Probabilistic COGS: Why AI ARR Is Disguised Consulting

Second Order Labs™ · June 2026

ABSTRACT

Much of what AI startups report as recurring revenue is consulting work in disguise, and when the forward deployed engineers leave, the ARR leaves with them.

Keywords: AI economics, SaaS valuations, venture capital, enterprise AI, gross margins

> *"In many cases, AI companies simply don't have the same economic construction as software businesses. At times, they can even look more like traditional services businesses."*
>
> — Andreessen Horowitz, The New Business of AI

I. The ARR That Disappears When the Engineers Leave

Pull apart an AI startup's revenue and it stops looking like software. Freeze the implementation team and the product gets worse. Stop tuning prompts and hal-

Public SaaS-style multiples on AI names still assume margins will expand like software margins. Series B investors and public-market buyers kept underwriting that story through 2024. The problem is simple: integration is not a phase you finish. The human work required to keep a probabilistic system useful inside bad enterprise data keeps coming back. That cost be-

build-once, sell-forever model and keeps gross margins closer to an agency than a software platform. Sooner or later, valuation follows the economics.

II. Why the SaaS Premium Was Always Misapplied

The SaaS premium exists because software has near-zero marginal cost. It does not fit AI very well, because AI often looks like a tech-enabled services business. Andreessen Horowitz put gross margins for AI companies in the 50 to 60 percent range, below the 60 to 80 percent-plus range for traditional SaaS.[1] That spread is not a footnote. It is the story.

The number matters. The reason matters more. Traditional SaaS spreads one engineering effort across thousands of customers. AI cannot, because each enterprise has its own data stack and its own error tolerance. Tomasz Tunguz gets to the point: AI startups face margin pressure from inference on the back end and implementation labor on the front.[2] Investor notes usually model inference costs explicitly but treat implementation labor as temporary onboarding. That assumption is doing a lot of work.

> *"In many cases, AI companies simply don't have the same economic construction as software businesses. At times, they can even look more like traditional services businesses." Andreessen Horowitz, The New Business of AI*

When the firm's own analysts say the model looks like services, the burden shifts. The debate is no longer whether margins are lower. The debate is whether they ever become software margins at all.

Figure 1: The visible product is light; the deployment machinery underneath carries the weight.

III. The Eval Trap and the ARR Mirage

The Eval Trap turns AI software into AI services. LLM output is probabilistic, so every enterprise ends up needing its own evaluation stack to catch failures before they hit production. LlamaIndex says it plainly: building a RAG pipeline is easy, but evaluating it and making it production-ready for a specific enterprise is hard.[3] In normal software, you build a feature once and ship it widely. In AI, the feature has to be validated against the client's data and the failure modes that matter to that client. Harrison Chase of LangChain makes the same point from the architecture side. Each company needs a slightly different setup, and there is no universal agent.[4] The eval does not get written once. It gets maintained, retuned, and checked again as the model drifts and the data changes. The labor stays attached to the contract.

Disguised Retainer Revenue is recurring revenue that depends on ongoing human labor, even though it is billed like subscription software. The test is simple: remove the people and see if the revenue survives. With real SaaS, it does. With a lot of enterprise AI, it does not. C3.ai shows the pattern in its filings. The company says services revenue comes from implementation projects and training engagements, and it also says the model requires significant investment in customer deployment.[7] A company like Salesforce does not carry forward-deployment labor as a core recurring delivery cost for each account. Palantir's S-1 is even clearer: its Forward Deployed Software Engineers handle deployment and configuration, and their costs are a significant part of cost of revenue.[8] Once a company books integration labor above the gross margin line, it is admitting that labor is part of the product. Many enterprise AI vendors do some version of this while still presenting ARR as if it were pure software.

> *Strip the forward deployed engineers out of an AI contract and the ARR doesn't shrink gracefully, it evaporates.*

Figure 2: Andreessen Horowitz's benchmark: AI margins sit a full tier below the SaaS band the premium assumes.

Revenue Type	Survives FDE Removal?	Gross Margin Profile	Correct Classification
Traditional SaaS seat	Yes	70-80%	Software
AI subscription (productized)	Yes	55-65%	Software
AI subscription (FDE-dependent)	No	30-50%	Services / Retainer
Pure professional services	No	20-35%	Services

IV. Selling Work Means Owning the Liability

Bessemer describes the shift as software-as-a-service turning into service-as-a-software.[9] The wording sounds cute. The accounting consequence is not. Sell a tool and the customer owns the outcome. Sell the work and you own the outcome. Sarah Tavel of Benchmark says the next wave will not sell software. It will sell work.[10] Higher-value pitch, worse margin profile. Labor does not go to zero.

Selling an AI agent is economically closer to hiring a worker than buying a seat license. TechCrunch captured the mood inside the industry: startups have realized that selling AI looks more like onboarding a digital employee.[11] An employee needs ongoing supervision and correction. So does an agent. Bret Taylor, co-founder of Sierra, says you cannot just drop an LLM into an enterprise. You have to connect it to systems of record, which means deep integration work.[12] Snowflake says the same from the data side. No AI strategy without a data strategy, and data prep is still the bottleneck for generative AI.[13] Early-stage AI vendors often spend weeks cleaning customer data pipelines before the product works well enough to renew.

Figure 3: An agent enters the enterprise like a new hire, not like a license key.

V. Who Actually Captures the Last Mile

The Palantir Illusion traps a lot of founders. Yes, Palantir built a real business with forward deployed engineers and its AIP Bootcamp model, putting top engineers directly in the room with customer data.[15] But Palantir can carry that model because it has large contracts that absorb expensive deployment labor. A seed-stage startup pointing to Palantir is copying the theater without the balance sheet.

Follow the money. Accenture reported more than $600 million in new generative AI bookings in one quarter, and its clients keep discovering that scaling AI starts with a lot of data and services work.[16] The consultancy wins because the last mile is a services problem, and Accenture never pretended otherwise. AI startups often do the same work while calling it software. Elad Gil's framing is right: many of them are acting as outsourced R&D and integration teams for enterprises.[17] Sequoia's David Cahn has pointed to the widening gap between the revenue implied by AI infrastructure spending and the revenue actually showing up.[18] Part of that gap is human throughput. Deployment cannot scale at the speed of capital because each enterprise needs custom integration and ongoing evaluation.

VI. What Founders and Investors Should Do Before the Correction

If revenue depends on FDE labor, investors will eventually value that revenue more like services than software. The only real question is timing.

Start with one number: subscription revenue that survives if you pull the forward deployed engineer. That is the software business. Everything else deserves a different label, a different margin expectation, and probably a different multiple.

Use a harder operating rule. If an account churns when you remove the FDE, that revenue is services. Track implementation labor in COGS against subscription revenue every quarter. If the ratio rises as you grow, you are scaling a consultancy. And wherever the same eval work shows up twice, turn it into product immediately. That is the only path out of labor-backed ARR.

Before anyone slaps a SaaS multiple on an AI company, they should ask for FDE-independent revenue. Once one public company is forced to split software ARR from labor-backed revenue, the sector's multiples will reset in a week. When boards start asking for that number, half the category will find out it never had software margins in the first place.

KEY FINDINGS

- Andreessen Horowitz pegs AI gross margins at 50-60%, well below the 60-80%+ that traditional SaaS valuations assume, exposing a structural mispricing.
- The Eval Trap forces AI startups to build bespoke human evaluation pipelines per client, breaking the build-once-sell-infinitely model that justifies software multiples.
- Forward deployed engineer costs land in cost of revenue, not OpEx, structurally depressing AI gross margins toward consulting levels.
- Selling an AI agent resembles onboarding a digital employee that needs training and alignment, not provisioning a self-sustaining SaaS seat.
- Consultancies like Accenture are capturing the GenAI windfall because the last mile of enterprise AI is fundamentally a services problem.

REFERENCES

[1] Andreessen Horowitz, "The New Business of AI: How Is AI Changing the Software Business?", gross margins for AI companies in the 50-60% range versus 60-80%+ for SaaS. https://a16z.com/the-new-business-of-ai-and-how-its-different-from-traditional-software/

[2] Tomasz Tunguz, "The Margin Structures of AI Companies", dual margin pressure of inference compute and human implementation.

[3] LlamaIndex, "The RAG Evaluation Landscape", building RAG is easy; making it production-ready per enterprise is hard.

[4] Harrison Chase, Latent Space Podcast, no one-size-fits-all agent; cognitive architecture differs per company. https://www.latent.space/p/langchain

[5] Anthropic Documentation, "Prompt Engineering", iterative, highly experimental process tied to specific use case and data. https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/overview

[6] UC Berkeley, "Operationalizing Machine Learning: An Interview Study", deployment as ongoing operational burden requiring continuous human intervention. https://arxiv.org/abs/2209.09125

[7] C3.ai, Form 10-K (FY2023), professional services revenue from implementation, training, advisory; significant deployment investment. https://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&CIK=0001577526&type=10-K&dateb=&owner=exclude&count=100

[8] Palantir Technologies, Form S-1, FDSE costs as a significant component of cost of revenue. https://www.sec.gov/Archives/edgar/data/0001321655/000119312520230013/d904406ds1.htm

[9] Bessemer Venture Partners, "State of the Cloud 2024", shift from software-as-a-service to service-as-a-software. https://www.bvp.com/atlas/state-of-the-cloud-2024

[10] Sarah Tavel, "Selling Work, Not Software", next generation of AI startups will sell work, not software. https://www.sarahtavel.com/p/a-few-sell-work-not-software-updated

[11] TechCrunch, "The rise of the forward-deployed engineer in AI", selling AI as onboarding a digital employee.

[12] Bret Taylor, Stratechery interview, wiring LLMs into systems of record requires deep, specific integration. https://stratechery.com/2025/an-interview-with-sierra-founder-and-ceo-bret-taylor-about-ai-agents-and-tech-history-lessons/

[13] Snowflake, Q4 2024 Earnings Call, no AI strategy without a data strategy; data prep as the bottleneck. https://investors.snowflake.com/events-and-presentations/event-details/2024/Q4-FY24-Snowflake-Earnings-Conference-Call/

[14] Stanford HAI, "Artificial Intelligence Index Report 2024", deployment cost often exceeds software cost. https://hai.stanford.edu/ai-index/2024-ai-index-report

[15] Palantir Q4 2023 Earnings Call (Alex Karp), AIP Bootcamps placing engineers directly with customer data. https://www.sec.gov/Archives/edgar/data/0001321655/000132165523000005/a2022q4ex991pressrelease.htm

[16] Accenture, Q2 Fiscal 2024 Earnings Call, over $600 million in new generative AI bookings; data foundation requires services. https://www.sec.gov/Archives/edgar/data/0001467373/000146737324000106/fy24q2earnings8-kexhibit.htm

[17] Elad Gil, "AI Startup vs Incumbent Value", startups acting as outsourced R&D and integration teams. https://blog.eladgil.com/p/ai-startup-vs-incumbent-value

[18] Sequoia Capital, "AI's $600B Question", widening gap between infrastructure build-out and actual AI revenue. https://sequoiacap.com/article/ais-600b-question/
