CLOUD CAPITAL

2026

The Cost of Compute: *AI Edition*

What 100 CFOs Reveal About AI Infrastructure's Impact on the P&L

Contents

Foreword

For years, SaaS finance teams have had a fairly well-understood economic model.

We know how to think about hosting costs. We know what healthy gross margins look like. We know how to forecast infrastructure as the business scales. And we've built our pricing models, budgets, and SaaS metrics around those assumptions.

AI is changing those assumptions.

AI is not simply another product feature. For many companies, it introduces an entirely new cost structure underneath the product. Compute, model usage, storage, data infrastructure, and other costs can move much more quickly and much less predictably than the infrastructure costs SaaS companies have historically managed.

That makes the Cloud Capital report particularly timely.

As AI becomes a bigger part of the product and internal systems, finance teams need to understand more than just how much the company is spending. We need visibility into what is driving those costs, which customers and products are creating them, and whether our pricing model supports the underlying economics. That theme—visibility, attribution, and pricing discipline—is central to the findings in this report.

I believe this will become an increasingly important part of the SaaS CFO playbook. Gross margin can no longer be something we simply review after the month closes. Finance needs to be involved earlier in conversations around AI product strategy, pricing, infrastructure, and unit economics.

The fundamentals of SaaS finance are not going away. But AI is adding another layer that requires additional frameworks.

The companies that understand that layer will be much better positioned to scale AI without losing sight of the economics of the business.

Ben Murray
The SaaS CFO

Introduction

I've watched a strange thing happen on the P&L of every startup we work with. A cost that barely registered five years ago has become, for a growing number of companies, one of the largest lines they carry, and one of the hardest to explain. Finance leaders who feel confident about nearly every other part of their business tell me AI infrastructure is the line they trust least. Engineering leaders tell me they cannot fully explain last month's bill. Both are describing the same problem from opposite ends.

That problem is why we commissioned this research, a survey of 100 CFOs and senior finance leaders at growth-stage software companies, run independently by Sapio Research.

The pattern is almost always the same. A team ships a new AI-powered feature, adoption is strong, usage climbs, and then the invoice arrives three or four times higher than Finance modeled, driven by usage no one forecasted and billed by a vendor Finance wasn't watching.

Our survey found that most of it isn't actually tokens, counterintuitive as that sounds. It's the compute, data, and storage that sit underneath, and that downstream spend is far more challenging to predict.

But the most surprising finding wasn't about the cost at all. A few years ago, simply having visibility into cloud and AI spend was an advantage. Today, it's about what companies do with that visibility. Top performers can trace their AI cost down to the individual customer and price to recover it, and the distance between them and everyone else was wider than I expected. It tracks how a company is built more than its size or age.

What ties it together is that AI infrastructure has stopped being a technical line item and become a financial one. It now sits close enough to the center of the P&L that how a company sees it, forecasts it, and prices for it is starting to separate durable AI economics from a cost that simply grows. The pages that follow dig into how the strongest companies do it, and what it would take to follow.

I hope you find it useful.

Edward Barrow
Chief Executive Officer, Cloud Capital

Executive **Summary**

Across every part of this survey, one picture holds: AI infrastructure cost is large, rising fast, hard to forecast, and unevenly managed.

For a meaningful share of companies, infrastructure has become one of the largest lines on the P&L, reaching a third of revenue for the companies most dependent on AI, and it's growing 25 to 50 percent or more a year. It is also hard to see coming. Nearly every company overran an AI cost forecast in the past year, and the overruns concentrate in compute and data storage rather than the model bill most teams watch. Forecasting that cost is now the top concern finance leaders name for the year ahead.

The companies that handle AI cost well share two capabilities: they attribute the cost precisely, down to the individual customer, and they price to recover it. Where those capabilities are present, AI is a healthy line of business. Where they are absent, it is a cost that grows faster than the tools to see it. The number of formal cost controls a company runs does little to separate the two groups; attribution and pricing power do.

The clearest fault line in the data runs between two kinds of company. **AI-native** companies, whose product is the AI, carry cost visibility as a structural feature of how they are built. **AI-enabled** companies, traditional software businesses that have added AI features to an existing product, more often absorb AI as a secondary cost they struggle to see or predict. That distinction explains who is managing AI economics well better than company size or company age does, and it holds even after accounting for scale. The report returns to it in every section, because the economics follow the architecture.

A note on scope. This survey measures the cost side of AI infrastructure (spend, growth, forecasting, attribution, pricing) and the margin outcome. It does not measure the drivers of margin change such as headcount, labor productivity, or overall pricing, so where margins are discussed, the analysis is bounded to what the data can support. Full methodology and respondent profile appear at the back.

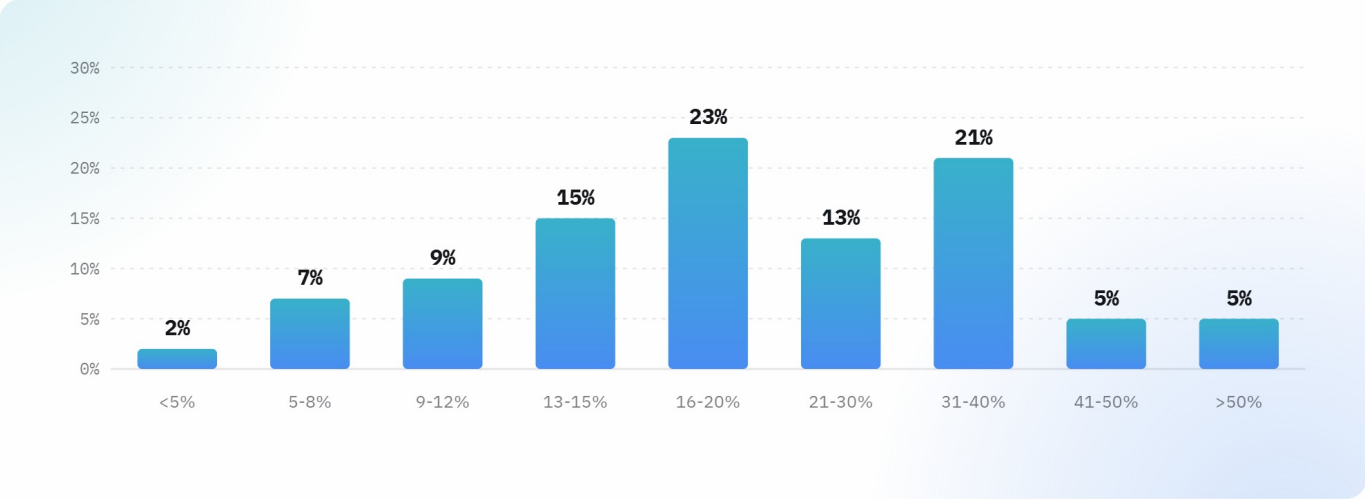
The Cost Is Large and *Growing Fast*

The first thing the data establishes is scale. For a meaningful share of companies, AI infrastructure is one of the largest cost lines they carry, and it is growing faster than almost anything else on the P&L.

Across the full sample, infrastructure spend (cloud plus AI) runs at a median of 16 to 20 percent of revenue. That median understates the story, because the market is split. One cluster of companies runs infrastructure at 13 to 20 percent of revenue; a second, nearly as large, runs it at 31 to 40 percent. A third of all companies spend more than 30 percent of revenue on infrastructure. The sample holds two distinct populations rather than one, and the split between them follows a clear pattern.

How much of revenue goes to infrastructure

Share of all companies at each level of infrastructure spend. The market splits into two groups: one near 16–20% of revenue, another near 31–40%.



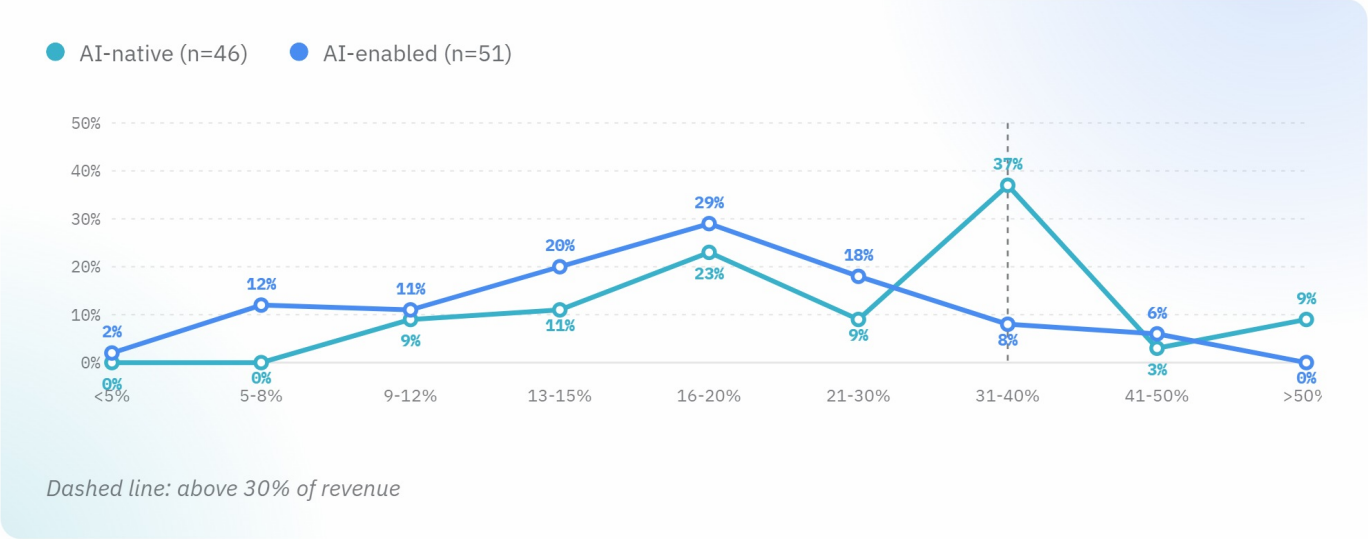
Source: June 2026 CFO survey, n=100, administered by Sapio Research: all respondents; bars show % of companies in each band.

The division runs along business model. For AI-native companies, infrastructure spend reaches roughly a third of revenue, and half of them spend more than 30 percent. AI-enabled companies sit materially lower, with a median in the 16 to 20 percent range and essentially none above 40 percent.

When the AI is the product, the infrastructure to run it is a dominant cost of goods; when AI is a feature added to an existing product, it is a smaller, though still growing, line.

Infrastructure spend as % of revenue, by company type

Share of each group at each spending level. Half of AI-native companies spend over 30% of revenue on infrastructure; almost no AI-enabled company does.

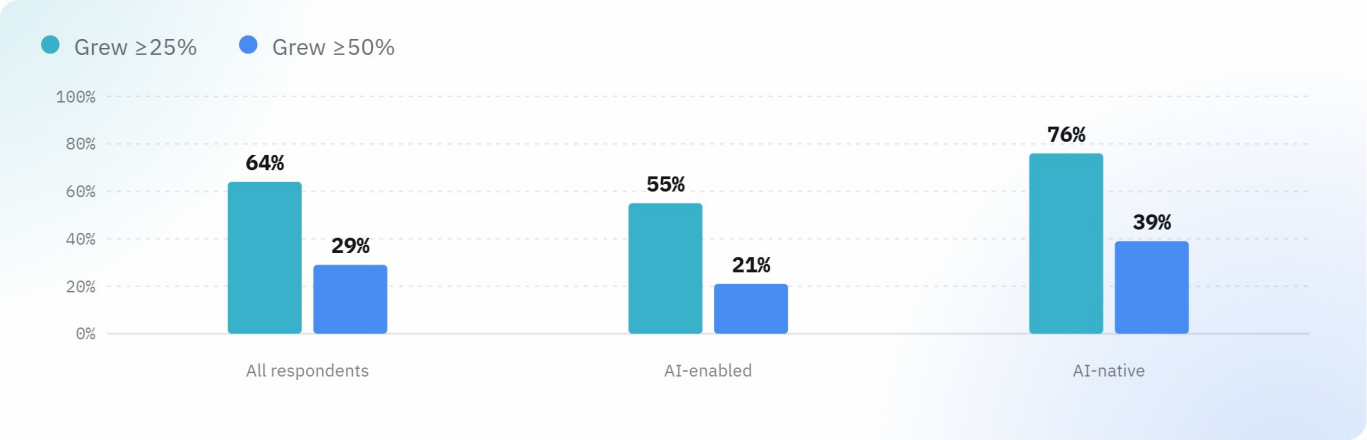


Source: June 2026 CFO survey, n=100, administered by Sapio Research: points show % of each cohort in each band. Bands are ordinal categories: connecting lines aid readability and do not imply values between bands.

The growth is near-universal and steep. Sixty-four percent of all respondents grew AI infrastructure spend by 25 percent or more over the past year, and 29 percent grew it by more than half. Among AI-native companies the curve is steeper still: 76 percent grew spend by 25 percent or more, and 39 percent grew it by more than half. The pattern is strongest among the most AI-intensive companies of all, those for whom AI represents 21 percent or more of cloud spend, which grew fastest in the sample. The growth, in other words, is concentrated precisely where the base is already largest.

AI infrastructure spend growth, past 12 months

Share of each group whose AI spend grew by 25% or more, and by 50% or more.

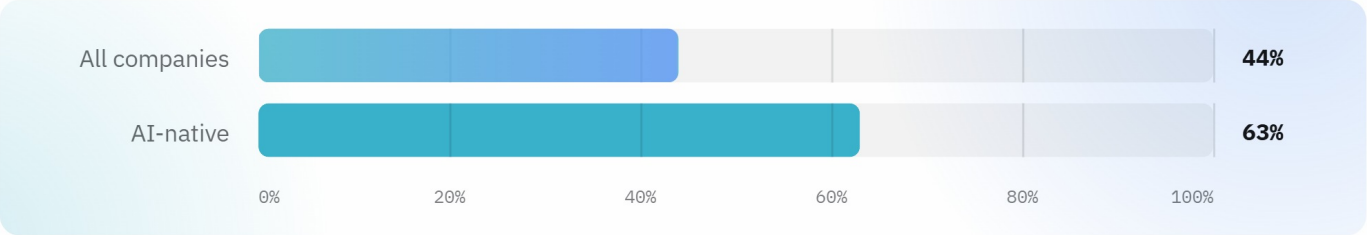


Source: June 2026 CFO survey, n=100, administered by Sapio Research: % of cohort whose AI spend grew by the stated amount.

AI is also a large and rising share of the cloud bill itself. For 44 percent of all companies, AI represents a major or dominant share of total cloud spend (21 percent or more); among AI-native companies that figure is 63 percent. For a large part of the market, AI has become the main driver of the infrastructure bill.

Share of companies where AI is a major part of cloud spend

Percentage of companies where AI is at least 21% of total cloud spend, by company type.



Source: June 2026 CFO survey, n=100, administered by Sapio Research: % of cohort where AI is 21%+ of cloud spend.

The takeaway for a finance leader reading this is straightforward and slightly uncomfortable. This is a cost line that is both large and compounding, and it is largest and fastest-growing precisely for the companies whose economics depend on it most. A cost that behaves this way cannot be managed as an afterthought. The rest of this report is about whether finance teams can see it, forecast it, and price for it, and what separates the ones who can from the ones who cannot.

The Cost Is Hard to Forecast, *and Margin Pressure Is Building*

Beyond its size, AI infrastructure cost is difficult to see coming. It overruns estimates for almost everyone, and it does so in a way that misleads the teams trying to forecast it.

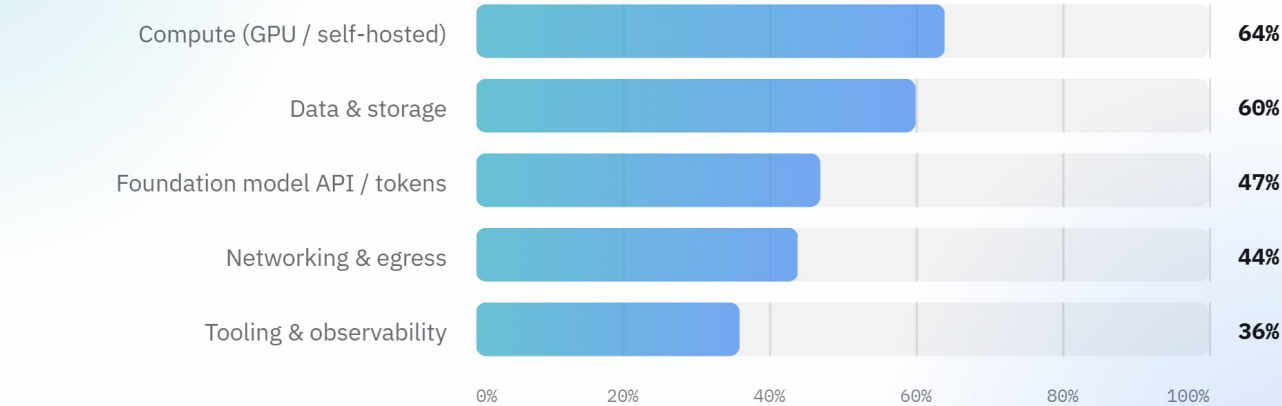
Ninety-six percent of respondents had at least one AI cost category exceed its forecast over the past year. Only 4 percent came through clean. The categories that most often break the forecast are the ones finance teams watch least closely: compute infrastructure (named by 64 percent among their top overruns) and data and storage (60 percent) both outrank foundation-model API and token costs (47 percent). The overruns hide in the infrastructure lines beneath the model bill everyone tracks.

96%

of respondents had an AI cost category exceed forecast in the past year.

Where forecasts break: categories exceeding forecast

Share of companies naming each category among their top forecast overruns.



Source: June 2026 CFO survey, n=100, administered by Sapio Research: "select up to three"; only 4% said nothing exceeded forecast.

The Cost Is Hard to Forecast, and Margin Pressure Is Building

(continued)

This ordering points to a blind spot worth sitting with. The token bill is the visible, explicit cost of AI. It has a published per-token price, it arrives on a dedicated invoice, and it is the number most teams reach for when they estimate what an AI feature will cost. The compute, data, and storage required to build, serve, and support that AI are larger and less legible, spread across the broader infrastructure bill and rarely tagged as "AI." The survey does not isolate the cause, but the pattern is consistent with a simple explanation: teams forecast the cost they are watching and under-scope the cost they are not, so the overruns land where attention is thinnest. The practical consequence is that AI cost is a full-stack infrastructure question. Treating it as a model-pricing question is what leaves forecasts exposed.

> *Finance leaders who feel confident about nearly every other part of their business tell me AI infrastructure is the line they trust least. Engineering leaders tell me they cannot fully explain last month's bill. Both are describing the same problem from opposite ends.*
>
> 
>
> **Ed Barrow**
> Co-Founder & CEO, Cloud Capital

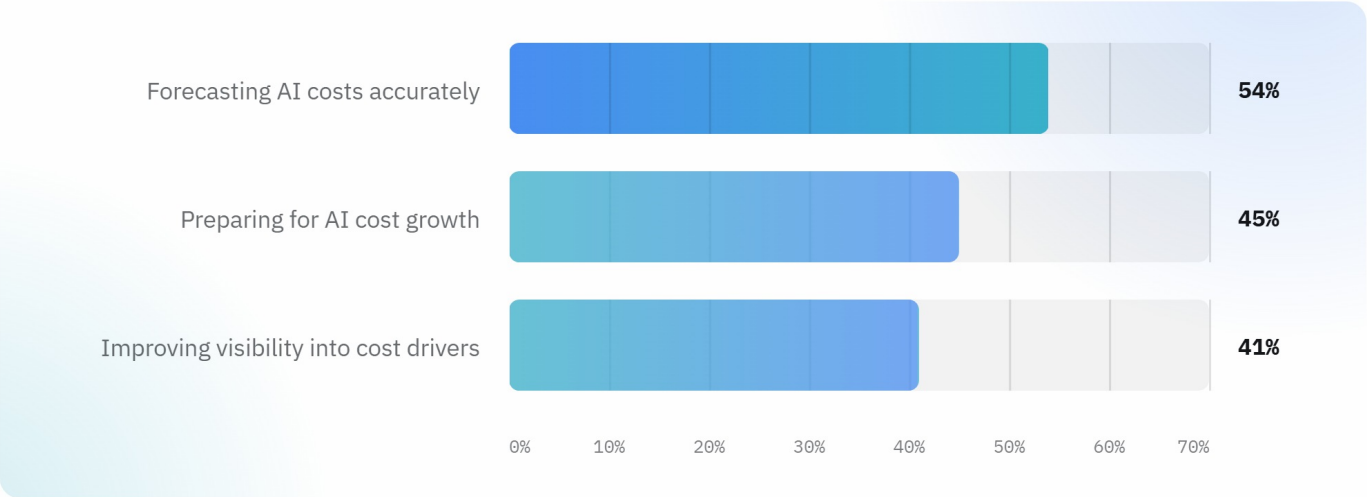
Because the cost behaves this way, forecasting has become the dominant anxiety. Forecasting AI costs accurately is the single most-cited concern for the next twelve months, named by 54 percent of respondents, ahead of preparing for cost growth (45 percent) and improving visibility (41 percent). This is the lived experience of a cost that grew 25 to 50 percent or more in a year and blows past its forecast at the category level for virtually every company.

The Cost Is Hard to Forecast, and Margin Pressure Is Building

(continued)

Top AI cost concerns, next 12 months

Share of finance leaders citing each, among their top three concerns.

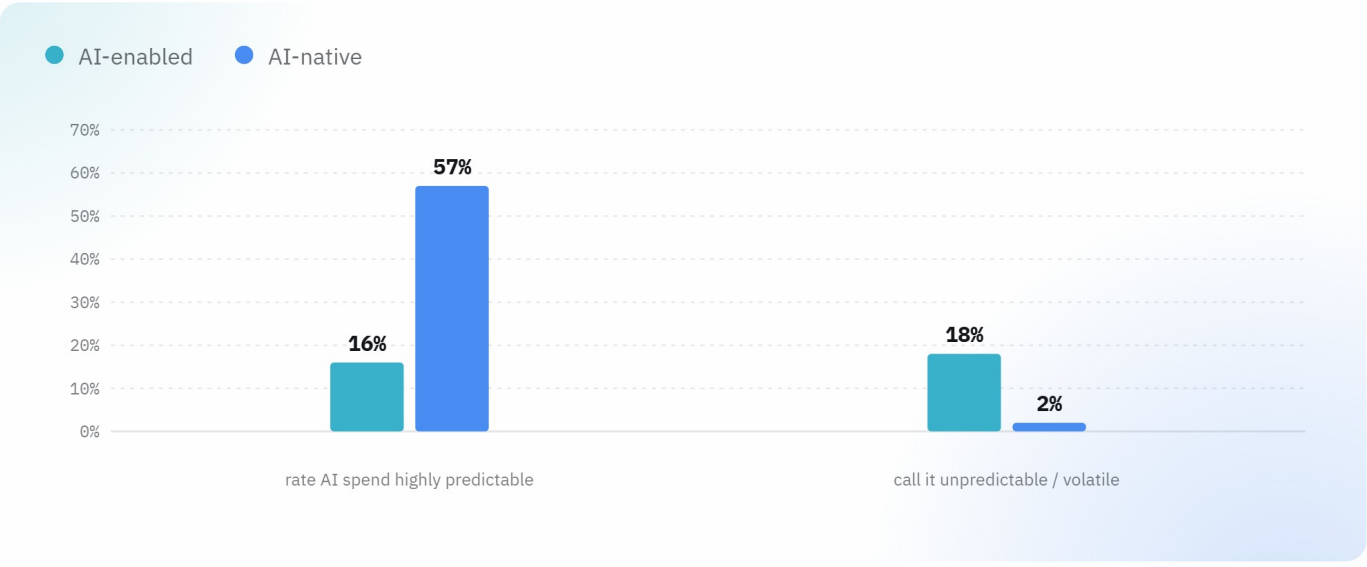


Source: June 2026 CFO survey, n=100, administered by Sapio Research: "select up to three"; top three of seven options shown.

The ability to forecast splits sharply by business model. Among AI-native companies, 57 percent rate their AI spend highly predictable and only 2 percent call it unpredictable. Among AI-enabled companies, just 16 percent rate it highly predictable and 18 percent call it unpredictable or volatile. The companies that treat AI as a core line have built the instrumentation to forecast it; the companies that added it to an existing product carry the forecast risk with the least ability to see it coming.

Forecast confidence splits by business model

Share of each group rating AI spend highly predictable, versus unpredictable or volatile.



Source: June 2026 CFO survey, n=100, administered by Sapio Research: highly predictable vs unpredictable/volatile.

The Cost Is Hard to Forecast, and Margin Pressure Is Building

(continued)

Two figures need to be read together, because apart they mislead. Ninety percent of respondents say their total AI spend lands within 10 percent of forecast, even though 96 percent had a category overrun. The reconciliation is that companies are absorbing constant category-level surprises in order to hold an aggregate number. The component lines are volatile; holding the total steady is active, effortful work, and it should be read as evidence of effort rather than reassurance that the cost is easy to predict.

This is also where the report addresses margins, and it does so carefully, because the data supports a specific and bounded claim. The pressure is real and widely felt: 61 percent of respondents, and 76 percent of AI-native companies, name AI infrastructure as a primary or significant source of downward pressure on gross margin.

That pressure has not yet shown up as falling reported margins. Margins across the sample are healthy and, by respondents' own accounts, mostly rising. The two facts coexist because companies are offsetting the pressure with levers that sit outside the infrastructure line, most likely pricing power, revenue mix, and productivity, which this survey does not measure directly. One offsetting lever is visible in the data: the AI-native companies reporting the most pressure are largely the same companies pricing AI accretively. For them, the margin holds because they can price for a cost they fully acknowledge is large.

The reasonable reading is that margins are holding because companies are pulling hard on levers that have limits, and the companies without pricing power are the exposed ones. The concern those companies report is a leading indicator, which is why forecasting and cost growth top the worry list while defending margin sits lower, at 23 percent. Finance leaders are trying to get ahead of the pressure before it reaches the bottom line.

> *Gross margin can no longer be something we simply review after the month closes. Finance needs to be involved earlier in conversations around AI product strategy, pricing, infrastructure, and unit economics.*
>
> 
>
> **Ben Murray**
> The SaaS CFO

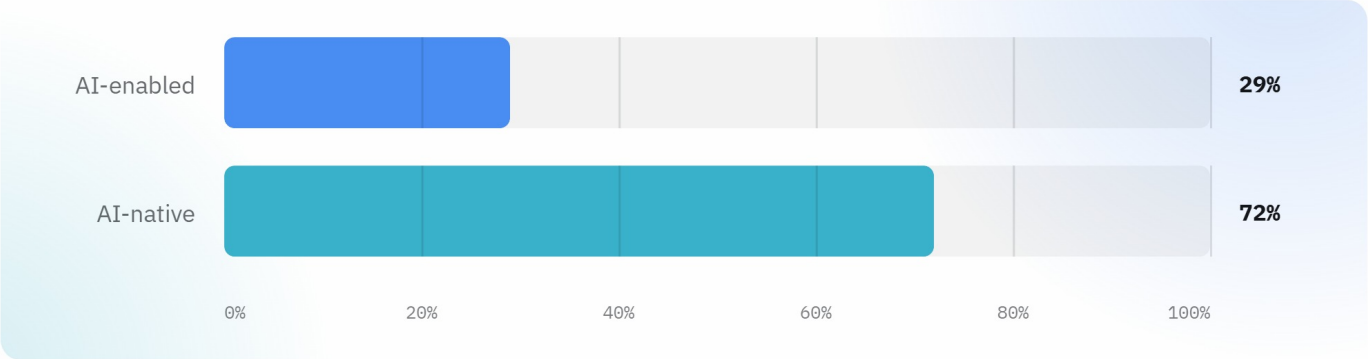
The Differentiator Is Attribution
and Pricing, Not Governance

Formal AI governance controls are widespread and evenly held. The average company runs about 2.8 controls, and that count is essentially identical for AI-native and AI-enabled companies. Yet the two groups reach very different outcomes on cost visibility and pricing recovery despite deploying the same amount of monitoring. Controls appear at the same rate whether or not a company can actually see and price its AI costs, which makes them close to table stakes. The capability that separates the two groups sits underneath the controls: cost attribution and pricing recovery.

On attribution, customer-level cost visibility exists in only about half the market (49 percent), and it splits hard by type. Seventy-two percent of AI-native companies can attribute AI cost down to the individual customer, versus 29 percent of AI-enabled companies. The gap holds regardless of company size, which marks it as a trait of how the business is built rather than a function of finance-team scale.

Which companies can break AI cost down by customer

Share that can trace AI cost to individual customers, by company type.



Source: June 2026 CFO survey, n=100, administered by Sapio Research: % of cohort with excellent customer-level cost attribution.

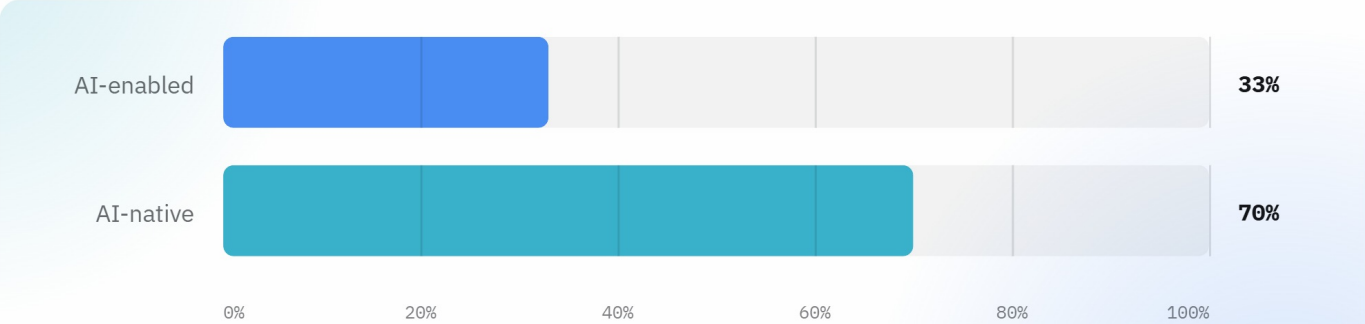
The Differentiator Is Attribution and Pricing, Not Governance

(continued)

On pricing, the gap is the cleanest single result in the survey. Seventy percent of AI-native companies say their AI pricing is accretive to gross margin, versus 33 percent of AI-enabled companies. More than twice as many AI-native companies successfully price to recover AI cost, and the gap widens further among the most AI-intensive companies, reaching 79 percent accretive among those with AI at 21 percent or more of cloud spend. The finding strengthens under a stricter definition rather than weakening.

Who prices AI to make money

Share of companies whose AI pricing adds to gross margin, by company type.



Among the most AI-intensive companies, 79% price AI profitably.

Source: June 2026 CFO survey, n=100, administered by Sapio Research: % of cohort saying AI pricing is accretive to gross margin.

The mechanism of pricing matters less than the power behind it. Bundled, add-on, and usage-based pricing all succeed or fail at roughly equal rates. Companies that bundle AI into their subscription without a separate charge actually show a higher accretive rate than those with explicit add-ons or usage meters, because the bundlers are disproportionately AI-native companies for whom the AI is the product and the subscription price already recovers its cost. Pricing power beats pricing structure.

The composition of governance controls reinforces the point. AI-native companies favor strategic controls that gate what gets built, such as approval workflows and model or vendor policy.

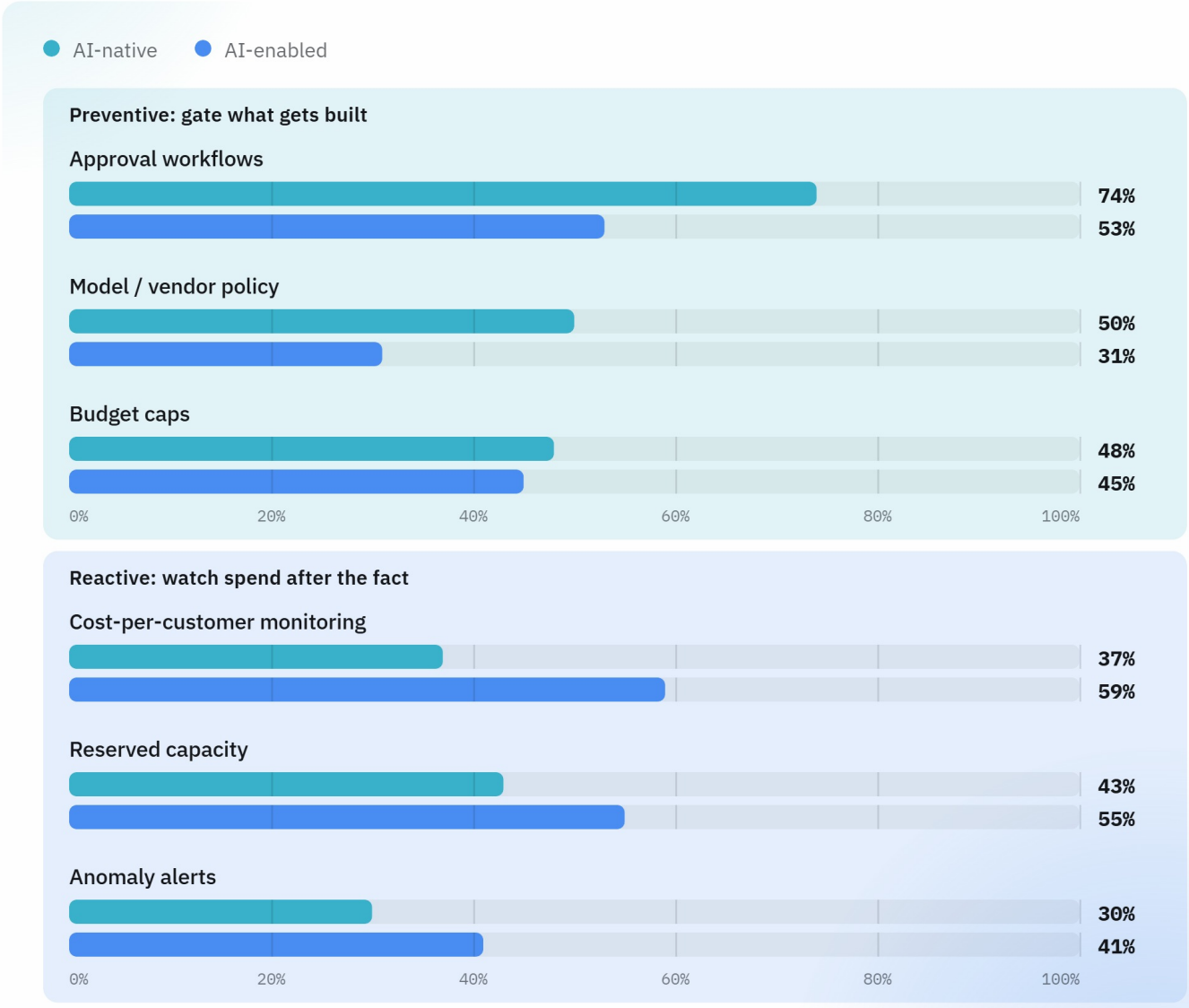
The Differentiator Is Attribution and Pricing, Not Governance

(continued)

AI-enabled companies favor reactive, operational controls that watch spend after the fact, such as cost-per-customer monitoring and anomaly alerts. Both groups run about the same number of controls, but the AI-native mix is built to prevent cost problems while the AI-enabled mix is built to catch them.

AI-native controls prevent, AI-enabled controls react

Share of each group using each control. Both run about 2.8 controls on average.



Source: June 2026 CFO survey, n=100, administered by Sapio Research: % of cohort with each control.

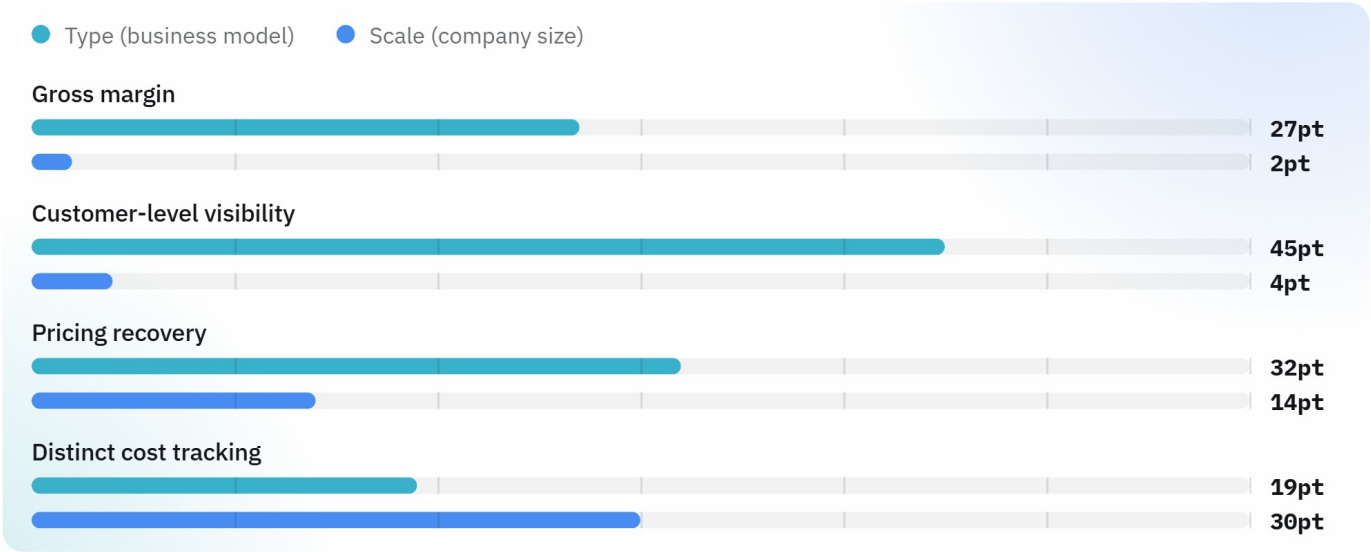
The Economics Follow
the Architecture

The differences across attribution, pricing, and forecasting all track a single underlying variable: whether cost visibility is designed into the product. In AI-native companies, the AI is the product, so its economics are instrumented from the start and cost attribution is inherent. In AI-enabled companies, AI is added to an existing product where cost visibility was not part of the original design. The gap between the two groups is structural, best understood as a property of how each kind of company is built.

This is the report's central claim, so it is worth addressing the obvious objection directly. One might suspect the AI-native advantage is really a size or maturity advantage in disguise, that the AI-native companies simply happen to be larger or further along and would look better on any measure. The data rules this out. When each finding is examined while holding company size constant, the gaps in margin, attribution, and pricing persist within every size band, which identifies them as effects of business model rather than scale. The one capability that does track size is distinct cost tracking, the basic act of separating AI from general cloud in the ledger, which builds with finance-team maturity. Everything else that shapes the economics tracks how the company is architected, not its scale.

The gaps track business model, not company size

How much of each gap comes from business model versus company size (percentage points).

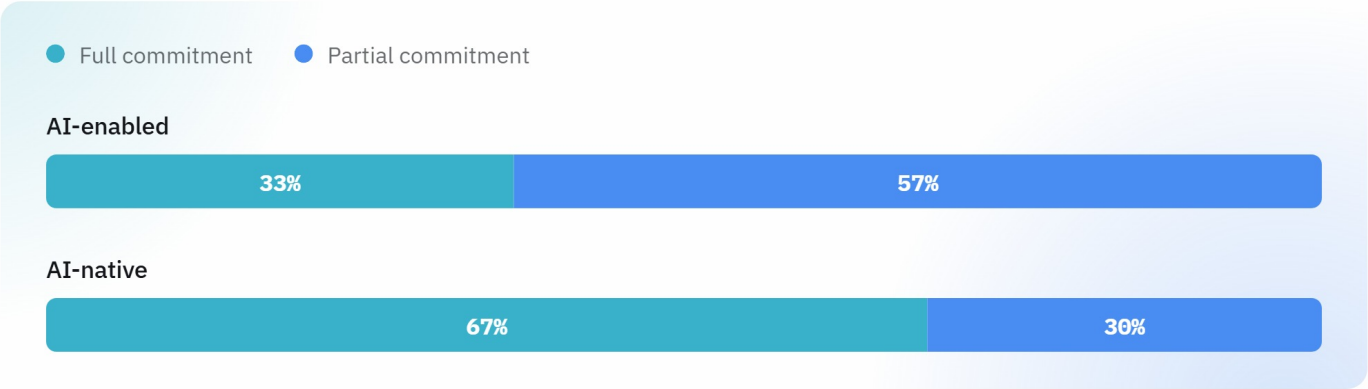


Source: June 2026 CFO survey, n=100, administered by Sapio Research: approximate percentage-point effect of each factor, holding the other constant.

The same split appears in how companies buy. Offered a 30 percent inference discount in exchange for a twelve-month volume commitment, 94 percent of companies took the deal in some form. AI-native companies committed at full volume more often, 67 percent versus 33 percent of AI-enabled companies, which more often took a reduced commitment. The groups that forecast AI spend most confidently were also the groups that committed most fully. The confidence to forecast a cost is what gives a company the confidence to commit to it, and commitment is where the largest savings on that cost are found.

Commitment behavior mirrors forecast confidence

How each group responded to a 30% discount offer for a 12-month commitment.



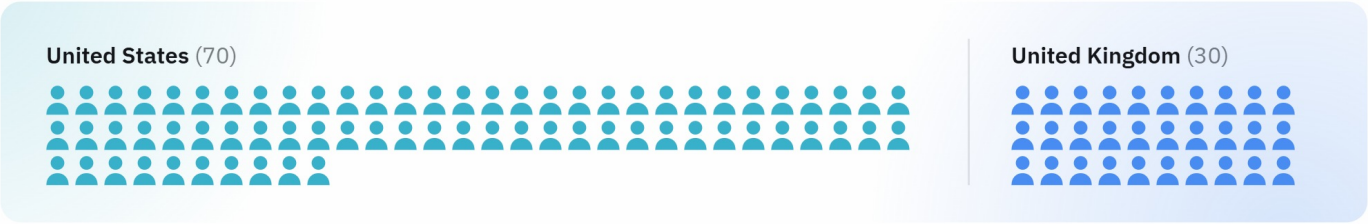
Source: June 2026 CFO survey, n=100, administered by Sapio Research: response to a 30% inference discount for a 12-month commitment.

The throughline of the report is that AI infrastructure economics are not evenly distributed, and the line that divides them is business model. The companies whose product is the AI see their costs, forecast them, price for them, and commit to them. The companies that added AI to an existing product more often carry a cost they cannot fully see, cannot confidently forecast, and cannot reliably price. Both are living with the same rising cost. Only one group is currently equipped to manage it.

About This *Research*

Cloud Capital partnered with Sapio Research, an independent insights firm, to ensure this report reflects decision-grade, unbiased data. The study was designed to capture how senior finance leaders at growth-stage technology companies are navigating the cost, forecasting, and margin pressures of AI infrastructure spend.

Fieldwork was conducted online in June 2026 using a structured email invitation and a secure survey environment. This ensured consistent screening, validated respondent identities, and eliminated channel-based sampling bias.



RESEARCH PARTNER	FIELDWORK DATES	METHOD	RESPONDENT PROFILE
Sapio Research	June 2026	Email invitation to online survey	Senior-most finance leader in the organization
SAMPLE SIZE	**COMPANY SIZE**	**COMPANY TYPE**	**GEOGRAPHIES**
100 qualified respondents	50 to 1,000 employees	AI-native and AI-enabled SaaS / technology	70% United States, 30% United Kingdom

We collected 100 fully qualified responses from the senior-most finance leaders at US- and UK-based companies employing 50 to 1,000 people, the scale at which cloud and AI infrastructure becomes one of the largest contributors to cost of goods sold and operating variability. By focusing exclusively on these leaders, the study captures strategic priorities and real-world decision-making rather than departmental sentiment.

These methodological choices give Cloud Capital and Sapio Research high confidence in the clarity, relevance, and reliability of the findings. Our intent is simple: to equip finance leaders with an objective, data-backed benchmark for navigating the economics of AI infrastructure.

CLOUD CAPITAL

The financial control layer for *cloud and AI infrastructure*

Committing more is the cheapest way to buy cloud, often 20 to 60% less. It's also the riskiest. Cloud Capital underwrites that risk, so finance and engineering can capture the savings without ever being on the hook.

Trusted by

panaseer · Optimizely · Metomic · coverflex · Tyk · MiQ

How we work *with customers.*

Commit more, save more, risk-free

The deepest cloud savings come from committing further than most teams dare. We recommend the commitments your usage supports and underwrite them. If usage shifts or you re-architect, that exposure is ours, not yours.

5% back from day one

Most cloud savings take long-term commitments or engineering time to capture. Every new customer gets an immediate 5% back on eligible AWS services. No commitment, no architectural changes, no catch.

See it and forecast it

A demand forecast grounded in real usage and mapped to actual cost drivers, across COGS and OpEx. For the first time, the whole line is in front of you, and you can take a real number to your board.

"Every percentage point matters for SaaS margins. Cloud Capital gave us the ability to capture savings without adding risk, and those results flow straight into operating leverage."

Myles Johnson
CFO, Optimizely

"Before Cloud Capital, we were focused entirely on building features and had almost no process for optimizing cloud costs. Now we can see efficiency in real time and forecast it with confidence."

Imre Nagy
VP Product & Engineering, Colossyan

AWS Partnership

Certified AWS Advanced Partner

Cloud Capital is an AWS-certified Advanced Partner. We access read-only billing data, fully aligned with AWS terms and conditions.

aws PARTNER Advanced Tier Services · aws PARTNER Cloud Operations Services Competency · aws Qualified Software