

Ten questions to ask an AI-native company before the term sheet, and **what evidence of agent unit economics looks like.**

Written for the investor who has to decide whether a company's adoption will turn into durable gross margin. Growth is visible in the deck. The cost of delivering that growth usually is not, because most companies report AI spend as one blended number. This playbook covers the three moments where that matters: before the term sheet, during portfolio support, and before the next round.

WRITTEN FOR

A partner or principal at a venture firm investing in AI-native companies from Seed to Series C, and the platform and portfolio support teams who help those companies report. You are the person who has to defend the margin assumption in the investment memo.

ALSO FOR

Founders preparing a round, who should read it next to the founder data room guide named at the end. LPs asking how the fund underwrites AI margins can read the metric set and the ten questions to see what the fund asks for.

NOT FOR

Operating partners at buyout firms working on control investments. Their questions are about AI COGS across a held portfolio and exit readiness, and the companion private equity playbook covers them. Nothing here is investment advice.

Nearly all of the money is going to AI, and the margin evidence has not kept up with it.

Two years ago the cost of serving an AI feature was a detail a founder could defer to Series B. It now decides whether the company in front of you is a software business or a services business with a software front end. Most companies can tell you what AI cost last month. Very few can tell you what one customer costs to serve.

87.5%

of US venture dollars in the first half of 2026 went to AI. Median valuation step-ups ran 2.2x for AI companies against 1.6x for everyone else.

Fortune citing PitchBook, August 2026

25% vs 60%

gross margin for Bessemer's fastest-scaling AI cohort, the Supernovas, often negative, against about 60% for the slower Shooting Stars.

Bessemer Venture Partners, State of AI, August 2025

43%

of companies track AI and other costs by customer. 22% track them by transaction.

CloudZero with Benchmarkit, February 2026

66%

of boards now require proof of AI return before approving more AI funding, according to a survey of 260 senior finance leaders.

CloudZero, June 2026

Read the second and third cards together. The fastest-growing AI companies carry the thinnest margins, and fewer than half of companies can say which customers that margin is coming from. An investor who underwrites the growth without the cost per customer is pricing a curve nobody has measured. Both CloudZero figures are vendor-commissioned research, and the Bessemer figure predates 2026. They describe the market, not any company you are looking at.

There are two pools of AI cost in every company you look at, and only one belongs in gross margin.

POOL ONE

The R&D pool

Prototypes, evals, model comparisons, engineers using AI coding tools, features that may never ship. This spend is supposed to be variable. It is how a young company finds the product, and an investor who pushes the founder to make it predictable is asking for less learning.

Know roughly how big it is, who owns it, and whether it is growing faster than headcount. That is enough.

POOL TWO

The production pool

Every model call behind something a customer pays for. The day a feature ships to a paying account, its AI cost stops being research and becomes cost of revenue. It should be attributable to the customer, measured per unit of work, and tied to what the providers invoiced.

This is the pool the ten questions are about, and the one your margin assumption rests on.

Almost no company reports the split, because the instrumentation they have produces one aggregate figure. A founder who arrives with two numbers and a rule for when spend crosses from one to the other has answered half of this playbook before you ask. Do not mark a company down for a large R&D pool. Mark it down when it cannot tell you which experiments became recurring product cost.

THE BENCHMARK REGISTER

What the published data says about AI gross margins. Context for the memo, not a target for the company.

Each row was produced with a different method on a different sample. Use them to frame the question, never as the bar a specific company has to clear.

ICONIQ Growth

Published July 2026

AI product gross margins were 45% in 2025, projected at 53% for 2026 and 59% for 2027, while AI products grow from 32% to a projected 42% of company revenue. Survey of about 300 executives at software companies building AI products, fielded in the second quarter of 2026.

ICONIQ Growth, 2026 State of AI: The Builder's Economy, July 2026

ICONIQ Growth via SaaS Mag

Published 15 May 2026

Average AI product gross margin of 52% in 2026, up from 41% in 2024, against roughly 80% for classic SaaS. The same article estimates a 12 to 17 point gross margin hit for traditional SaaS that adds AI features. That range is the article's synthesis across public disclosures, not a single survey result.

SaaS Mag, The AI COGS Problem, 15 May 2026. Crystil benchmark bm-025

Published guidance for AI founders puts a defensible gross margin at seed at 50 to 60%, with a demonstrated path to expansion, against the 70 to 90% traditional SaaS benchmark.

CRV, LLM Inference Costs: What AI Founders Should Know, 16 July 2026

CORRECTION ONE

A partner reads 53% as the number a portfolio company should hit. It is a projected average across companies of every size and architecture. A seed company at 40% with cost per customer falling every month can be a better position than a Series B at 60% with no per-customer view at all. The trajectory and the evidence carry more weight than the level.

CORRECTION TWO

A partner assumes falling token prices will fix the margin. Token prices have fallen roughly tenfold a year, from \$60 per million in 2021 to about \$0.06, while consumption has exploded (Forbes citing EpochAI, July 2026). Vista Equity Partners finds enterprise agents use 5 to 30 times more tokens than single requests for the same task type (May 2026). Cheaper tokens and costlier tasks can arrive together.

MOMENT ONE. BEFORE THE TERM SHEET

Seven things to test before the margin assumption goes into the memo.

A company at seed will not have all seven. What you are reading is whether the founder knows which are missing and has a plan to measure them.

01 Agent unit economics

The unit the customer buys, whether a seat, a completed task, a resolved ticket or a processed document, and what one of them costs to deliver, including retries, tool calls, retrieval and any human review. A cost per token is a provider statistic. It is not a unit economic.

02 Gross margin trajectory

Gross margin with production AI cost inside cost of revenue, monthly, for as long as the company has paying customers. The direction and the explanation for it matter more than any single month.

03 Cost per customer and per outcome

Cost to serve each account as a distribution, not an average, and the cost of one completed outcome. ICONIQ reports one workflow projected at \$0.10 per run that reached \$1.50 or more as agents retried (July 2026). An average would never have shown it.

04 The R&D and production split

Two numbers, an owner for each, and a written rule for when a workload moves from one pool to the other.

05 **Pass-through model spend**

How much of cost of revenue is spend paid straight through to model providers, and how much of it the pricing recovers. Among consumption-priced AI builders, 26% pass all inference cost through to customers, 58% split it and 15% absorb it entirely (ICONIQ, July 2026, a small sample of 72).

06 **Concentration of cost**

What share of production AI cost sits with the top few customers, and whether any of them are served below cost. TechCrunch reported in April 2026 that Cursor still loses money on individual developer accounts while holding positive margins on enterprise. Segment economics can diverge sharply inside one product.

07 **Provider price and rate exposure**

Share of spend by provider and model, the margin effect of a price change in either direction, and what happens to the product when a rate limit is hit. On 30 July 2026 OpenAI cut one model's price by 80%, and Anthropic's introductory Sonnet 5 pricing ended on 31 August 2026, reverting up 50% (CloudZero and Digital Applied trackers, August 2026).

MOMENT TWO. DURING PORTFOLIO SUPPORT

One monthly metric set, reported the same way by every portfolio company.

The value of a portfolio standard is comparability over time inside each company, not a league table across companies with different architectures. Agree the definitions at the first board meeting after close and ask for the same eight lines every month. A company that cannot yet produce a line reports it as not measured, with a date.

METRIC	DEFINITION	WHY IT MATTERS	GOOD EVIDENCE
Cost per customer	Production AI cost attributed to each paying account, reported as median and top decile.	Shows whether growth adds margin or adds cost faster than revenue.	Attributed at the time of the call, not allocated afterwards by revenue share.
Cost per outcome	Total model cost of one completed unit of work, retries and tool calls included.	The number pricing has to cover. It moves when agents loop.	Calls grouped into the workflow they served, with completion defined in writing.
AI gross margin	Gross margin with production AI cost inside cost of revenue.	The line the next investor will rebuild from source.	The same treatment every month, with the R&D pool excluded by rule.

R&D pool spend	AI cost outside production, including coding tools, evals and experiments.	Healthy when owned and visible. A problem only when it hides production cost.	A named owner and a monthly total, not a cap.
Pass-through ratio	Share of production model spend recovered in customer pricing.	Shows whether a provider price change lands on the customer or the margin.	Model spend mapped to each pricing plan.
Cost concentration	Share of production AI cost carried by the top five accounts, and the count of accounts served below cost.	One heavy account can swing a small company's margin.	Named accounts, with the revenue for each beside the cost.
Provider mix	Share of production spend by provider and model.	Sizes the exposure to a price change, a deprecation or a rate limit.	Taken from tracked usage and checked against each provider invoice.
Record variance	Gap between tracked AI cost and what providers invoiced, with each gap explained.	A number that does not tie to the bill is an estimate.	A monthly reconciliation finance signs, with the largest gap named.

MOMENT THREE. BEFORE THE NEXT ROUND

The evidence a later-stage investor's diligence will ask for, assembled before the process starts.

The next lead will rebuild the margin from source records rather than accept the deck. A company that keeps the monthly metric set running walks in with an update, not a reconstruction.

- **Twelve months of cost per customer.** Monthly, as a distribution, with the explanation for each change in direction. A rising line with a reason and a plan is stronger than no line.
- **The pool rule and its history.** The written rule for R&D and production, and the workloads that crossed over, with the date each one became cost of revenue.
- **The pricing to cost map.** Each plan, its price, the model cost behind it, the break-even usage, and how many accounts are past it today.
- **The provider sensitivity.** Share of spend by provider and what a published price move, such as the 50% Sonnet 5 reversion, would have done to gross margin.
- **The reconciliation history.** Tracked cost against provider invoices, month by month, with every gap named and the largest one explained.
- **The owner and the cadence.** One name, the monthly review the numbers go to, and the decisions they changed in the last two quarters.

Ask in this order. The first two decide whether the others can be answered at all.

Each question is asked by the investor or the founder, the CTO or the CFO. Each comes with what a defensible answer sounds like, what a plausible one sounds like, and the follow-up that separates them. The plausible answers are not dishonest. They are what a competent founder says when the instrumentation cannot do better, and they are the answers that become an open item in the memo.

QUESTION 01 ASK THE FOUNDER AND CTO

What unit do you sell, and what does one cost you to deliver?

The agent unit economics question, and the foundation for the other nine. Enterprise agents use 5 to 30 times more tokens than single requests for the same task type (Vista Equity Partners, May 2026), and 68% of AI builders found the true cost of internal AI very or moderately difficult to predict (ICONIQ, July 2026). A cost per call says nothing about the cost of the work the customer is paying for.

A DEFENSIBLE ANSWER

A named unit that matches the pricing, a cost per unit with retries, tool calls and retrieval inside it, and the boundary of what is included stated plainly.

A PLAUSIBLE ANSWER

"Our blended cost per call is low and falling." True, and it is the provider's unit, not the customer's.

THEN ASK

Walk me from one real customer outcome last month to the calls that delivered it and what they cost.

QUESTION 02 ASK THE CTO

How much of your AI spend is R&D and how much sits behind something a customer pays for?

The margin in the deck depends on the answer. Software companies building AI products spent 11% of revenue on AI for internal productivity in 2025, projected to reach 16% in 2026 (ICONIQ, July 2026). Mixed into one line with production cost, that spend makes both numbers unreadable.

A DEFENSIBLE ANSWER

Two monthly figures, an owner for each, and a written rule for when a workload becomes cost of revenue, applied from the day it ships to a paying account.

A PLAUSIBLE ANSWER

"It is all R&D at this stage." Possible at pre-seed. Not once customers are paying for an AI feature.

THEN ASK

Which workloads crossed from R&D into production in the last two quarters, and on what date?

QUESTION 03 ASK THE CFO OR FOUNDER

What is your gross margin trajectory with AI cost in cost of revenue, month by month?

The question the valuation turns on. ICONIQ reports AI product gross margins of 45% in 2025, projected at 53% for 2026 (July 2026). CRV puts the defensible seed band at 50 to 60% with a demonstrated path to expansion (July 2026). The path is what CRV asks for, which means a trend line, not a single quarter.

A DEFENSIBLE ANSWER

Monthly gross margin since the first paying customer, with production AI cost inside it, the same treatment every month, and a reason for each change in direction.

A PLAUSIBLE ANSWER

"Margins will expand as model prices fall." A forecast resting on the provider's roadmap, not the company's.

THEN ASK

Which three changes you made moved the margin most, and by how much each?

QUESTION 04 ASK THE CFO

How much of your cost of revenue is model spend paid through to providers, and how much of it does your pricing recover?

Inference consumes roughly 23% of revenue at scaling-stage AI B2B companies (ICONIQ snapshot via SaaS Mag, January 2026). Among consumption-priced builders, only 26% pass all of it through to customers, 58% split it and 15% absorb it (ICONIQ, July 2026, sample of 72). Where the cost lands decides who pays for the next provider price change.

A DEFENSIBLE ANSWER

Model spend as a share of cost of revenue, mapped to each pricing plan, with the recovered share stated per plan.

A PLAUSIBLE ANSWER

"Our pricing covers our costs." An aggregate statement about a cost that varies by customer and by plan.

THEN ASK

On which plan is model spend the largest share of price, and how many customers are on it?

QUESTION 05 ASK THE FOUNDER

Which customers cost the most to serve, and how concentrated is your AI cost?

The tail question. Only 43% of companies track costs by customer and 22% by transaction (CloudZero with Benchmarkit, February 2026). Segment economics can diverge inside one product. TechCrunch reported Cursor still losing money on individual developer accounts while holding positive margins on enterprise (April 2026, sourced reporting rather than company disclosure).

A DEFENSIBLE ANSWER

The top five accounts by AI cost, by name, with revenue beside each, the share of production cost they carry, and a count of accounts served below cost.

A PLAUSIBLE ANSWER

"Usage is fairly even across the base." An average offered where a distribution was asked for.

THEN ASK

Is any customer unprofitable today, and how would you know if one became so next month?

QUESTION 06 ASK THE FOUNDER

Are you charging a flat price for anything whose cost underneath is usage-based?

The pricing shape that breaks first at scale. Consumption pricing among AI builders rose from 35% to 42% in six months, and outcome-based pricing from 18% to 23% (ICONIQ, July 2026). GitHub moved every Copilot plan to usage-based billing on 1 June 2026 (GitHub, April 2026). When a category leader changes its pricing shape, the cost shape underneath is usually why.

A DEFENSIBLE ANSWER

The plans with that shape named, the break-even usage for each, and the number of accounts already above it.

A PLAUSIBLE ANSWER

"Heavy users are rare." Possibly true today, and it says nothing about the customers you are about to fund the company to win.

THEN ASK

What does gross margin look like if the median customer's usage doubles on the current price?

QUESTION 07 ASK THE CTO

What happens to your margin and your product if a provider changes its price or its rate limits?

Concentration risk, and it moves in both directions. AI builders now run about 3.3 model providers on average, and 81% cite Anthropic (ICONIQ, July 2026). In one 32-day window an OpenAI model's price fell 80% and Anthropic's introductory Sonnet 5 pricing reverted up 50% (CloudZero and Digital Applied, August 2026). A rate limit does not show on the invoice. It shows up as a customer outage.

A DEFENSIBLE ANSWER

Share of production spend by provider and model, the margin effect of a stated price move already calculated, and the fallback when a limit is hit, tested.

A PLAUSIBLE ANSWER

"We are model-agnostic and could switch in a week." Probably true in principle, and silent on cost, quality and what breaks.

THEN ASK

When did you last switch a production workload between models, and what did it do to cost per outcome and quality?

QUESTION 08 ASK THE CTO

If one customer's usage tripled tonight, would you know before the invoice?

The controls question. 72% of organizations had an unexpected AI cost spike in the past year, a third were caught off guard more than once, and only 20% could identify the cause of a spend doubling within hours (Harness, July 2026, a vendor-sponsored survey of 700 leaders). In a company with a few large customers, one spike is a margin event.

A DEFENSIBLE ANSWER

A threshold per customer and per workflow, a named person who receives the alert, and an example of a spike caught early.

A PLAUSIBLE ANSWER

"We review spend every month." Monthly review finds the spike after it has been paid for.

THEN ASK

What was the largest spike in the last six months, how long did it take to find the cause, and who found it?

QUESTION 09 ASK THE CFO

Does your tracked AI cost tie to what the providers actually invoiced?

The question that decides whether every other number is a measurement or an estimate. Only 13% of organizations have even basic visibility into actual AI spending (Harness, July 2026). A figure computed from logs has no way of knowing it is wrong. The invoice does.

A DEFENSIBLE ANSWER

A monthly comparison of tracked cost to each provider invoice, done by a named person in finance, with the last gap and its cause stated.

A PLAUSIBLE ANSWER

"Our dashboard tracks it in real time." Real time is not the same as reconciled.

THEN ASK

When did the record last fail to match an invoice, by how much, and why?

Who owns these numbers?

Close on accountability. 52% of organizations say there is no clear owner of AI cost (Harness, July 2026). In a startup the cost is created by engineers and reported by whoever does finance, and without one owner the cost per customer is nobody's forecast.

A DEFENSIBLE ANSWER

One name. At a small company that is often the founder or the CTO, said plainly, with the monthly review the numbers go to.

A PLAUSIBLE ANSWER

"Engineering and finance share it."
Shared ownership of a forecast reads as unowned.

THEN ASK

What decision did these numbers change last quarter?

READING THE ANSWERS

Count the defensible ones, not the confident ones.

Score an answer as defensible only if it came with a document, a name or a date attached. Confidence is not evidence, and in an investment memo it becomes the assumption you are asked about at the next partners' meeting. "We do not measure that yet, and here is when we will" is a stronger answer than a confident figure that falls apart under one follow-up.

EIGHT TO TEN

Rare, and worth saying so in the memo.

The remaining gap is usually question nine, because tying tracked cost to the invoice is the step most tooling leaves to finance. Close it and the margin line can be rebuilt from source.

Ask for the monthly metric set from the first board meeting after close.

FOUR TO SEVEN

Normal, and a support plan.

Do not ask for every gap to close before the round. Questions one and two are the foundation, and questions five and nine are the ones the next lead will probe hardest.

Put a name and a date against those four in the first hundred days.

THREE OR FEWER

An instrumentation gap, not a judgement on the team.

Most of the market sits here, since fewer than half of companies attribute cost by customer at all. The risk is that the gap is found by the next investor, on their timetable.

Start with question two. The split changes three other answers.

If an LP asks how the fund underwrites AI margins, show the metric set and these ten questions. What an LP can test is a definition applied the same way across the portfolio, not a target margin. Adams Street found 28% of LPs cite technology disruption as a key risk, up from 17% a year earlier (March 2026).

Agent unit economics. Ten questions for the founder.

Print this, or send it to the founder before the meeting so both sides are working from the same sheet.

- | | | |
|----|--|--------------|
| 01 | What unit do you sell, and what does one cost you to deliver? | FOUNDER, CTO |
| 02 | How much of your AI spend is R&D and how much sits behind something a customer pays for? | CTO |
| 03 | What is your gross margin trajectory with AI cost in cost of revenue, month by month? | CFO, FOUNDER |
| 04 | How much of cost of revenue is model spend paid through to providers, and how much does pricing recover? | CFO |
| 05 | Which customers cost the most to serve, and how concentrated is your AI cost? | FOUNDER |
| 06 | Are you charging a flat price for anything whose cost underneath is usage-based? | FOUNDER |
| 07 | What happens to your margin and your product if a provider changes its price or its rate limits? | CTO |
| 08 | If one customer's usage tripled tonight, would you know before the invoice? | CTO |
| 09 | Does your tracked AI cost tie to what the providers actually invoiced? | CFO |
| 10 | Who owns these numbers? | FOUNDER |

Score an answer as defensible only if it arrived with a document, a name, or a date attached. After close, ask for the eight-line monthly metric set from the first board meeting.

WHERE THIS CAME FROM

We measure AI cost at the unit an investor underwrites.

Crystil is an independent third party, SOC 2 compliant. Its Cost Tracking SDK is one line of code. It prices every model call by input, cached and output tokens and the model used, and attributes the cost to a customer, organization or team, with an optional department or team beneath it.

Workflows group many model calls into one logical task, so a completed outcome has one cost.

Customers sets cost per customer beside revenue per customer to give margin per customer.

Pulse routes a team's AI coding tools, such as Claude Code, Codex and Gemini CLI, through Crystil with no SDK, which makes the R&D pool visible.

We do not reconcile provider invoices, set target margins or value companies. We provide the attributed record a finance team reconciles against its invoices. Questions one, two, four and five are the ones our record is built to answer.

[Talk to an AI measurement expert](#)

crystil.com

COMPANION GUIDES

For founders: the founder data room guide, Raise readiness 2027, with a ninety-day plan. **For buyout operating partners:** the private equity portfolio playbook.

Sources

1. Fortune citing PitchBook, 12 August 2026. Bessemer Venture Partners, The State of AI 2025, 13 August 2025, published before 2026.
2. CloudZero with Benchmarkit, FinOps In The AI Era, 12 February 2026, and CloudZero, survey of 260 finance leaders, 24 June 2026. Both vendor-commissioned.
3. ICONIQ Growth, 2026 State of AI: The Builder's Economy, July 2026. The pass-through figure rests on 72 respondents.
4. SaaS Mag citing ICONIQ Growth, The AI COGS Problem, 15 May 2026, Crystil benchmark bm-025. ICONIQ Growth, State of AI Bi-Annual Snapshot, January 2026, via SaaS Mag and SaaSTr.
5. CRV, LLM Inference Costs: What AI Founders Should Know, 16 July 2026.
6. Forbes citing EpochAI, 28 July 2026. Vista Equity Partners, AI Inference Economics: A Framework for Investors, 27 May 2026.
7. TechCrunch, reporting on Cursor, 17 April 2026. Sourced reporting, not company disclosure.
8. CloudZero and Digital Applied pricing trackers, 5 August 2026. GitHub, GitHub Copilot is moving to usage-based billing, 27 April 2026.
9. Harness, 2026 State of AI in FinOps, 29 July 2026. Survey of 700 engineering and FinOps leaders by Sapio Research, vendor-sponsored.
10. Adams Street Partners, 2026 Global Investor Survey, 11 March 2026.

Crystil. v1.0, 7 October 2026. Every figure is published and dated. Nothing here is modelled, estimated or invented. This document describes what published sources say. It is not legal, tax or investment advice.