsatzerlöse erbrachten Leistungen“ aus und darunter das „Bruttoergebnis vom Umsatz“.

Wer nach HGB im Gesamtkostenverfahren bilanziert, hat die Kennzahl, um die dieser ganze Streit geführt wird, nicht im Abschluss. Sie muss konstruiert werden – und die Konstruktion ist eine Entscheidung darüber, was in die Umsatzkosten wandert. Das ist kein deutscher Sonderfall am Rand. Es ist der Hinweis darauf, dass die Kennzahl eine Darstellungskonvention ist und keine Naturgröße.

Alle argumentieren mit derselben Art von Beleg – und es sind keine Abschlüsse

Die Annahme, um die es hier geht, wurde präzise aufgeschrieben – zwanzig Jahre bevor jemand darüber stritt. Sundararajans Modell der Preissetzung für Informationsgüter formuliert sie in §2.1: „Variable costs of production to the firm – creating copies or providing access to the product – are zero.“ Das ist die Voraussetzung eines Theorems, kein Slogan – und die kursierende Behauptung lautet, die Branche erfülle sie nicht mehr. Ob das stimmt, ist eine Messfrage, und fast ausschließlich Leute, die etwas zu verkaufen haben, haben sie beantwortet.

Es lohnt außerdem, genau zu sagen, was die Annahme gebrochen haben soll. Maschinelles Lernen sitzt seit einem Jahrzehnt in den Umsatzkosten dieser Unternehmen – Betrugserkennung, Abwanderungsmodelle, Suchranking, Empfehlungen – zu Grenzkosten nahe null, denn ein trainiertes Modell ist billig auszuführen. Geändert hat sich die generative Inferenz, bei der jeder Aufruf beim Anbieter auf die Uhr geht. „Künstliche Intelligenz in den Umsatzkosten“ ist also nicht ein Phänomen mit einer Kostenkurve – und eine Sammelzeile, die die billige Sorte seit Mitte der 2010er klaglos mitträgt, wird nun auf das Eintreffen der teuren gelesen.

Der Kompressionsfall ist überall und er ist konkret. SaaSMag beziffert ihn auf „roughly \$230,000“ Inferenzkosten je 1 Mio. USD KI-Umsatz und einen Pfad in die „60s and low 70s“. Die Zahl stammt von ICONIQ Growth. Der 80-Dollar-Arbeitsplatz, der durch 15 Dollar Inferenz fünfzehn Punkte verliert, stammt von Ben Murray. Die 65-Prozent-Obergrenze für LLM-native Unternehmen stammt von Bessemer. Verfolgt man irgendeine Zahl in dieser Argumentation zurück, endet sie bei einem Wagniskapital-Benchmark, einer Anbieterangabe oder einem weiteren Praktiker, der eine davon zitiert.

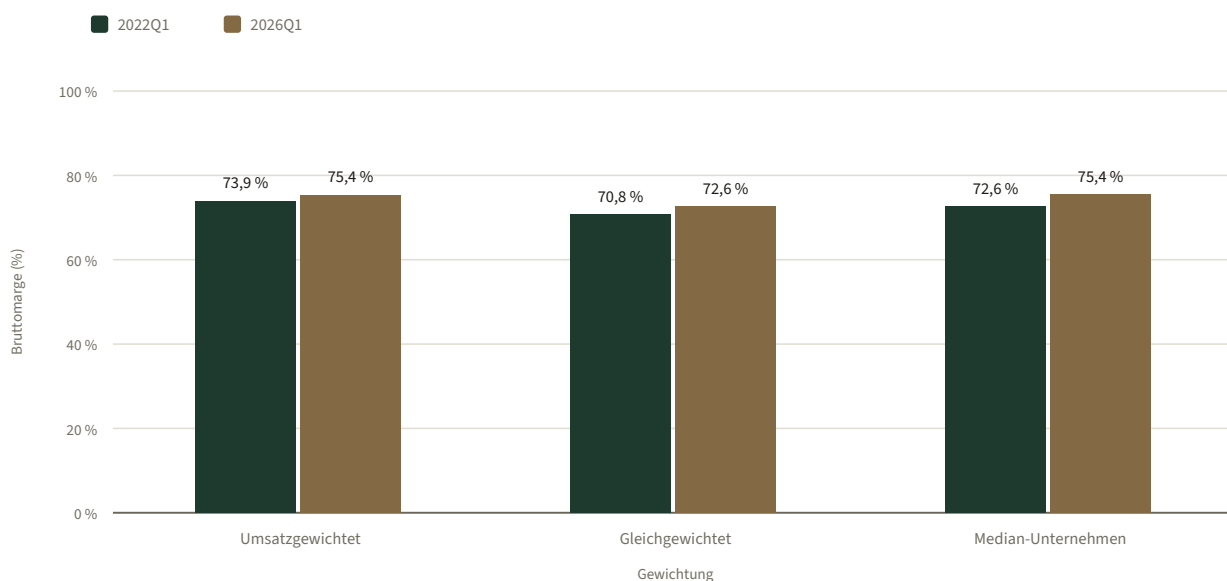
Das andere Lager macht dasselbe. Die Aleph-x-Benchmark-Benchmarks 2026 berichten, die Bruttomarge sei „remarkably stable at 79–81% across all four years“ geblieben, KI habe sie nicht nennenswert gedrückt. Das sind 342 Unternehmen – selbst berichtet, private und börsennotierte gemischt, ohne Testat. Es ist der beste Beleg dieser Seite, und er ist eine Umfrage.

Meritech Analytics verkauft historische Quartals-Bruttomargen für – in ihren Worten – „every public software company, ever“, als Datenwerkzeug und nicht als Argument zu dieser Frage. Ich habe die testierte Reihe nirgends neben der Behauptung veröffentlicht gefunden – und das ist eine Aussage über eine Suche, kein Beweis der Abwesenheit: Meritech mag sie für Abonnenten längst gezeichnet haben, und ein Newsletter, den ich nicht gelesen habe, mag sie gebracht haben. Sagen lässt sich: Kein Argument in dieser Debatte zitiert eine. Die Daten sind kostenlos, quartalsweise und in einem Nachmittag gerechnet.

Was die Abschlüsse sagen

Ich habe Umsatz, Umsatzkosten und Bruttoergebnis je Quartal für die 64 Werte des BVP Nasdaq Emerging Cloud Index aus den XBRL-Daten der SEC gerechnet. Es ist dieselbe Grundgesamtheit, die dieses Vorhaben einen Tag zuvor für den Preisseiten-Korpus festgelegt und veröffentlicht hatte. Ich habe sie gewählt, weil sie ein Index von Cloud-Unternehmen und keine Liste von Firmen mit KI-Funktionen ist. 53 haben eine lückenlose Reihe über siebenzehn Quartale von 2022Q1 bis 2026Q1: 901 Unternehmensquartale. Panel, Code und Ausschlüsse sind zusammen mit diesem Text veröffentlicht.

Abbildung 1 Das Panel im Überblick: drei Gewichtungen, eine Richtung



SEC EDGAR XBRL Company Facts; ausgeglichenes Panel von 53 Werten des BVP Nasdaq Emerging Cloud Index, 2022Q1 bis 2026Q1, 901 Unternehmensquartale. Panel, Code und Ausschlüsse: isoglu.com/de/forschung/bruttomargen-panel/.

Vergleicht man je Unternehmen den Durchschnitt 2022 mit den letzten vier Quartalen, gewinnt das Median-Unternehmen 1,93 Punkte, und vierzig von dreiundfünfzig verbessern sich (Wilcoxon-Vorzeichen-Rang-Test, $p = 0,0002$). Ohne C3.ai, dessen Einbruch die größte Einzelbewegung im Panel ist, steigt der Mittelwert von 1,63 auf 2,45 Punkte. Mit Zwei- oder Sechs-Quartals-Blöcken statt vier: +2,18 und +1,49. Die Richtung bewegt sich nicht.

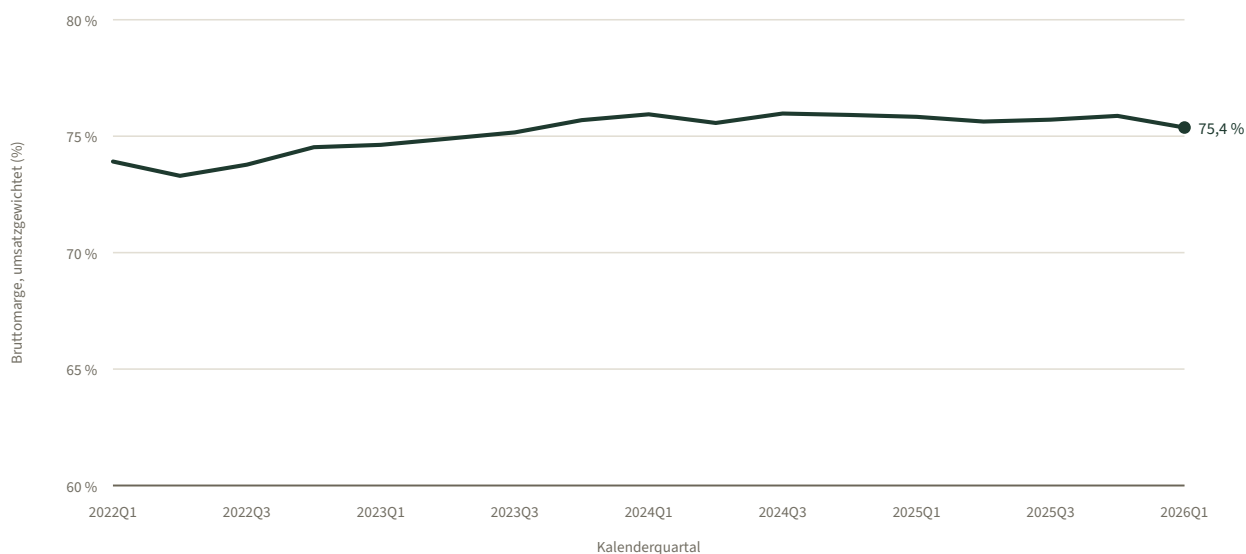
Eine kursierende Behauptung ist konkret genug für einen direkten Test. Mehrere börsennotierte Vertical-SaaS-Unternehmen sollen im vierten Quartal 2025 6 bis 9 Punkte Margenverlust im Jahresvergleich ausgewiesen haben, ausdrücklich KI-Funktionen zugeschrieben. In diesem Quartal, gegen das Vorjahresquartal, über alle neunundfünfzig Werte, die beide Quartale eingereicht haben – nicht nur über das ausgeglichene Panel, denn die Behauptung nennt ein Segment, und einige dieser Namen sind junge Börsengänge: kein einziges Unternehmen fiel um 6 bis 9 Punkte. Keines. Der stärkste Rückgang außerhalb von C3.ai war Sprinklr mit 5,3 Punkten. Das Median-Unternehmen legte um 0,31 Punkte zu, und derselbe Test für Q1 2026 und für das Gesamtjahr 2025 ergibt dasselbe: null.

Zu prüfen sind die in dieser Behauptung genannten Vertical-SaaS-Werte, also hier einzeln: Toast +1,0, ServiceTitan +3,9, nCino +1,7, Q2 +2,7, Intapp +1,8, AppFolio +1,5, Weave +0,6, Veeva –0,4, Agilysys –0,6, Procore –1,1, Alkami –2,2. Der stärkste Rückgang unter ihnen beträgt 2,2 Punkte. Die Behauptung braucht sechs.

Der ehrliche Teil: Der Anstieg ist zum Stillstand gekommen

Der Zuwachs ist eine Geschichte der Jahre 2022 bis 2024, und sie ist ausgelaufen. Vom Durchschnitt 2024 zu den letzten vier Quartalen liegt das Median-Unternehmen 0,28 Punkte im Plus, der Mittelwert 0,43 Punkte im Minus, der Test ergibt $p = 0,62$. Die umsatzgewichtete Linie liegt seit acht Quartalen zwischen 75,4 und 76,0 Prozent.

Abbildung 2 Die Zeile, die alle zitieren, über siebzehn Quartale gezeichnet



Aus dem veröffentlichten Panel mit dem beiliegenden Code gerechnet: umsatzgewichtete Bruttomarge je Quartal, ausgeglichenes 53er-Panel, SEC EDGAR XBRL Company Facts, 2022Q1 bis 2026Q1. Die Achse ist aus dem in der Lesenotiz genannten Grund auf 60–80 beschnitten; in voller Skala verschwindet die Bewegung.

Der zutreffende Satz lautet also nicht „die Margen steigen“. Er lautet: Die Margen stiegen während der Effizienzkorrektur und sind flach, seit KI-Funktionen in Serie ausgeliefert werden. Das ist vereinbar mit einer realen, exakt kompensierten Kompression. Es ist ebenso vereinbar damit, dass nichts passiert. Diese Daten können das nicht trennen – und die Daten der anderen auch nicht.

Warum die Kennzahl das nicht klären kann – mit den Zahlen des Kompressionslagers

Nimmt man ICONIQ ernst, verschlingt Inferenz 23 Prozent des KI-Umsatzes. Die Bruttomarge fällt dann um $0,23 \times$ den KI-Anteil am Gesamtumsatz. Sind 5 Prozent des Umsatzes KI, sind das 1,15 Punkte. Bei 10 Prozent: 2,3 Punkte.

Für die fünfzehn Punkte von 80 auf 65 Prozent müssten 65 Prozent des gesamten Umsatzes aus KI stammen.

Das ist kein feiner Einwand, sondern die Arithmetik der Behauptung selbst. Und die Effektstärken, die plausibel sind, liegen unter der Nachweisgrenze dieses Designs: simuliert mit der beobachteten Streuung zwischen Unternehmen hat das Panel 84 Prozent Teststärke gegen 3 Punkte mittlere Kompression, 50 Prozent gegen 2 Punkte und 17 Prozent gegen 1 Punkt. Eine echte Kompression von einem Punkt sähe genau so aus wie mein Befund.

Beide Lager zitieren einen Mischwert aus einer großen alten und einer kleinen neuen Basis und werten dessen Stabilität oder Bewegung als Beleg. Er ist weder das eine noch das andere.

Der stärkste Einwand gegen das eben Gesagte lautet, dass es hier zwei Mechanismen gibt und ich nur einen bepreist habe. Wird KI als eigener Umsatz verkauft, mit Inferenz dagegengerechnet, gilt die obige Arithmetik und die Mischlinie bewegt sich kaum. Wird KI dagegen in die ohnehin verkauften Arbeitsplätze gebündelt – kein neuer Umsatz, die Inferenzrechnung wandert in die Umsatzkosten –, gibt es überhaupt keinen Umsatzausgleich, und die Marge fällt um den vollen Inferenzanteil. Drei Prozent des Umsatzes für Inferenz sind drei Punkte Marge, ganz ohne 65-Prozent-Schwelle.

Dieser zweite Mechanismus ist mit ziemlicher Sicherheit der häufigere, und er ist derjenige, der meiner Widerlegung schadet. Er ist zugleich derjenige, den diese Daten tatsächlich sehen können, weil er die Mischlinie unmittelbar bewegt – und über vier Jahre ist er dort oberhalb von etwa einem Punkt nicht vorhanden. Der ehrliche Zwischenstand lautet also: Der Mechanismus, den das Kompressionslager beschreibt, kann die Zahl, die es nennt, nicht erzeugen; und der Mechanismus, der es könnte, ist in den Abschlüssen nicht in einer Größenordnung sichtbar, die dieses Design nachweisen kann.

Das Unternehmen, das der These am ähnlichsten sieht, ist kein Beleg für sie

C3.ai fiel von 70,3 auf 29,3 Prozent, der mit Abstand größte Rückgang im Panel – ein KI-Unternehmen. Und doch nicht das, wonach es aussieht. Die Umsatzkosten bewegten sich kaum: 41,2 Mio. USD im Quartal bis April 2025, 40,3 Mio. USD im Quartal bis April 2026. Der Umsatz halbierte sich, von 108,7 auf 51,6 Mio. USD. Das ist ein Nachfrageeinbruch gegen eine fixe Kostenbasis, also operativer Hebeleffekt und keine variablen Inferenzkosten. Der Fall, der der Erzählung am nächsten kommt, hat die falsche Form für sie.

Mechanisch stimmig sieht das Muster nur schmaler und leiser aus. Cloudflare ging von 77,8 auf 71,2 Prozent, während die Umsatzkosten um das 3,9-Fache wuchsen gegen 3,0-faches Umsatzwachstum; DigitalOcean von 62,9 auf 56,1 Prozent. Beide verkaufen Rechenleistung. Das ist ausdrücklich kein geprüfter Befund – ich habe es nach dem Blick in die Daten bemerkt, die Gruppe war nicht vorab festgelegt, und die zehn Unternehmen mit mehr als einem Punkt Rückgang stehen für 14,4 Prozent des Panel-Umsatzes. Es ist die Hypothese, die eine nächste Untersuchung vorab registrieren würde, kein Ergebnis.

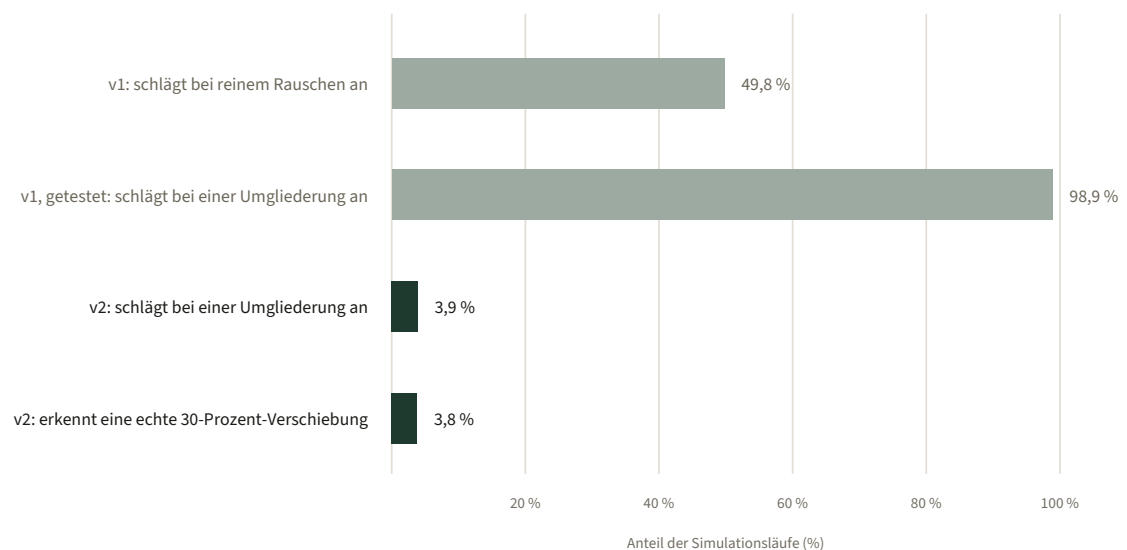
Ich wollte ein Messinstrument liefern. Es ist zweimal gescheitert – und das ist der nützliche Teil

Das Margenniveau beantwortet die Frage „wie profitabel“, nicht die Frage „werden die Kosten variabel“. Die zweite Frage ist eine Elastizität: Regression des Logarithmus der Umsatzkosten auf den Logarithmus des Umsatzes über die Quartale, wobei eine feste Marge exakt 1 ergibt und alles darüber bedeutet, dass die Kosten überproportional wachsen.

Ich habe das gebaut und gegen synthetische Daten ohne jeden Effekt laufen lassen – der Schritt, der entscheidet, ob ein Instrument echt ist. Die naive Lesart, Elastizität über 1, schlägt bei Zufallszahlen in rund 50 Prozent der Fälle an. Ein Münzwurf im Regressionsmantel. Die signifikanzgeprüfte Fassung übersteht das – und schlägt dann bei einer reinen bilanziellen Umgliederung in 98,9 Prozent der Fälle an: einem einmaligen Kostensprung ohne jede Änderung der Kostenstruktur. Jedes Unternehmen, das je ein Support-Team in die Umsatzkosten verschoben hat, würde sie auslösen.

Also habe ich sie in ersten Differenzen neu gebaut, wo ein einmaliger Sprung eine Beobachtung verunreinigt statt die ganze Steigung zu kippen. Sie besteht den Umgliederungstest mit 3,9 Prozent – und erkennt dann eine echte Verschiebung hin zu 30 Prozent variablen Kosten in nur 3,8 Prozent der Fälle, unterhalb ihrer eigenen Falsch-positiv-Rate. Robust und blind.

Abbildung 3 Vier Tests, zwei Instrumente, kein Überlebender



diagnostic-v1-failed.py, 20.000 Wiederholungen: schlägt bei 49,8, 50,7 und 49,6 Prozent der Läufe über drei Rauschstufen an, gezeichnet ist die erste. diagnostic-v2-failed.py, 6.000 Wiederholungen: eine einmalige 8-Prozent-Umgliederung der Kosten und eine anwachsende 30-Prozent-Verschiebung der variablen Kosten mit Bootstrap-Intervall. Beide als Fehlschläge veröffentlicht: isoglu.com/de/forschung/bruttomargen-panel/.

Beide Fassungen liegen im Repositorium, als Fehlschläge benannt, weil sonst jemand die erste veröffentlicht. Ein Test, der eine Änderung der Kostenstruktur nicht von einer Umgliederung unterscheiden kann, ist kein konservativer Test. Er ist ein Erzeuger selbstsicherer Fehltritte, und weder Vorbehalte noch eine größere Stichprobe reparieren das.

Wie sollten Finance und RevOps die tatsächlichen KI-Bereitstellungskosten messen?

Die Frage hat eine Antwort. Sie steht nur in keiner Gewinn- und Verlustrechnung, weil die Umsatzkosten eine einzige Sammelzeile sind und kein Abschluss die Inferenz gesondert ausweist – weshalb jede veröffentlichte Behauptung dazu, in beide Richtungen, eine als Messung verkleidete Schlussfolgerung ist. Dieses Vorhaben ist von der anderen Seite schon einmal gegen dieselbe Wand gelaufen: Was KI wirklich verändert räumt am Ende ein, dass drei seiner vier vorgeschlagenen Auswertungen Daten brauchen, die der Leser nicht hat. Das Muster verdient einen Namen: Die meisten veröffentlichten KI-Messungen sind durch das begrenzt, was zufällig erfasst wird – nicht durch das, worauf es ankommt.

Sie steht in Ihrer Cloud-Rechnung. Drei Größen, die Sie alle bereits haben:

Ihr Inferenzaufwand als Anteil am Umsatz der Produkte, die ihn nutzen – nicht am Gesamtumsatz. Ihr Anbieter rechnet nach Dienst ab; falls der Modell-API-Aufwand heute keinem Produkt zugeordnet ist, ist das eine Frage von wenigen Stunden Kennzeichnung, kein neues System. Nur diese Fassung der 23-Prozent-Zahl bedeutet etwas, weil sie Ihre ist.

Kosten je aktivem Konto auf der KI-Oberfläche, monatlich, über sechs Monate. Kompressions- und Stabilitätsthe-se sagen entgegengesetzte Steigungen voraus, und für Ihre Preissetzung gilt nur Ihre. Die Token-Preise sind gefallen, die Nutzung je Konto ist gestiegen; welcher Effekt gewonnen hat, ist eine lokale Tatsache.

Welche Verträge das bepreisen, was skaliert. Wenn die Inferenzkosten mit der Nutzung steigen und der Vertrag eine Pauschale je Arbeitsplatz ist, haben Sie eine Option gegen Ihre eigene Kostenbasis geschrieben. Das ist eine bereits getroffene, heute gültige Preisentscheidung – ablesbar an Ihren eigenen Auftragsformularen, noch heute Nachmittag.

Abbildung 4 Die drei Größen aus Ihrer eigenen Cloud-Rechnung

ANTEIL INFERENZAUFWAND	KOSTEN JE AKTIVEM KI-KONTO	VERTRÄGE, DIE SKALIERUNG BEPREISEN
Am Umsatz der Produkte, die ihn nutzen, nicht am Gesamtumsatz.	Auf der KI-Oberfläche, in diesem Monat.	Bepreist der Vertrag das, was skaliert: ja oder nein.

Sechs monatliche Durchgänge. Die Antwort ist der Trend; ein einzelner Monat ist Rauschen.

Eigenes Arbeitsblatt.

Und wenn jemand die 60-bis-70-Prozent-Zahl ins Gespräch bringt, ist die nützliche Frage nicht, ob sie stimmt. Sie lautet, aus wessen testiertem Abschluss sie stammt. Nach allem, was bislang vorliegt: aus keinem.

Grenze

Grenze. Die Berichtslinie widerlegt eine einfache Margengeschichte, identifiziert aber keine kausalen KI-Kosten. Trianguliere Nutzungs-, Infrastruktur- und Produktdaten.

Evidenzbasis. Die Kostenanalyse des Beitrags stützt sich außerdem auf diese zusätzliche Quelle: Sundararajan 2004. Der Link identifiziert das konkrete Werk; es stützt den hier diskutierten Mechanismus und seine Geltungsgrenzen, nicht jede einzelne Aussage isoliert.

Sundararajan, A. (2004). Nonlinear pricing of information goods. *Management Science*, 50(12), 1660–1673.
<https://doi.org/10.1287/mnsc.1040.0291>

Handelsgesetzbuch (HGB), § 275 Abs. 2 und Abs. 3 – Gliederung der Gewinn- und Verlustrechnung. Bundesministerium der Justiz.
https://www.gesetze-im-internet.de/hgb/__275.html

ZITIERWEISE

Isoglu, S. (2026, 10. August). Was KI mit den Umsatzkosten gemacht hat, steht nicht in der Zeile, die alle zitieren. Sinan Isoglu.

<https://isoglu.com/de/insights/journal/was-ki-mit-den-umsatzkosten-gemacht-hat/>

WEITERLESEN

Revenue Operations & KI

Die Funktion ohne deutschen Namen

<https://isoglu.com/de/insights/journal/die-funktion-ohne-deutschen-namen/>

Revenue Operations & KI

Die Zahl, die Sie melden

<https://isoglu.com/de/insights/journal/die-zahl-die-sie-melden/>

Revenue Operations & KI

Was KI in Revenue Operations wirklich verändert

<https://isoglu.com/de/insights/journal/was-ki-wirklich-veraendert/>



Sinan Isoglu, MBA (Quantic)

Führungskraft für Umsatzwachstum, Dozent und Doktorand