JAN 14, 2026

Our Predictions for 2026

What 2025 Got Wrong and the Bets We're Making for This Year



By Jenny Xiao, Jay Zhao, and Liang Wu

Leonis Capital · leoniscap.com/research/our-predictions-for-2026/



Why We're Not Writing Another "State of AI" Report

For the last few years, we opened the year with a "State of AI" report: a broad survey of models, infrastructure, applications, and capital flows. We know how to write those reports. We know how to make them comprehensive, well-cited, and directionally correct.

But they also suffered from the same structural problem most AI reports now have: they try to say something about *everything* and, in doing so, avoid saying anything sharp about what actually matters. In a field moving as quickly as AI, breadth has become a form of risk aversion.

This year, we're doing the opposite.

Instead of attempting another exhaustive "State of AI," we are narrowing our scope to the slice of work that actually drives outcomes for us as investors. This piece is not meant to catalog the industry. It is meant to surface where our expectations broke, what those breaks revealed about the underlying structure of AI, and where we are willing to make explicit bets going forward.

Specifically, we focus on three things:

- What genuinely surprised us in 2025, given what we believed at the end of 2024.
- What those surprises reveal about the structure of the AI industry, technically, economically, and organizationally.
- Three contrarian, testable predictions about 2026 that we are willing to stand behind.

What Surprised Us in 2025, By 2024 Standards

If you froze the world at the start of 2024 and showed most founders and investors the actual state of AI at the end of 2025, several developments would have been genuinely surprising yet obvious in hindsight.

Surprise 1: Agentic systems largely flopped, not because of intelligence, but because state and runtime were unsolved

In 2024, autonomous agents sat near the peak of the hype cycle. "AI employees," multi-agent systems, and the end of SaaS were treated as imminent. We shared more of this optimism than we should have.

By 2025, the gap between narrative and reality was clear.

Retrieval alone proved insufficient. No robust, stateful runtime layer existed. Agents collapsed under unpredictability, reliability, and coordination failures. Systems that looked impressive in controlled demos broke down in production environments.

The companies that actually shipped value did not build "AI employees." They built deterministic workflows with tightly bounded autonomy. The failure was not about model intelligence; it was about systems design. Agents failed because the infrastructure required to support persistent state, error recovery, and long-running execution never materialized.

Surprise 2: Open-source models start to seriously challenge closed-source models in enterprise settings

In 2024, the dominant belief was that open-source models were fast, cheap, and could handle narrow or low-stakes tasks, but would be insufficient for serious enterprise workloads.

By early 2025, that belief no longer held.

Open-source models like DeepSeek reached o1-level performance. The quality gap effectively vanished for 80-90% of enterprise use cases, especially anything not requiring bleeding-edge agentic reliability or ultra-long reasoning chains. Open-source fine-tuned models became the default choice for narrow, production-relevant workloads, even as closed-source models continued to pull ahead on the capability frontier. [1]

For many buyers, benchmark rankings matter far less than auditability, control, deployment flexibility, and data governance. The question quietly shifted from *"Can open source catch up?"* to *"In which workloads is closed source still worth the premium?"* That inversion was not widely anticipated at the end of 2024.

Surprise 3: TPU meaningfully cracked Nvidia's moat

In 2024, CUDA's dominance looked close to permanent. Alternative silicon was treated as a rounding error: interesting in theory, but irrelevant at scale. The consensus view was that Nvidia's moat would erode slowly, if at all, and only from the bottom up.

What happened in 2025 was different.

Anthropic trained Opus primarily on TPUs. OpenAI also started renting TPUs to run ChatGPT. Google began actively selling TPUs to external labs, which felt like an unthinkable move a year earlier when most observers assumed TPUs existed primarily to reinforce GCP's internal moat. AWS Trainium also emerged as a credible training layer.

The surprising part was not that alternatives improved, but *where* the pressure appeared. The erosion of Nvidia's moat began with top-tier workloads and the largest buyers, not with scrappy upstarts nibbling at the edges. That was a platform-level shift, not incremental competition.

Why People Misread 2025, and Why They Will Misread 2026 for the Exact Same Reasons

We don't think the mistakes people made in reading 2025 were one-off errors. They reflect deeper, recurring issues in how Silicon Valley reasons about AI progress.

Two factors matter the most.

First, **AI narratives systematically lag reality by 12–24 months**. Most people build their understanding from what's most visible: benchmarks, model launches, press releases, and social media commentary. These signals are easy to track and debate, but they're downstream indicators reflecting what has already crystallized, not what's quietly shifting underneath. This lag is compounded by AI's sharp discontinuities, which defy the smooth forecasting curves most people default to. In 2024, the reference points were model performance, GPU shortages, and imminent agent automation. These narratives persisted into 2025.

Second, **people consistently overestimate technical progress while underestimating institutional constraints**. In 2024, much of the discourse treated models like racehorses: who had the longest context window, who gained a few points on a benchmark, who shipped the next frontier system first. Meanwhile, Silicon Valley consistently underestimated institutional inertia. Enterprises, governments, and compliance-heavy organizations don't move on model release cycles; they move on procurement timelines, risk committees, and legacy system dependencies.

These dynamics explain not just why 2025 was misread, but why 2026 will be too. Narratives will gravitate toward visible technical breakthroughs while real change accumulates in boring places: infrastructure economics, institutional adoption, workflow integration, etc. Those who anchor on demos and benchmarks will keep being surprised. Those who pay attention to incentives, cost curves, and organizational reality will not.

Our Predictions for 2026

With the above reflections in mind, we want to make three predictions for 2026, one at each layer of the AI ecosystem: infrastructure, foundation models, and applications. Our goal is not to forecast breakthroughs, but to trace where incentives, economics, and institutional constraints are likely to push the AI industry next.

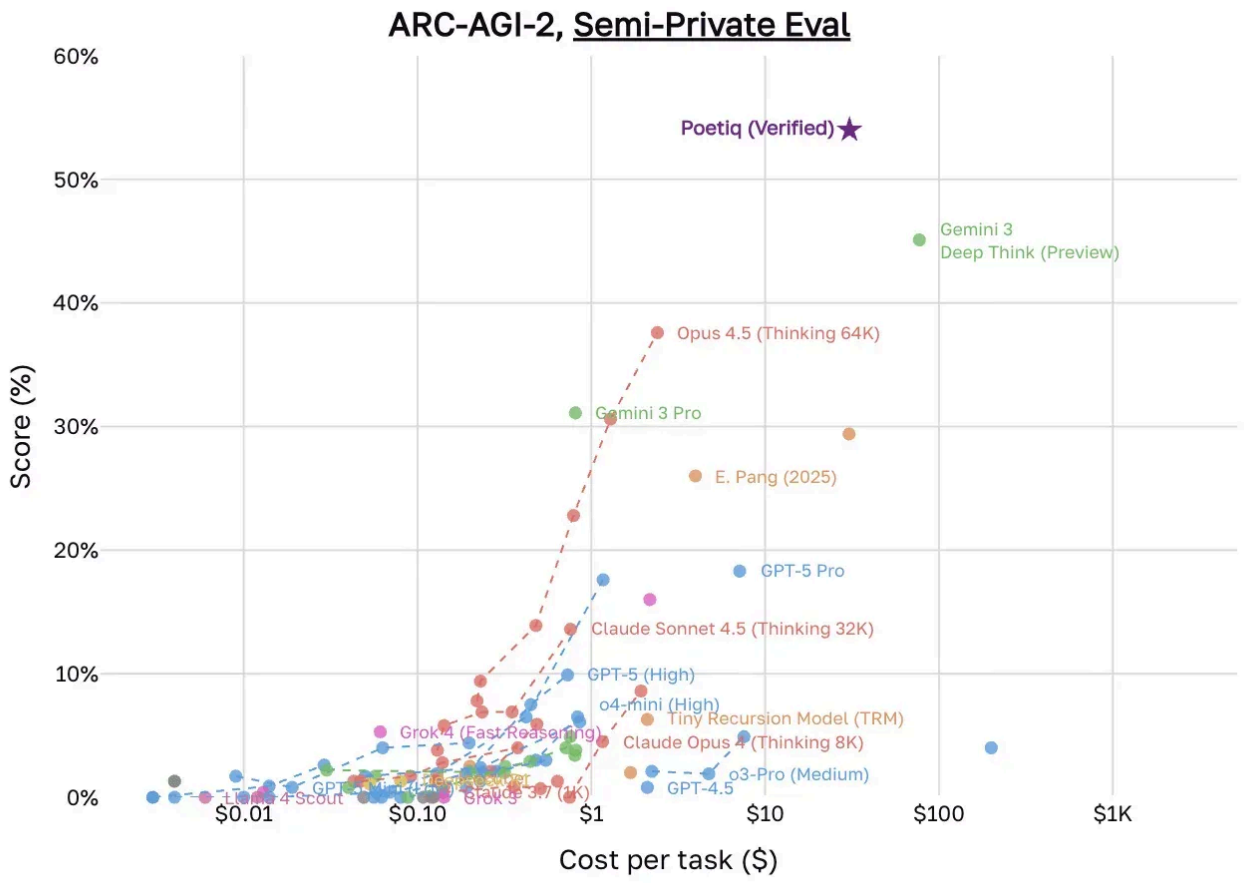
Prediction 1: Agent harnesses unbundle AI performance from model training.

Indicator: By the end of 2026, at least three organizations that do not train their own foundation models will have topped major AI benchmark leaderboards.

Agent harnesses improve model performance not by training new models, but by systematically extracting more capability from models that already exist. By structuring work over time, breaking tasks into steps, checking results, and iterating, these systems allow foundation models to handle more complex, long-running work. **The result is a shift in where performance comes less from training and more from the systems that orchestrate models at runtime.**

This shift is visible, but it remains underestimated by much of Silicon Valley. Historically, frontier gains in AI were tightly linked to larger models, more data, and more compute spent during training.

The rise of agent harnesses means that small teams with relatively limited capital can compete on model performance with foundation model labs. **Poetiq is an early example.** In December 2025, the company topped the ARC-AGI-2 benchmark, becoming the first system to exceed 50% accuracy, while outperforming Google's Gemini 3 Deep Think at roughly half the per-task cost. The team had fewer than ten people, and they achieved these results without training or fine-tuning. Instead, their system focused entirely on learned test-time reasoning: constructing optimized inference pipelines through iterative refinement, dynamic reasoning chains, and verification passes.



Source: Poetiq

Frontier labs are already internalizing this shift. In November 2025, engineers at Anthropic published a detailed account of how to build effective harnesses for long-running agents. Google followed suit in December 2025 with the launch of the Gemini Conductor, formalizing a "Context-Driven Development" (CDD) workflow. Both labs' work demonstrates that frontier models already contain sufficient latent capability, but reliably realizing that capability requires structured test-time systems that include persistent state, incremental task decomposition, verification, and clean handoffs.

The economic implications are significant. If capability gains shift from training scale to inference-time orchestration, value capture also shifts to systems that can reliably extract and operationalize model intelligence. Under this lens, AI scaling becomes a software and systems problem rather than a pure capital-intensive arms race to develop the best model. This could create **a new infrastructure layer between models and applications**.

The open question for 2026 is whether this shift remains an internal competency of frontier labs or becomes a standalone category. In the short run, we expect self-improving agent harnesses to attract billions in venture funding as investors underwrite test-time orchestration as a primary driver of performance. However, over the long run, these capabilities are more likely to be absorbed by model labs and GPU clouds, where control over models, infrastructure, and inference economics confers a structural advantage. Whether agent-harness startups can establish defensible positions before that absorption occurs will determine whether any independent companies endure.

Prediction 2: Foundation model labs will deliberately de-emphasize their API businesses and reserve frontier capabilities for first-party products

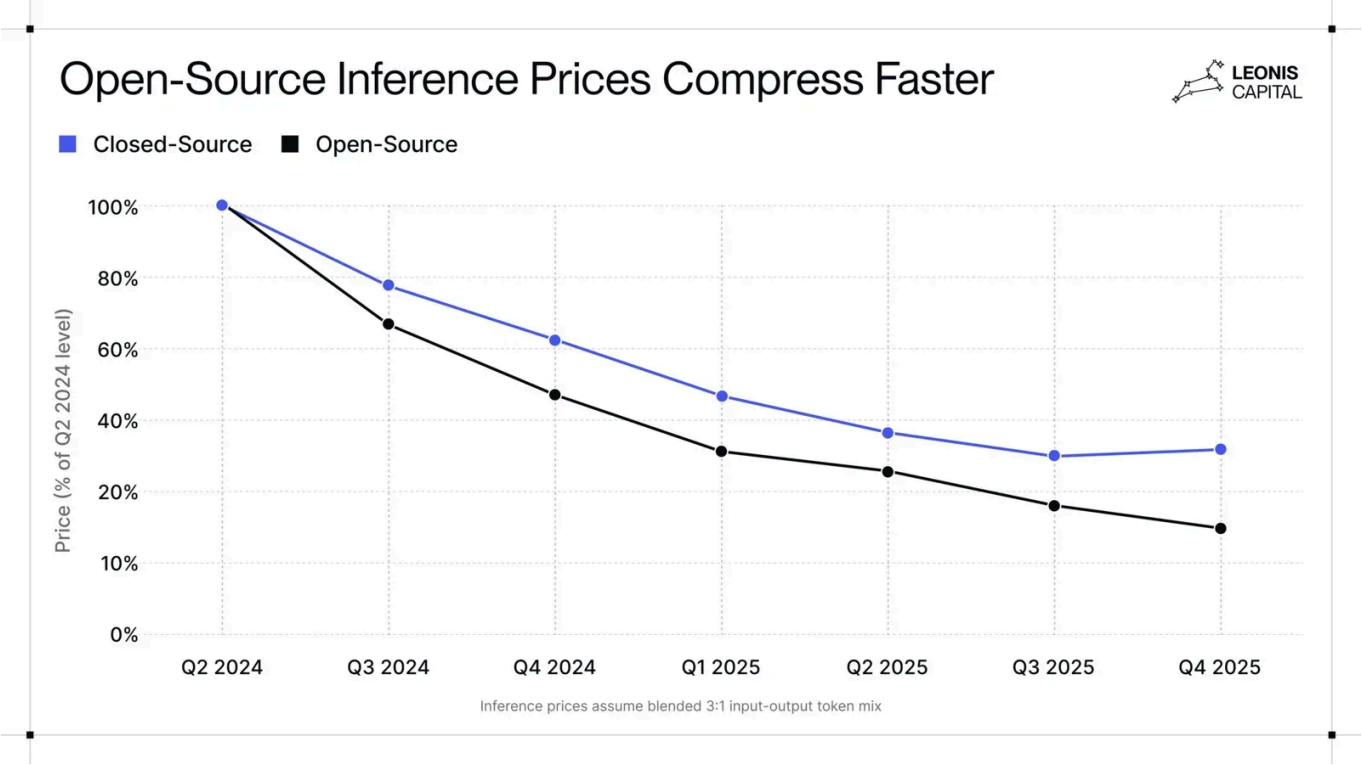
Indicator: By the end of 2026, at least one leading foundation model lab will release a frontier-tier capability that is available only through a first-party product and not via its public API for at least six months.

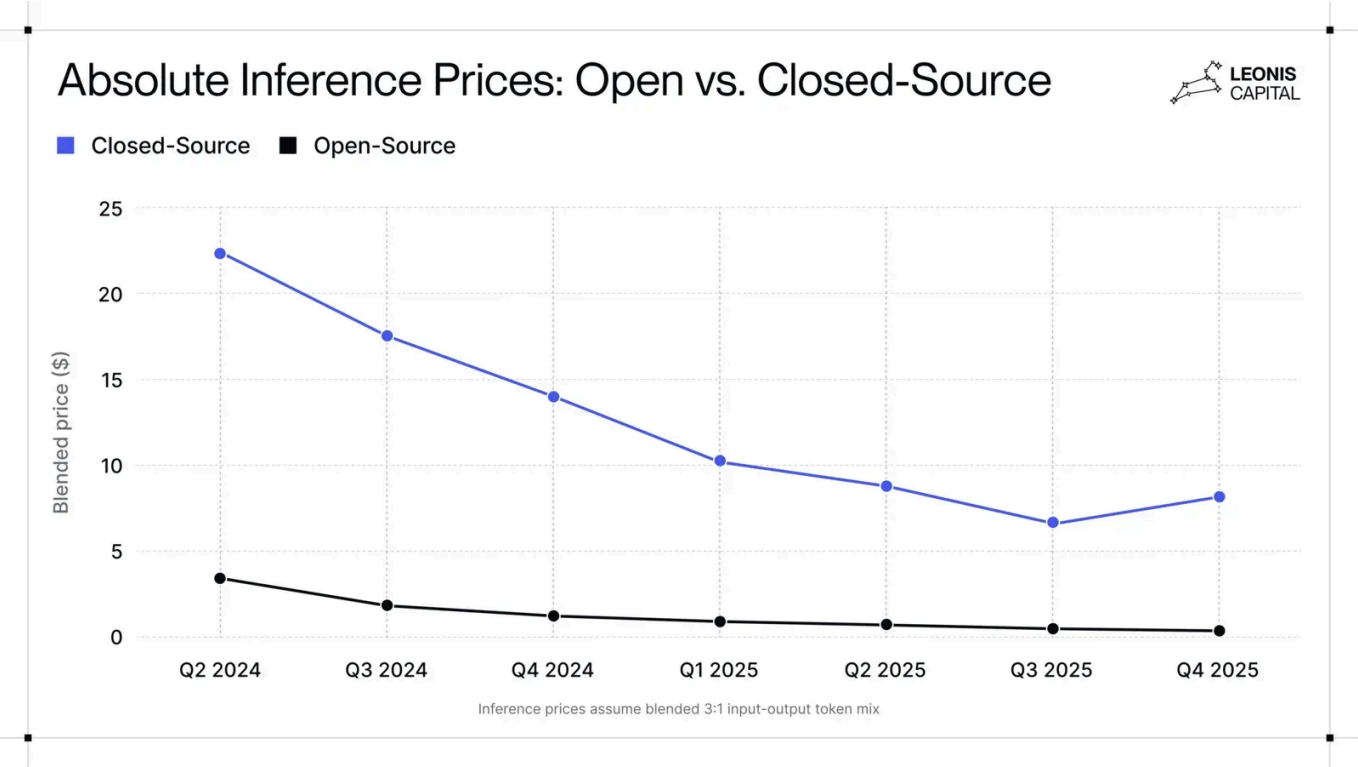
We expect 2026 to mark a quiet but consequential shift in how leading foundation model labs allocate attention and capital. As open-source models narrow the performance gap, model APIs will lose both differentiation and pricing power. **First-party products, by contrast, offer structurally higher margins *and* more user data to improve the models.** A rational move for foundation model labs is to de-emphasize their APIs and increasingly reserve frontier capabilities for their own products.

As open-source models become "good enough" for the majority of enterprise and developer use cases, customers increasingly treat models as interchangeable. This means that **APIs stop being**

sticky or defensible, eroding their pricing power. This shift is already visible in enterprise behavior: a recent survey by the Linux Foundation found that among organizations that have adopted AI, nearly 90% incorporate open-source models somewhere in their stack.

At the same time, cost dynamics increasingly favor open-source. Our analysis finds that while both open-source and closed-source models have become meaningfully cheaper over the past two years, **open-source prices have fallen far faster: approximately 6.7x versus 2.8x for closed models**. This widens the effective price differential between the two and indicates future margin compression for closed-source models.





Source: Leonis Capital research

This pushes foundation model APIs toward **infrastructure-like economics**: high volume, low margin, and intense price competition. The marginal value of being slightly better on benchmarks keeps falling, while inference costs and open-source pressure cap upside. The consequence is not just margin compression but valuation fragility: **an API-heavy revenue mix actively undermines frontier model valuations**. [2]

As model differentiation collapses, a strategic tension becomes unavoidable: large API ecosystems give rise to downstream companies that own the user. Developer tools, vertical copilots, and agent platforms increasingly control workflows, switching costs, and customer loyalty, while treating models as interchangeable backends. As of Q3 2025, Cursor and GitHub Copilot together accounted for nearly half of Claude's coding-related API revenue. As these large customers multi-home and bargain aggressively, **APIs begin to look less like platforms and more like subsidies for future competitors**.

In addition to margin loss, there is also a risk of **learning decay**. An API-first distribution model deprives model providers of the task-level feedback needed to improve reasoning and long-horizon reliability. At a moment when high-quality RL data is scarce and expensive (and labs are spending hundreds of millions to acquire it), this loss of signal is counterproductive. In that context, prioritizing first-party products is the best way to tighten the learning loop.

These pressures (margin compression, downstream competition, and a tightening learning bottleneck) are already shaping lab behavior. OpenAI's internal "code red" increasingly

prioritizes the ChatGPT experience as part of a broader push toward owning end-user surfaces. This trajectory aligns with our deep dive on how OpenAI is becoming an "everything platform," in which product distribution, and not standalone APIs or model performance, becomes the primary locus of value capture. Anthropic has taken a similar but more gradual path. The launch of Claude Code signals a deliberate move to internalize coding workflows, and Claude CoWork extends that capability into the rest of knowledge work. While Anthropic remains more API-dependent today and is therefore likely to transition more slowly, the directional incentives are similar.

Taken together, these forces imply a structural shift rather than an accidental outcome. By late 2026, we expect leading foundation model labs to cap or slow API growth, reserve frontier capabilities for first-party products, and deprioritize broad developer ecosystems as a primary growth vector. API usage may continue to grow, but its strategic importance will diminish. **Value capture, learning, and differentiation will concentrate in first-party products.**

Prediction 3: Outcome-based pricing finally takes off in domains where AI replaces human labor

Indicator: At least one AI startup with ≥$50M ARR will publicly confirm they derive the majority of revenue from outcome-based or success-fee pricing structures, rather than consumption or seat-based pricing.

We first argued for the rise of outcome-based pricing in early 2025, but didn't pinpoint a specific timeline. The idea has then been discussed by many in the industry, but 2026 is likely the point at which it meaningfully takes off, not because enterprises suddenly prefer it, but because **the alternative stops working**.

The first enabling condition is straightforward: **model performance has crossed a threshold where AI can reliably replace human labor in bounded domains**. AI is now good enough, cheap enough, and consistent enough to take responsibility for discrete units of work: tickets resolved, documents processed, claims adjudicated, reports generated. Without this baseline reliability, outcome-based pricing is impossible.

The more important drivers, however, are structural rather than technical:

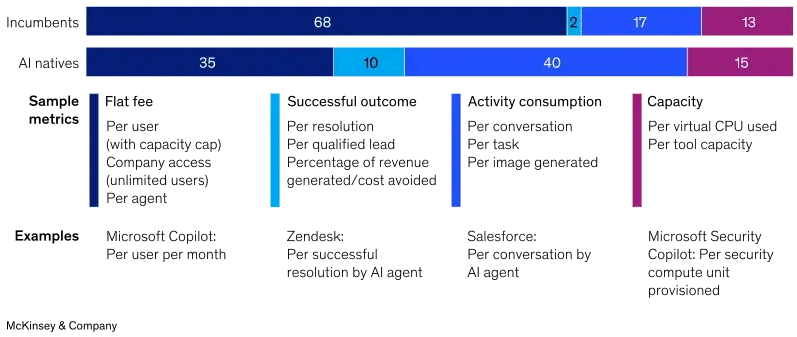
- **Enterprise buyers increasingly do not want to buy new software at all.** Large organizations already have entrenched vendors, long contracts, and internal inertia. Even when AI-native tools are meaningfully better, switching costs (organizational, political, and operational) are often higher than the perceived upside of incremental productivity gains.
- **Generic AI SaaS is being compressed from both sides:** by traditional SaaS incumbents with distribution moats bundling AI features on one end, and by highly customized, service-heavy solutions on the other.

In this environment, **outcome-based pricing becomes the only viable wedge** because it reframes the purchase not as software adoption, but as direct replacements for headcount, outsourcing, or agency spend.

This structural pressure is already visible in pricing behavior. A McKinsey report from November 2025 found that roughly 10% of AI-native companies have adopted outcome-based pricing. While still a minority relative to traditional pricing models, outcome-based pricing is already **five times more prevalent among AI-native companies than among traditional software firms**. This indicates that outcome-based pricing is more aligned with AI-native products.

In the emerging, AI-driven software era, providers have a wide range of consumption pricing metrics to build their businesses around.

Software pricing meter type and sample metrics, by company type, %



McKinsey & Company

We are already seeing **early adoption among AI-native startups**. Sierra ($100M ARR) promotes outcome-based per-resolution pricing but operates a "blended model" that combines it with volume-based pricing for routine tasks. CEO Bret Taylor has called outcome-based pricing "the future of software," though the company hasn't disclosed what percentage of revenue comes from each model. Fellow customer support platform Decagon has also scaled to ~$35M ARR on a combination of outcome-based and volume-based pricing. In debt collections and brokering, AI agents already operate on de facto outcome-based models: providers take a percentage of what's recovered, with minimal upfront fees from customers.

Even traditional software companies are beginning to experiment, though more cautiously. Intercom (~$343M ARR) now charges 99 cents per AI resolution on top of its seat-based fees. While the company still derives the majority of its revenue from traditional seats, it is able to adopt outcome-based pricing because it operates in a domain where outcomes are measurable and attributable (tickets are either resolved or not). Intercom's trajectory is also instructive. Fin was originally priced at $1.90 per resolution but dropped to 99 cents after customers pushed back, arguing that the price wasn't competitive with outsourced human agents. This suggests that **buyers already benchmark AI against human labor costs in fields where the AI can own the outcome**.

Most traditional SaaS products lack this structural advantage. The vast majority of traditional enterprise SaaS products are productivity boosters whose marginal contribution is difficult to

isolate. Unless they meaningfully update their products or business model, we expect traditional SaaS to struggle to adopt outcome-based pricing in any meaningful way.

Outcome-based pricing requires outcome ownership. You can't price outcomes unless you can reliably execute end-to-end in what we call "**Systems of Action**," which requires much deeper integration with customer workflows, like embedding into systems of record, handling edge cases, absorbing variability, and often co-designing processes with their customers. This means longer onboarding cycles, heavier integration work, and more forward-deployed engineering, but the payoff justifies it: fewer customers, significantly higher revenue per customer, and far lower churn once embedded. For customers, this dynamic forces AI vendors to have skin in the game and aligns them toward products that actually work.

As a result, in 2026, we expect enterprise AI companies to look less like traditional SaaS and more like **hybrids of software, services, and embedded operators**. They face their own set of challenges, such as negotiating outcome definitions with their customers and absorbing operational complexity. Moreover, they will need to manage gross margin pressures because their revenue is contingent on outcomes, but their costs are tied to usage. The winners will not be those with the cleanest product metrics or fastest time-to-value, but those willing to take on variability and risk in exchange for durable, outcome-linked revenue.

What 2026 Means for Founders and Investors

For founders, on the technical side, inference-time systems may become the biggest performance booster, perhaps even more so than proprietary data. The boring infrastructure (context, retries, state management, etc) is now a competitive moat, not technical debt. On the business side, if you're in labor replacement domains, outcome-based pricing will become table stakes. The challenge lies in how to adopt it. As startups become embedded operators and "systems of action," forward-deployed engineering should become a core capability, not a temporary GTM tactic.

For investors, first-party products from foundation labs will increasingly compete with your portfolio companies, especially as they reserve frontier capabilities for themselves. This forces an uncomfortable question: Are your companies building defensible businesses or subsidizing their own future competitors? Short-term growth can obscure whether a company survives once foundation model players enter the arena.

Taken together, the common thread in our advice is to **position for 2027-2028 outcomes, not 2026 narratives.**

FOOTNOTES

1. For example, the WeirdML benchmark shows that the gap between open and closed models is persistent over time. Despite the recent strides that DeepSeek and Qwen have made, the performance gap could be widening on complex tasks as closed-source models have accelerated at a steeper rate.
2. We talked about this extensively in a prior article. API-centric model businesses are structurally fragile: once a top open-source model matches or exceeds their performance, pricing power can disappear rapidly, driving the economic value of the closed-source model toward a “zero-value threshold.”