REVENUE OPERATIONS & AI

What AI did to cost of goods sold is not visible in the line everyone quotes.

Two camps argue about AI and software margins from survey data. The audited filings agree with neither, and could not have settled it anyway.

Sinan Isoglu, MBA (Quantic)

Commercial growth leader, lecturer and doctoral researcher

PUBLISHED	UPDATED	READING TIME	WORDS
10 August 2026	20 August 2026	11 min	2,523

https://isoglu.com/insights/journal/what-ai-did-to-cost-of-goods-sold/

CONTENTS

MANAGEMENT SUMMARY

The argument that AI is turning software cost of goods sold variable: that the 80% gross margin is gone and the new range is 60–70%: rests entirely on benchmark surveys and vendor disclosures. Nobody had computed it from audited filings. On the 64 constituents of the BVP Nasdaq Emerging Cloud Index, 53 of which have a complete run of quarters, blended gross margin rose from 73.9% to 75.4% between 2022Q1 and 2026Q1, and 40 of 53 firms improved. The specific claim that public vertical-SaaS companies shed 6–9 points in Q4 2025 is false: none did. But the rise stopped in 2024, and the deeper problem is that this line has no power to detect the effect it is being quoted for: on the compression camp's own cost figure, AI would have to be 65% of all revenue to move it 15 points.

Keywords: Cost of goods sold · Artificial intelligence · Machine learning · AI inference cost · SaaS gross margin · Unit economics · Revenue operations

Somebody in your next budget meeting will say that AI has broken software's gross margin: that inference is a real marginal cost, that the 80% era is over, and that the new range is sixty-something. They will be quoting a number, and that number will have come from a survey. On the audited filings it did not happen: across 53 listed cloud companies, blended gross margin went from 73.9% to 75.4% between the first quarter of 2022 and the first quarter of 2026, and forty of the fifty-three improved.

That is the smaller half of the finding. The larger half is that the gross-margin line could not have shown the compression even if it were happening, which means both camps in this argument have been reading a number that cannot answer their question.

Everyone is arguing from the same kind of evidence, and it is not filings

The assumption under all of this was written down precisely, twenty years before anyone was arguing about it. Sundararajan's model of information-goods pricing states it in §2.1: "Variable costs of production to the firm: creating copies or providing access to the product: are zero." That is the antecedent of a theorem, not a slogan, and the claim now in circulation is that the industry stopped satisfying it. Whether it did is a measurement question, and it has been answered almost entirely by people with something to sell.

It also pays to be precise about which thing is supposed to have broken it. Machine learning has been sitting inside these companies' cost of revenue for a decade: fraud scoring, churn models, search ranking, recommendations: at a marginal cost near enough to zero, because a trained model is cheap to run. What changed is generative inference, where every call meters at a provider. So "artificial intelligence in cost of goods sold" is not one phenomenon with one cost curve, and a blended line that has quietly carried the cheap kind since the mid-2010s is now being read for the arrival of the expensive kind.

The compression case is everywhere and it is specific. SaaSMag puts it at "roughly 230 , 000" o f in f er e n cecos t f or e v er y1M of AI revenue, and a glide path to "the 60s and low 70s". The figure is ICONIQ Growth's. The 80 se a tt ha tl oses f i f t ee n p o in t s t o a15 inference bill is Ben Murray's. The 65% ceiling for LLM-native companies is Bessemer's. Trace any number in that argument and it terminates in a venture capital benchmark, a vendor snapshot, or another practitioner quoting one of those two.

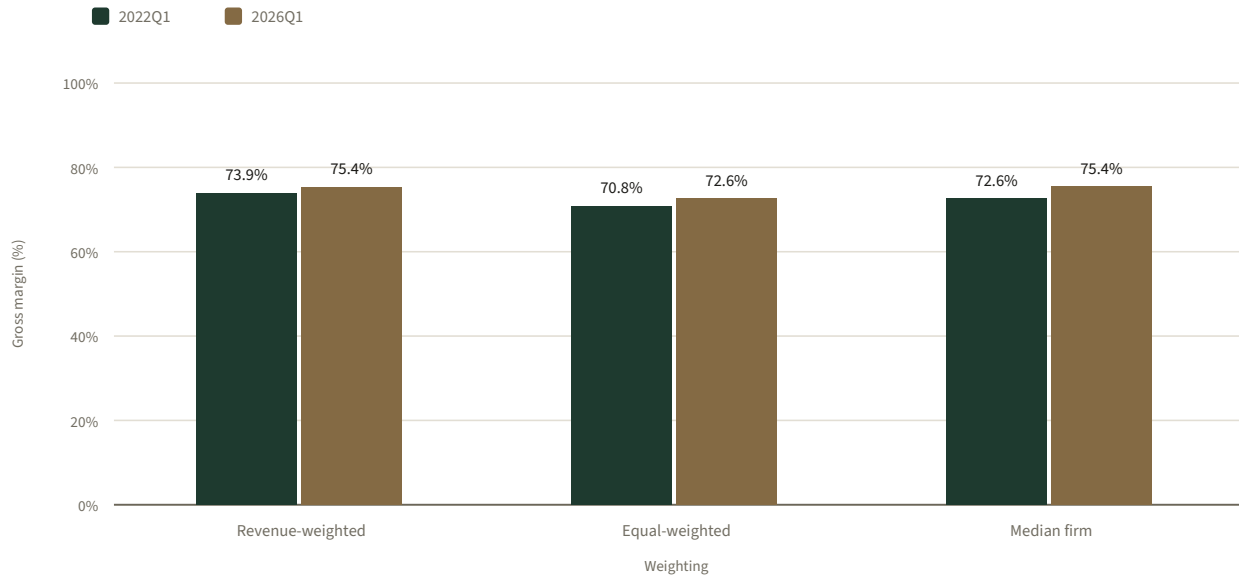
The other camp does the same thing. The 2026 Aleph × Benchmarkit benchmarks report software gross margin "held remarkably stable at 79–81% across all four years" and conclude AI has not materially compressed margins. That is 342 companies: self-reported, private and public mixed, no audit attestation. It is the best evidence on that side and it is a survey.

Jamin Ball's Clouded Judgement publishes a median gross margin of 76% for public software, from Bloomberg, Pitchbook and filings, as a single snapshot. Meritech Analytics sells historical quarterly gross margins for: its words: "every public software company, ever", as a data tool rather than as an argument about this question. I could not find the audited series published next to the claim anywhere, and that is a statement about a search, not a proof of absence: Meritech may well have charted it for subscribers, and a newsletter I did not read may have run it. What I can say is that no argument in this debate cites one. The data is free, it is quarterly, and it takes an afternoon.

What the filings say

I computed quarterly revenue, cost of revenue and gross profit for the 64 constituents of the BVP Nasdaq Emerging Cloud Index from SEC XBRL company facts: the same frame this site fixed and published for the pricing-page corpus a day earlier, chosen because it is an index of cloud companies rather than a list of firms with AI features. Fifty-three have a complete run of seventeen quarters from 2022Q1 to 2026Q1: 901 firm-quarters. The panel, the code and the exclusions are published alongside this piece.

Figure 1 The panel summary: three weightings, one direction



SEC EDGAR XBRL company facts; balanced panel of 53 BVP Nasdaq Emerging Cloud Index constituents, 2022Q1 to 2026Q1, 901 firm-quarters. Panel, code and exclusions: isoglu.com/research/gross-margin-panel/.

Comparing each firm's 2022 average with its latest four quarters, the median firm gained 1.93 points and forty of fifty-three improved (Wilcoxon signed-rank, p = 0.0002). Drop C3.ai, whose collapse is the largest single move in the panel, and the mean gain rises from 1.63 to 2.45 points. Take two-quarter blocks instead of four, or six: +2.18 and +1.49. The direction does not move.

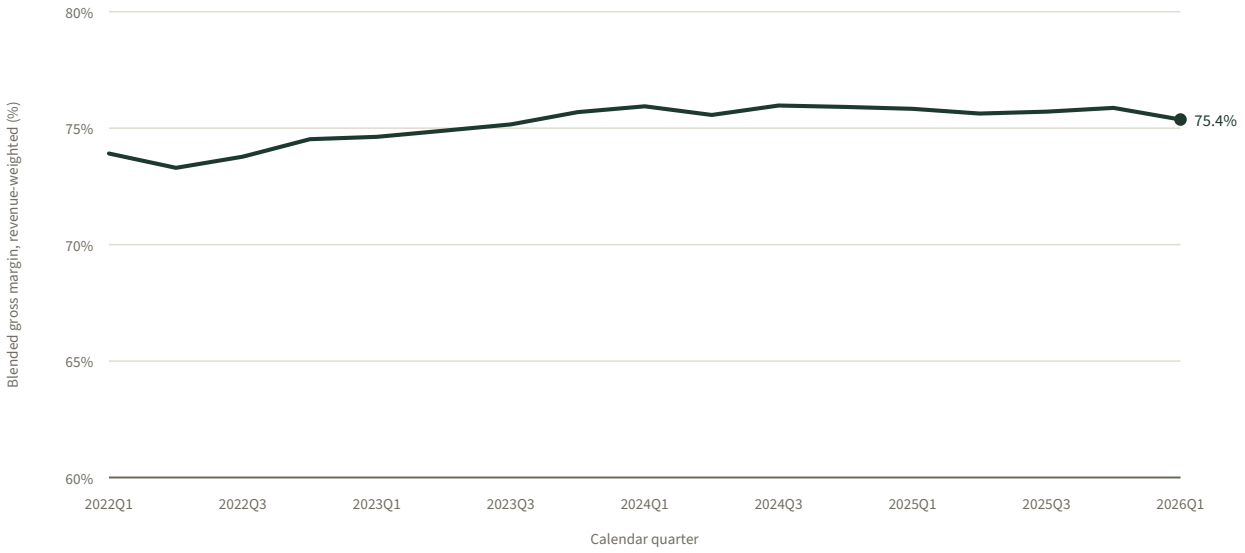
One circulating claim is specific enough to test directly. Multiple public vertical-SaaS companies are said to have disclosed 6–9 points of year-over-year gross margin compression in Q4 2025, attributed to AI feature cost. In that quarter, against the same quarter a year earlier, across all fifty-nine constituents that filed both, not only the balanced panel, because the claim names a segment and some of those names are recent listings: no company fell between 6 and 9 points. None. The steepest decline outside C3.ai was Sprinklr at 5.3. The median firm was up 0.31 points, and the same test on Q1 2026 and on full-year 2025 returns the same answer: zero.

The vertical-SaaS constituents named in that claim are the ones to check, so here they are individually: 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. The steepest decline among them is 2.2 points. The claim needs six.

The honest part: the improvement stopped

The gain is a 2022–2024 story and it has run out. Measured from the 2024 average to the latest four quarters, the median firm is up 0.28 points, the mean is down 0.43, and the test returns p = 0.62. The revenue-weighted line has sat between 75.4% and 76.0% for eight straight quarters.

Figure 2 The line everyone quotes, drawn over seventeen quarters



Computed from the published panel with the code beside it: revenue-weighted quarterly gross margin across the 53-firm balanced panel, SEC EDGAR XBRL company facts, 2022Q1 to 2026Q1. The axis is truncated to 60–80 for a reason the deck states; at full scale the move disappears.

So the truthful sentence is not margins are rising. It is: margins rose through the efficiency correction, and have been flat since AI features started shipping in volume. That is consistent with compression being real and exactly offset. It is also consistent with nothing happening. This data cannot separate those, and neither can anyone else's.

Why the line cannot settle it: on the compression camp's own numbers

Take ICONIQ's figure seriously: inference consumes 23% of AI revenue. Blended gross margin then falls by 0.23 × the AI share of total revenue. If AI is 5% of your revenue, that is 1.15 points. At 10%, 2.3 points.

To lose the fifteen points that takes 80% to 65%, AI would have to be 65% of all revenue.

That is not a subtle objection. It is the argument's own arithmetic, and it says the headline outcome cannot be produced by the mechanism offered for it at any AI revenue share these companies plausibly have. Meanwhile the effect sizes that are plausible sit under this design's detection threshold: simulated at the observed spread between firms, the panel has 84% power against a 3-point mean compression, 50% against 2 points, and 17% against 1 point. A one-point true compression would look exactly like what I found.

Both camps are quoting a blended average of a large old base and a small new one, and calling its stability or its movement evidence. It is neither.

The strongest objection to what I just said is that there are two mechanisms here and I have only priced one. If AI ships as separate revenue with inference against it, the arithmetic above holds and the blended line barely moves. But if AI ships bundled into the seats you already sell: no new revenue, the inference bill added to cost of revenue: then there is no revenue offset at all, and margin falls by the full inference share. Three percent of revenue spent on inference is three points of margin, no 65% threshold required.

That second mechanism is almost certainly the more common one, and it is the one that damages my rebuttal. It is also the one this data can actually see, because it moves the blended line directly, and across four years it is not there above about a point. So the honest scoreboard is: the mechanism the compression camp describes cannot produce the number they quote, and the mechanism that could is not visible in the filings at a magnitude this design can detect.

The company that looks most like the thesis is not an example of it

C3.ai fell from 70.3% to 29.3%, by far the largest decline in the panel, and it is an AI company. It is also not what it appears to be. Its cost of revenue barely moved: 41.2 M in t h e q u a r t er e n d in g A p r i l 2025 ,40.3M in the quarter ending April 2026. Revenue halved, from 108.7 M t o51.6M. That is a demand collapse against a fixed cost base, which is operating deleverage, not variable inference cost. The one case that most resembles the story is the wrong shape for it.

Where the pattern does look mechanically right is narrower and quieter. Cloudflare went from 77.8% to 71.2% as its cost of revenue grew 3.9× against 3.0× revenue growth; DigitalOcean from 62.9% to 56.1%. Both sell compute. I am not offering that as a tested finding: I noticed it after looking, the cohort was not specified in advance, and ten firms down more than a point account for just 14.4% of panel revenue. It is the hypothesis a next study would pre-register, not a result.

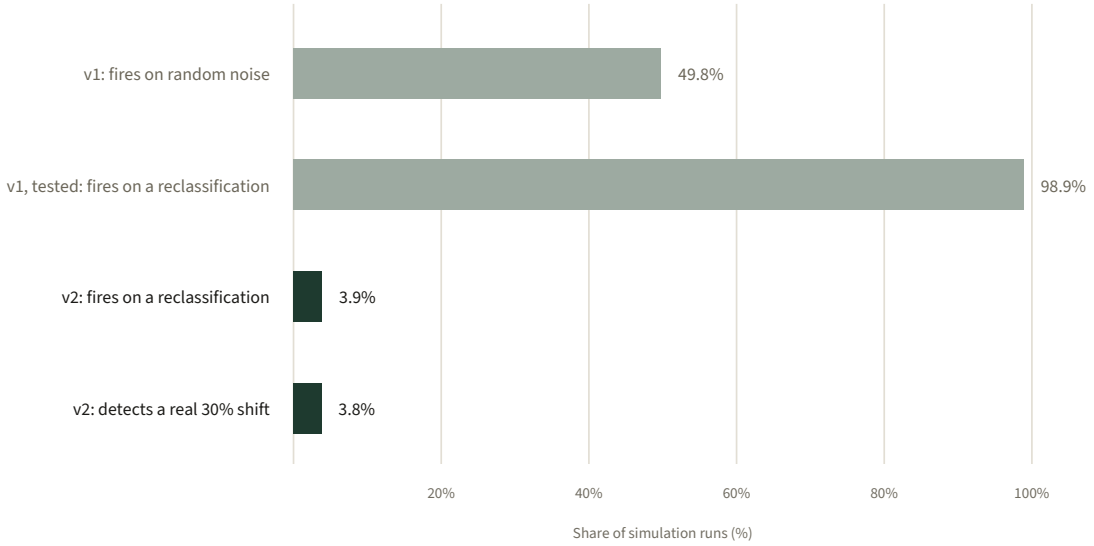
I tried to build you a diagnostic. It failed twice, and that is the useful part

Margin level answers how profitable, not is cost becoming variable. The second question is an elasticity: regress log cost of revenue on log revenue across quarters, where a fixed margin gives exactly 1 and anything above means cost is growing superlinearly.

I built it and ran it against synthetic data with no effect present, which is the step that decides whether an instrument is real. The naive reading: elasticity above 1: fires about 50% of the time on random numbers. It is a coin flip wearing a regression. The significance-tested version survives that, and then fires 98.9% of the time on a pure accounting reclassification: a one-off step in cost with no change in cost structure whatsoever. Every company that has ever moved a support team into cost of revenue would light it up.

So I rebuilt it in first differences, where a one-off step contaminates one observation instead of tilting the whole slope. It passes the reclassification test at 3.9%, and then detects a genuine variable-cost shift reaching 30% of revenue only 3.8% of the time, below its own false-positive rate. Robust and blind.

Figure 3 Four tests, two instruments, no survivor



diagnostic-v1-failed.py, 20,000 replications: fires on 49.8, 50.7 and 49.6% of runs across three noise levels, the first drawn here. diagnostic-v2-failed.py, 6,000 replications: a one-off 8% cost reclassification, and a ramped 30% variable-cost shift with a bootstrap interval. Both published as failures at isoglu.com/research/gross-margin-panel/.

Both versions are in the repository, named as failures, because the alternative is that someone publishes the first one. A test that cannot tell a cost-structure change from a restatement is not a conservative test. It is a generator of confident wrong answers, and neither caveats nor a larger sample fixes it.

How should finance and RevOps track AI cost to serve?

The question has an answer. It is simply not in anybody's income statement, because cost of revenue is one blended line and no filer breaks out inference, which is why every published claim about it, in both directions, is an inference dressed as a measurement. This site has run into the same wall from the other side before: what AI actually changes in revenue operations ends by admitting that three of its four proposed cuts need data the reader does not have. The pattern is worth naming: most published AI measurement is bounded by what happens to be recorded, not by what matters.

It is in your cloud bill. Three things, all of which you already hold:

Your inference spend as a share of the revenue of the products that use it, not of total revenue. Your provider bills by service; if model API spend is not tagged to a product today, that is a tagging job of a few hours, not a new system. This is the only version of the 23% figure that means anything, because it is yours.

Cost per active account on the AI surface, by month, for six months. The compression case and the stability case make opposite predictions about the slope, and yours is the only slope that governs your pricing. Token prices have fallen while usage per account has risen; which one won is a local fact.

Which contracts price the thing that scales. If inference cost rises with usage and the contract is a flat seat, you have written an option against your own cost base. That is a pricing decision already made, currently in force, and readable off your own order forms this afternoon.

Figure 4 The three numbers off your own cloud bill

INFERENCE SPEND SHARE Of the revenue of the products that use it, not of total revenue.	COST PER ACTIVE AI ACCOUNT On the AI surface, this month.	CONTRACTS PRICING WHAT SCALES Does the contract price the thing that scales: yes or no.

Six monthly passes. The trend is the answer; a single month is noise.

Author's own worksheet.

And when someone brings the 60-to-70-percent number to the meeting, the useful question is not whether it is right. It is which company's audited filing it came from. On the evidence so far, the answer is none of them.

Boundary

Boundary. The filing line can reject a simple margin-collapse story, but it cannot identify AI cost causally. Treat the result as a measurement boundary and triangulate with usage, infrastructure and product-level cost data.

Evidence base. The cost-of-goods analysis also draws on this additional source: Sundararajan 2004. The link identifies the exact work; it supports the mechanism and boundary conditions discussed here, not every claim in isolation.

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

END OF ESSAY

HOW TO CITE THIS ESSAY

Isoglu, S. (2026, August 10). What AI did to cost of goods sold is not visible in the line everyone quotes. Sinan Isoglu. https://isoglu.com/insights/journal/what-ai-did-to-cost-of-goods-sold/

KEEP READING

Revenue operations & AI

The function without a German name

https://isoglu.com/insights/journal/the-function-without-a-german-name/

Revenue operations & AI

The number you call

https://isoglu.com/insights/journal/the-number-you-call/

Revenue operations & AI

False positives in churn models have a cost

https://isoglu.com/insights/journal/false-positives-in-churn-models-have-a-cost/

Sinan Isoglu, MBA (Quantic)
Commercial growth leader, lecturer and doctoral researcher