BAIN & COMPANY

Technology Report 2026

AI—the Most Consequential Technology of Our Lifetimes

Authors and acknowledgments

David Crawford, chairman of Bain & Company's Global Technology, Media, and Telecommunications (TMT) practice; Anne Hoecker, leader of the Global TMT practice; and Dana Aulanier, vice president of the TMT practice, prepared this report.

Bain Partners Cory Boles, Padraic Brick, Mark Brinda, Alessandro Cannarsi, Thibaud Chabrelie, Aaron Denman, Greg Fiore, Frank Ford, Jonathan Frick, Arun Ganti, Peter Hanbury, Jonny Holliday, Alec Koh, David Lipman, Stanley Liu, Eric Maltiel, Justin Murphy, Sandeep Nayak, Christopher Perry, Bill Radzevych, Josh Sandberg, Michael Schallehn, and Jue Wang; Associate Partners Tatum Quinn, Kristie Tagawa, and Vito Trinchera; Expert Partners Syed Ali, Mukarram Bhaiji, Purna Doddapaneni, and Chris McLaughlin; and Practice Manager Sushant Bansal wrote its chapters.

The authors also wish to thank Partners Peter Bowen, Ann Bosche, Arjun Dutt, Tamara Lewis; Associate Partner Garima Nijhawan; Bain Capability Network (BCN) Director Eva Gupta; BCN Manager Vaishali Sharma; BCN Senior Manager Arpit Jain; BCN Project Leaders Anjali Mishra and Venkatesh Mishra; BCN Associates Sahil Garg, Aryan Gupta, and Tejal Janbandhu; BCN Analysts Praapti Bansal, Kankshit Kumar, Manya Singh, and Tejas vardhan T; Practice Directors Tarun Gupta and Alexander Smyth; Operations Manager Raghav Sharma; and Practice Editors Jeff Bauter Engel, Adam Jones, and David Sims for their editorial support.

This work is based on secondary market research, analysis of financial information available or provided to Bain & Company and a range of interviews with industry participants. Bain & Company has not independently verified any such information provided or available to Bain and makes no representation or warranty, express or implied, that such information is accurate or complete. Projected market and financial information, analyses and conclusions contained herein are based on the information described above and on Bain & Company's judgment, and should not be construed as definitive forecasts or guarantees of future performance or results. The information and analysis herein does not constitute advice of any kind, is not intended to be used for investment purposes, and neither Bain & Company nor any of its subsidiaries or their respective officers, directors, shareholders, employees or agents accept any responsibility or liability with respect to the use of or reliance on any information or analysis contained in this document. This work is copyright Bain & Company and may not be published, transmitted, broadcast, copied, reproduced or reprinted in whole or in part without the explicit written permission of Bain & Company.

Net Promoter®, NPS®, NPS Prism®, Net Promoter System®, and the NPS-related emoticons are registered trademarks of Bain & Company, Inc., NICE Systems, Inc., and Fred Reichheld. Net Promoter Score℠ is a service mark of Bain & Company, Inc., NICE Systems, Inc., and Fred Reichheld.

Copyright © 2026 Bain & Company, Inc. All rights reserved.

Contents

AI—the Most Consequential Technology of Our Lifetimes 2

Value Evolution 4

New Innovation Is Required to Fund AI's $6 Trillion Buildout 5

Software Investing in the Age of AI and Slower Growth 9

Strategic Battlegrounds 15

AI Data Center Boom: Can We Build It If They Come? 16

$100-Billion SaaS Opportunity Hiding in Cross-System Labor 23

Hardware Strikes Back in the AI Era 32

Tech Services: Cracking the Code on AI-led Growth 39

Cybersecurity's New AI Imperative: Attacking the Backlog of Vulnerability Alerts 44

Operational Transformations 49

The AI-Native Enterprise: Absorption Is the New Advantage 50

The Half-Finished Redesign: How AI Reshapes Software Organizations 58

Managing Token Spending without Choking Off Opportunity 62

The Missing Architecture for Agentic Software Development 67

How AI Is Changing Data Monetization 73

AI—the Most Consequential Technology of Our Lifetimes

No technology has ever disrupted as quickly or as broadly as AI has. The market capitalization of the world's 10 most valuable public companies far outpaces the rest of the market, and 9 of them are focused on AI.

Last year's agenda (where's the ROI, where should we pilot) has given way to more difficult challenges: How will we fund the buildout, integrate AI more rapidly, and preserve IP? And how will industry structure and regulation evolve to facilitate innovation, competitiveness, and safety?

So, AI must do more than boost labor productivity on existing tasks; it needs to unlock new sources of growth and value. Superintelligence should be used for superhuman endeavors, not just better execution of existing, mundane work.

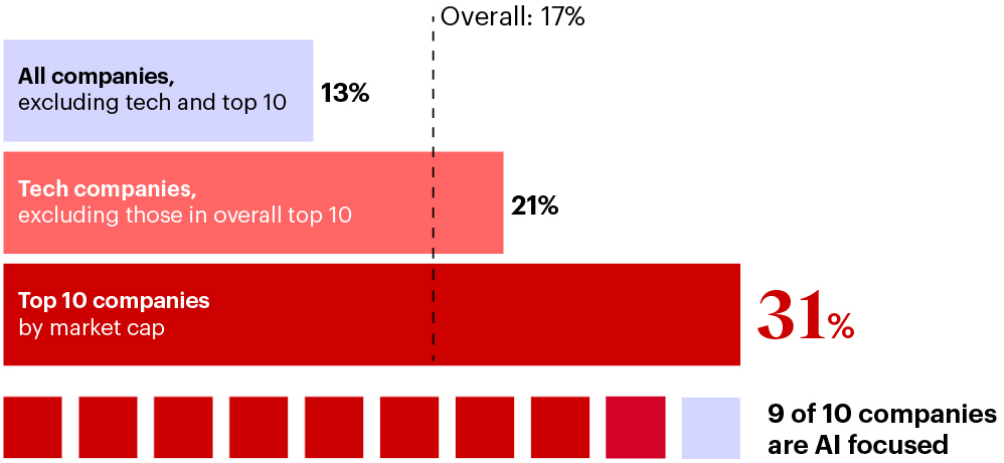
Productivity and growth depend on how quickly existing organizations can absorb AI and new AI-native organizations can grow. It's no longer only about access to the best technological capabilities; it's about who can put AI to work fastest and most effectively.

As technology absorption rates emerge as a strategic variable, enterprise leaders are committing to deep, lasting change. The technology stack is being rebuilt around AI at its core. Opex is migrating from headcount to token spend. The organizational pyramid is flattening as individual contributors take on something closer to a managerial role, directing agents rather than executing tasks themselves.

The enterprise and global economy are being reengineered from the ground up. This report maps the work.

David Crawford
Chairman of Bain's Global Technology, Media, and Telecommunications practice

2024–2026 CAGR for global public companies



Note: Market cap data as of January 1st , top 10 companies taken as per 2026
Source: CapIQ

Value Evolution

New Innovation Is Required to Fund AI's $6 Trillion Buildout. 5

Software Investing in the Age of AI and Slower Growth. 9

VALUE EVOLUTION

New Innovation Is Required to Fund AI's $6 Trillion Buildout

To justify the investment, AI must do more than boost productivity; it will need to unlock new sources of growth and value.

By David Crawford, Kristie Tagawa, Cory Boles, and Tatum Quinn

At a Glance

- AI infrastructure investment is racing ahead, but generating enough economic value to justify it will require trillions of dollars in new AI-driven revenue.
- Productivity gains from existing enterprise and consumer applications won't be enough; entirely new markets must emerge to close the funding gap.
- The winners will be those that create breakthrough AI applications that transform industries and expand the global economy.

The unprecedented speed and scale of the AI buildout, with billions flowing into chips, data centers, networks, and power systems, have focused attention on the challenge of building capacity. But the more important question may be whether enough economic value can be created to justify it.

Consider the scale of investment and the gap between that and the revenue model necessary to fund it.

- The arms race among hyperscalers (Microsoft, Google, Amazon, Meta, and Oracle) is accelerating: Their capital expenditures could reach $780 billion in 2026, nearly five times the level of just three years earlier.
- Leading-edge AI data centers today are approaching 1 gigawatt (GW) of power capacity. By 2027, many are expected to approach 2 GW facilities, with 9 GW campuses emerging by the end of the decade *(see Figure 1)*.

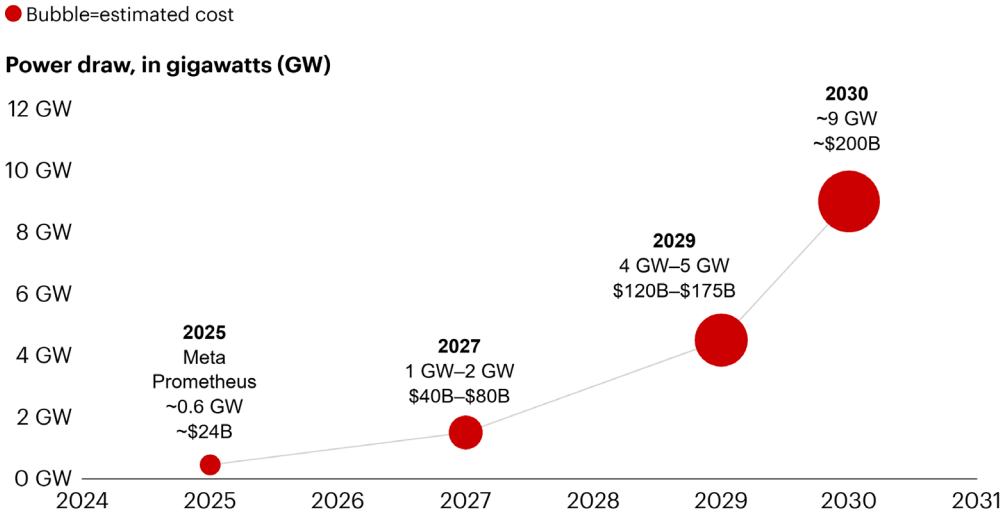
By 2031, annual spending on AI infrastructure could reach $1.5 trillion, including new data center infrastructure and compute capacity as well as ongoing upgrades to the installed base of GPUs, memory, and networking equipment.

If we assume that capital expenditures amount to about 25% of industry revenue (an ambitious but reasonable percentage based on trends among cloud providers), sustaining this level of investment would require an AI market approaching $6 trillion annually.

Some of that revenue is already coming into focus. Consumer AI products, through subscriptions and advertising, could generate an estimated $200 billion to $400 billion by 2031. Enterprise adoption could contribute another $1 trillion to $1.4 trillion in gains to providers alone as AI delivers meaningful productivity gains to enterprises across software development, sales, marketing, customer service, and IT operations.

Figure 1: The size and cost of AI data centers is accelerating, doubling approximately every 12 to 16 months

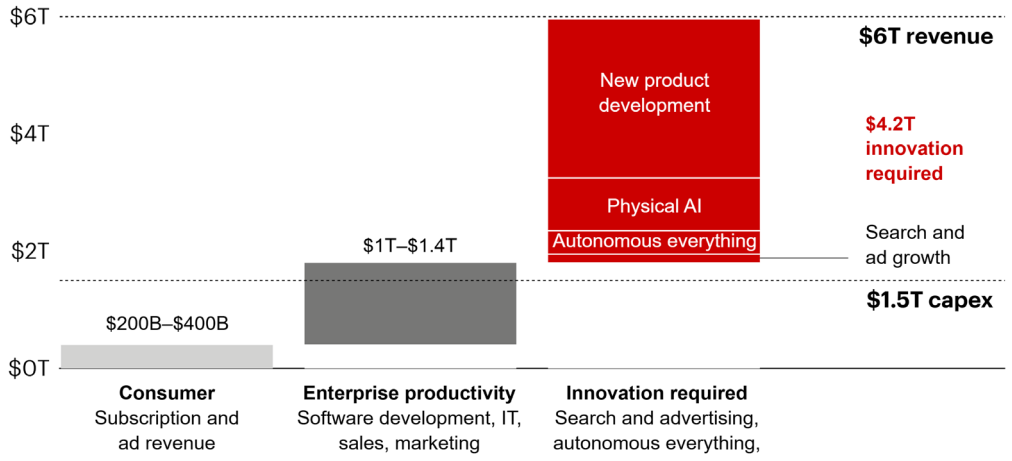
Power capacity and estimated cost of the largest single AI data center



Note: Each data point is an independent estimate for the single largest AI data center at that time; power and cost shown as midpoints of published ranges; 2027, 2029, and 2030 values extrapolated from a doubling trend
Source: Epoch AI

Figure 2: Four categories could supply revenue to fund AI's global market by 2031

Global market for AI, in trillions of US dollars



Sources: Bain & Company; Nvidia

Together, the consumer and enterprise AI market could total between $1.2 trillion and $1.8 trillion, leaving about $4.2 trillion of new revenue to reach the $6 trillion market that we estimate will be necessary to fund the buildout *(see Figure 2)*.

That revenue must come from new sources of economic value.

Sources of new value

Dramatic innovation will be required to deliver the revenue necessary to fund the gap. Bain's research finds four key categories that are likely to help deliver this growth.

- **Search and advertising ($100 billion to $200 billion):** By integrating ads into their chatbot products and encouraging the trend of using AI to replace a large portion of traditional Internet search, model providers could unlock $100 billion to $200 billion or more in revenue.
- **Autonomous everything ($400 billion):** Using AI to autonomously operate automobiles, trucks, and drones, as well as other industrial automation initiatives, represent a $400 billion market opportunity by increasing equipment uptime while reducing training and operating costs. Autonomous vehicles could provide significant value as new cars and trucks for consumers, as robotaxi services, and by automating logistics services.

- **Physical AI ($900 billion):** Advanced AI models can enable highly realistic simulations and digital twins of physical processes, helping companies improve productivity, test modifications, and accelerate the deployment of autonomous systems. Similarly, AI-powered robotics, including humanoid robots, can operate in unstructured environments, unlocking a wide range of new applications across the physical economy, from manufacturing to surgery. As a result, the physical economy could represent a $900 billion opportunity across priority sectors (including automotive, electronics, semiconductors, and aerospace and defense), assuming a 10% reduction in R&D and manufacturing costs from higher yields and faster factory ramps.

- **New product development:** Beyond AI's uses today, new applications could help close the gap. These could include AI-driven drug discovery that makes treatments for rare diseases economically viable, always-available mental health support that addresses billions of dollars in unmet demand, materials science breakthroughs that unlock next-generation batteries and semiconductors, and autonomous scientific research that accelerates progress in fields from neuroscience to fusion energy—all examples of new opportunities resulting from abundant intelligence.

Innovation and entrepreneurship must accelerate

Enterprise productivity is the tip of the spear, the first gains we're seeing from AI deployment, but it won't be nearly enough. The economics required to generate ROI from AI infrastructure are demanding trillions in new revenue, not just cost savings.

The economics required to generate ROI from AI infrastructure are demanding trillions in new revenue, not just cost savings.

The industry needs a wave of application innovation comparable with what mobile and cloud unlocked, not just productivity gains on existing workflows.

The infrastructure is being built ahead of the demand curve, and funding it sustainably will require adding approximately 1% to the annual global GDP growth rate. The question is whether the applications arrive in time to pay for it.

VALUE EVOLUTION

Software Investing in the Age of AI and Slower Growth

As AI reshapes software economics and growth slows, private equity firms need a new playbook to underwrite risk, measure performance, and create value.

By David Lipman, Christopher Perry, Jonny Holliday, and Thibaud Chabrelié

At a Glance

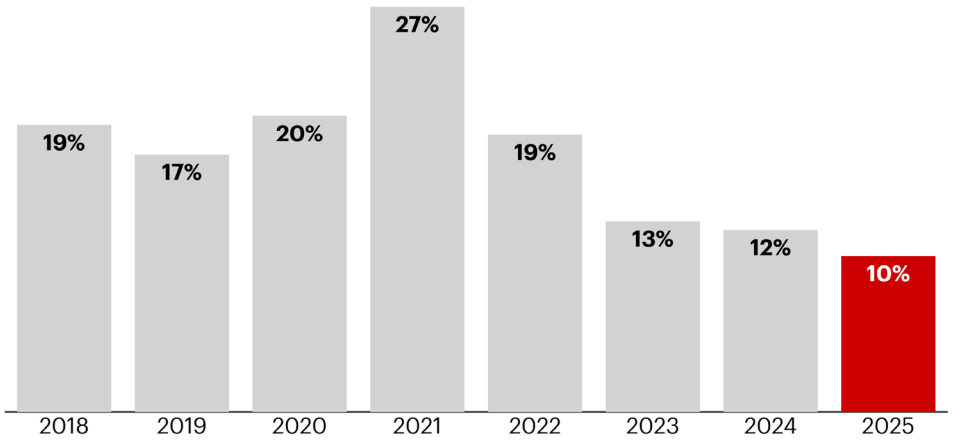
- As evidence builds that software is losing some of the structural advantages that made it a preferred investment, portfolio companies must adopt new strategies for navigating through the uncertainty.
- Diligence must adapt to consider not only the market risks posed by AI-native insurgents but also the costs for incumbents to raise their AI game.
- AI transformation is an imperative for every company in the portfolio, but the broader question is how AI should change the product offering.

If software investing has delivered anything to private equity investors so far in 2026, it is the disquieting certainty that the world has changed—probably forever.

Revenue growth that was running around 20% annually is now trending at half that. Net revenue retention (NRR) has dropped about 9 points since 2021 *(see Figure 1)*. Dealmaking is at a crawl. Aging portfolios are suddenly a thing in tech investing. And looming over everything else is the specter of artificial intelligence, which, at best, threatens the once-unassailable SaaS value proposition and, at worst, raises fears that some software use cases are veering toward obsolescence.

Figure 1a: Revenue growth has been cut in half in recent years ...

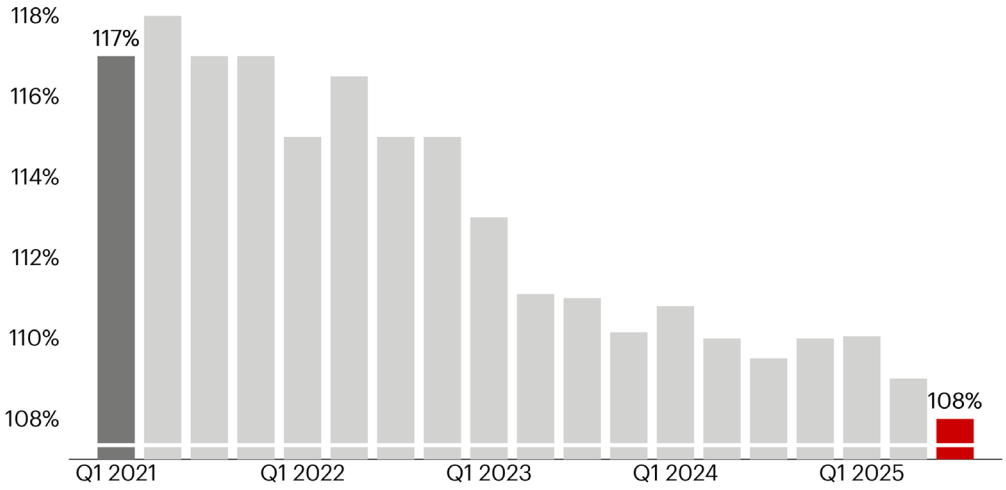
Median year-over-year revenue growth for US public software companies



Notes: 2025 data as of February 2, 2026; last 12 months (LTM) revenue used for 2025 if Q4 reporting unreleased; different company sets per year; includes firms with Software (Primary) as industry classification and more than $50 million in LTM revenue in the given year
Source: S&P Market Intelligence

Figure 1b: ... and net revenue retention is in retreat

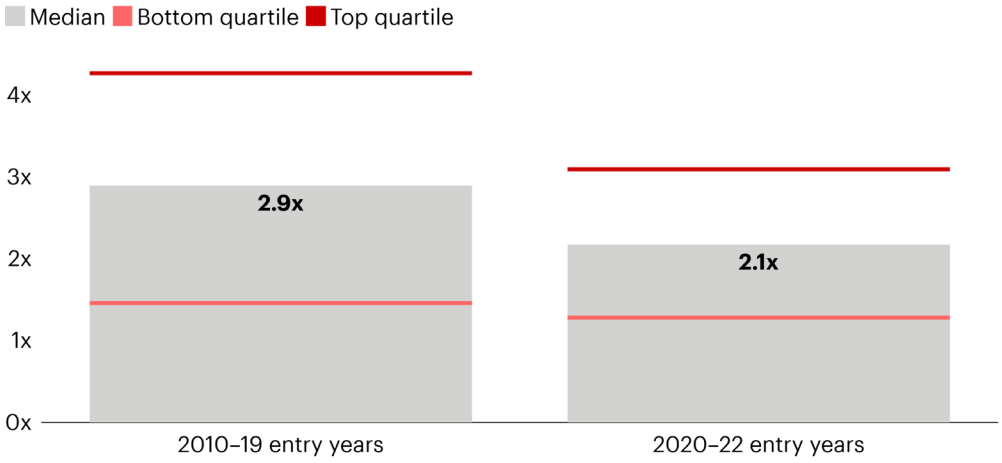
Median net revenue retention for US public enterprise SaaS companies



Notes: n=85; excludes any company/quarter combinations without available data
Source: AlphaSense

Figure 2: It will be challenging for postpandemic tech buyouts to achieve returns in line with the prior decade, given high prices paid and less-certain growth

Gross deal multiple on invested capital for tech buyouts in North America



Notes: Includes fully realized deals and deals that have realized at least 25% of their returns; includes all deal sizes; all figures in USD
Source: SPI by StepStone (April 2026)

Perhaps most ominous is the building evidence that software is losing some of the predictability that made it a preferred asset class for the buyout community, including reliable margin profiles, common expansion mechanics, and similar diligence frameworks. The software assets that are transacting in the private markets continue to attract high prices, largely because sponsors have focused their efforts on selling their A-plus companies. Yet deals completed after the Covid-19 pandemic, including more seasoned investments from 2020 to 2022, are, to date, generating returns below pre-Covid averages *(see Figure 2)*. While many of these investments are not yet fully realized, and performance could improve over time, that will be a challenge given the high prices sponsors paid for assets and the sector's slowing growth.

What comes next?

Patterns are emerging that suggest a number of practical actions private equity investors can take right now to keep pressing forward. We already know enough about AI's disruptive power, in fact, to start making no-regrets shifts in every phase of the PE value proposition. Proactive general partners (GPs) are taking steps to:

- pragmatically reorient due diligence to more reliably underwrite AI risks and opportunities;
- refocus value-creation playbooks to zero in on how AI technology can transform offerings through a deeper understanding of customer workflows; and

- demonstrate bankable progress by marshaling data and metrics that can provide the next owner with compelling AI proof points, not just rosy narratives or random prototypes.

The new due diligence imperative

For more than a decade, software value creation has centered on maximizing predictable, recurring revenue. SaaS companies grew by adding seats, adding premium features, and cross-selling products. High margins and sticky customer relationships—reinforced by embedded workflows, user habits, and accumulated data—created durable competitive moats.

AI is changing all that. Beyond enabling AI-native challengers to leapfrog established software providers, it fundamentally changes software economics, eroding some of the industry's structural advantages. The compute required to power AI adds substantial costs to every query, shifting revenue models from seat-based subscriptions toward usage and outcomes. Margins are lower, and future growth is harder to predict.

That shift makes traditional SaaS metrics less reliable in due diligence. Annual recurring revenue (ARR), NRR, and revenue multiples were reliable proxies for value when software moats were durable and margins consistently high. Today, as the link between growth and valuation weakens, investors need new ways to assess which companies can create lasting value in an AI-driven market.

Diligence in this new environment has to start with an assessment of two things: How much can AI impact the user workflows the software supports, and how much risk is there that AI could displace the software altogether within those workflows? Stress-testing requires specificity, since product moats, workflow moats, and data moats each come under different degrees of pressure from AI.

Strong diligence will also capture upside: How AI-ready is the company itself? Is it rapidly deploying AI internally to improve efficiency? Is it gaining traction with AI on the product side, either by adding features or by launching new products that customers value? The key here is evidence: The market is already bifurcating between companies that can demonstrate measurable AI traction and those that are still spinning a narrative without numbers behind it. In diligence, the key question is no longer "do they have an AI strategy?" It's "can they show me the proof points?"

AI transformation during ownership

Using AI internally to improve efficiency and transform workflows is rapidly becoming table stakes for any company in your portfolio, software or otherwise. The opportunity to truly inflect performance still involves some experimentation and faith in the technology, but standing still isn't an option.

The more complex question is how to reshape the product roadmap to generate AI-based revenue—and how to scale meaningful innovation at speed. That can be a tall order for an incumbent software company accustomed to optimizing for a traditional seat-based platform, especially given all the daily requirements of managing the core business.

The companies seeing the most success tend to have a light-bulb moment when they recognize that simply helping humans do tasks faster with incremental product enhancements is increasingly missing the point in an AI world. If an agentic system, for instance, can actually manage a workflow end to end, the goal should be enabling a step change in measurable outcomes, not just user efficiency.

That means turning the traditional product development approach on its head. Instead of focusing solely on how customers use discrete products within workflows, the most innovative companies are looking to connect the dots across the entire process. They are going deep on the customer's broad objective and how their solution can be rebuilt to achieve that outcome—sometimes freeing up workers for more productive pursuits, sometimes replacing them altogether.

The lesson here is that AI is not a game of incrementalism. Capturing the opportunity starts with zero-basing product assumptions and reimagining what's possible.

The lesson here is that AI is not a game of incrementalism. Capturing the opportunity starts with zero-basing product assumptions and reimagining what's possible. The answer won't always be a complete rebuild. But you need to decide how you can take customer outcomes to a new level by matching AI technology to a deeper understanding of the relevant workflows.

And you'll need to assess what that will require in terms of adding talent and making organizational changes to support rapid execution.

Proving it

It's no surprise that GPs and their portfolio company management teams are expending massive bandwidth to develop new AI features and products. But they're probably not spending enough energy to develop the means to track progress and measure impact with concrete data. That helps explain the wide gap in value expectations we're often seeing between buyers and sellers.

Because the traditional SaaS KPI stack (ARR, NRR, and gross margin) was built for a different value proposition (a world of near-zero marginal cost, seat-driven expansion, and predictable retention), it fails to capture AI impact reliably. ARR, for instance, is a great measure of the predictable revenue derived from subscription seats. But AI is priced on actual usage and outcomes, which can be bursty and unreliable.

Understanding the true performance of these products requires separately tracking at least three distinct revenue buckets: traditional AI and machine learning (predictive models, risk scores), AI add-ons

(copilots, assistant features), and agentic products (workflow automation, autonomous execution). Each has a distinct margin profile, growth trajectory, and competitive dynamics.

The cost side, too, is very different. AI products have significant variable costs per use that don't exist for traditional SaaS solutions. A full understanding of what you're spending requires tracking hosting and infrastructure costs, third-party model costs, and the fully loaded cost of each employee devoted in full or in part to AI-related R&D.

The bottom line is that if you don't change your metrics, you can't evaluate AI impact. The revenues and costs directly attributable to AI first need to be separated from the core business and then broken down into their constituent parts. Without that, it's impossible to answer with precision the three essential questions on everyone's mind: Is AI driving incremental revenue? Is AI changing cost structures? And are AI-related products scaling efficiently?

Armed with firm answers, portfolio companies can design the best product roadmap and allocate resources accordingly, and GPs can start building the kind of evidence-based exit story buyers are demanding.

Twenty twenty-six will likely be remembered as the year AI truly redefined the software industry—the kind of no-turning-back moment that challenges all previous assumptions. Nothing about that is easy. But the chaos won't last forever. The leaders coming out of this transformation are already hard at work rethinking how to underwrite risk, inflect portfolio company performance, and measure results in a world upended by AI.

One thing is clear: There's no time to waste.

Strategic Battlegrounds

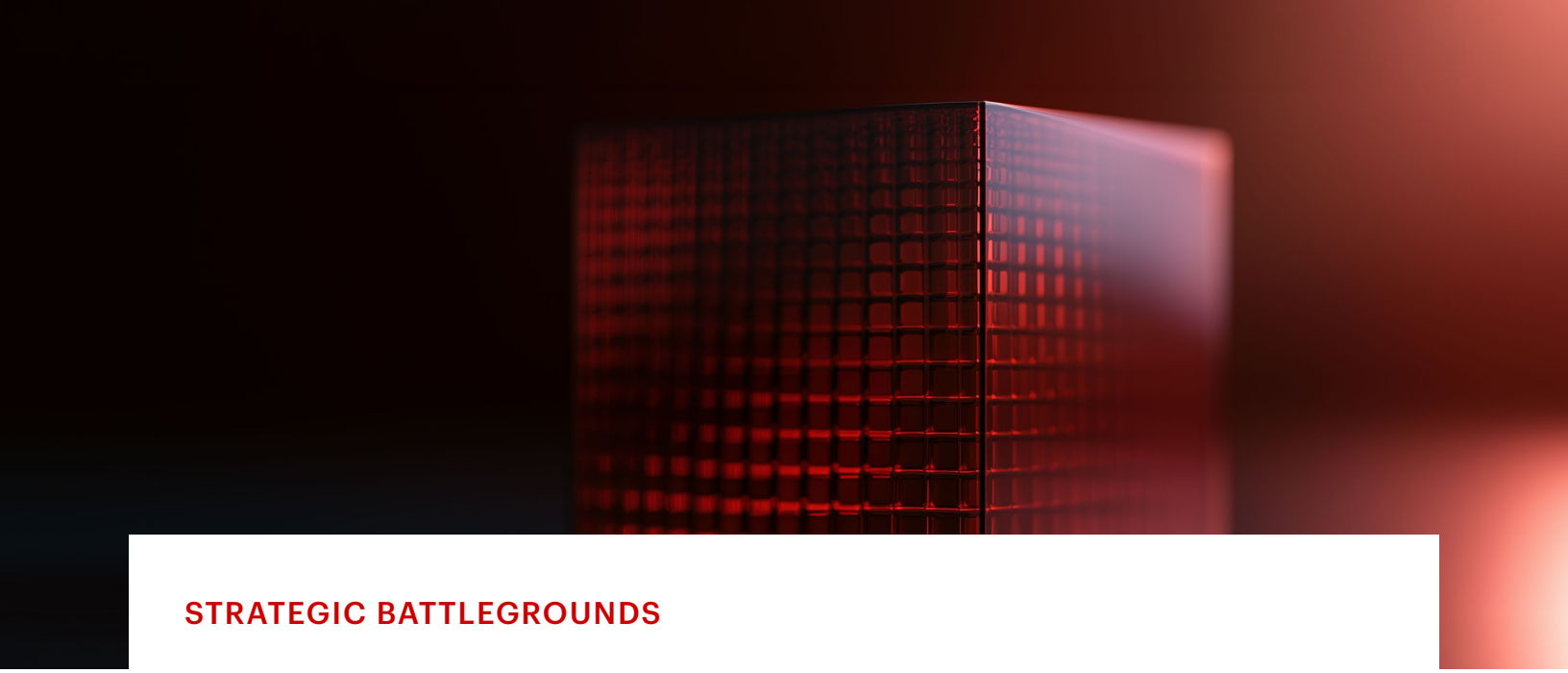
AI Data Center Boom: Can We Build It If They Come? 16

$100-Billion SaaS Opportunity Hiding in Cross-System Labor . . . 23

Hardware Strikes Back in the AI Era 32

Tech Services: Cracking the Code on AI-led Growth 39

Cybersecurity's New AI Imperative: Attacking the Backlog of Vulnerability Alerts 44



STRATEGIC BATTLEGROUNDS

AI Data Center Boom: Can We Build It If They Come?

Meeting AI compute demand will require system-level solutions, not the site-by-site workarounds plugging gaps today.

By Tatum Quinn, Peter Hanbury, David Crawford, Aaron Denman, Padraic Brick, and Alessandro Cannarsi

At a Glance

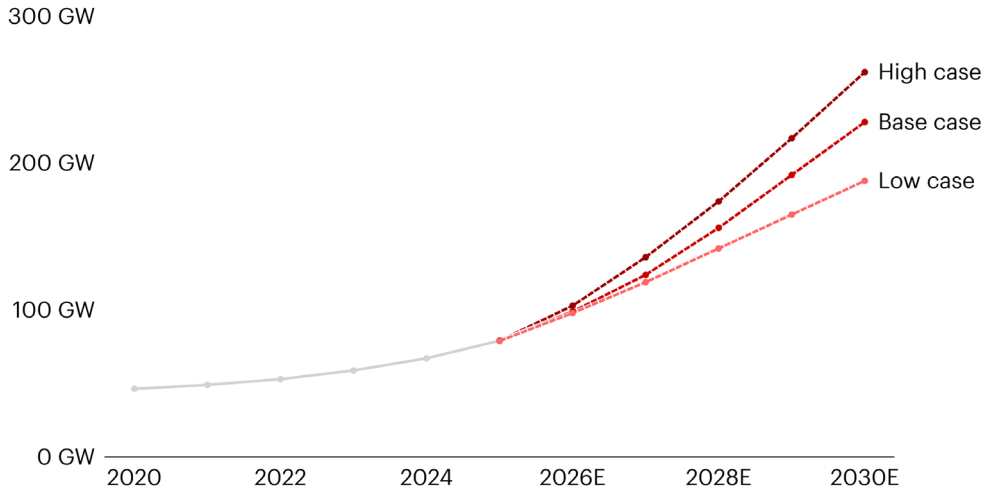
- Data centers have become critical infrastructure; Bain projects $5 trillion to $6.5 trillion of buildout, adding about 150 gigawatts or more by 2030.
- That would nearly triple global capacity over five years, but power, chips, skilled labor, and permitting are all constrained at once.
- Decisions made today set capacity several years from now, and the shortage won't resolve on its own.
- Closing the gap will take structural moves (new power and technology, public-private coordination, resource sharing), reshaping the industry.

Data centers powering AI have rapidly become a new layer of critical infrastructure worldwide that are central to technology innovation, economic growth, and national sovereignty.

US capital investment in data centers as a share of GDP now rivals what it spent building its railroads, telecom networks, and electrical grid: generational projects. The data center boom is moving faster.

Figure 1: Data center critical IT capacity is expected to grow by 20% to 30% over the next five years

Global data center installed base of IT power, in gigawatts (GW)



Notes: Reflects net growth in installed base; excludes GPU refresh, retrofit, and replacement of retired capacity
Source: Bain Data Center Model, July 2026

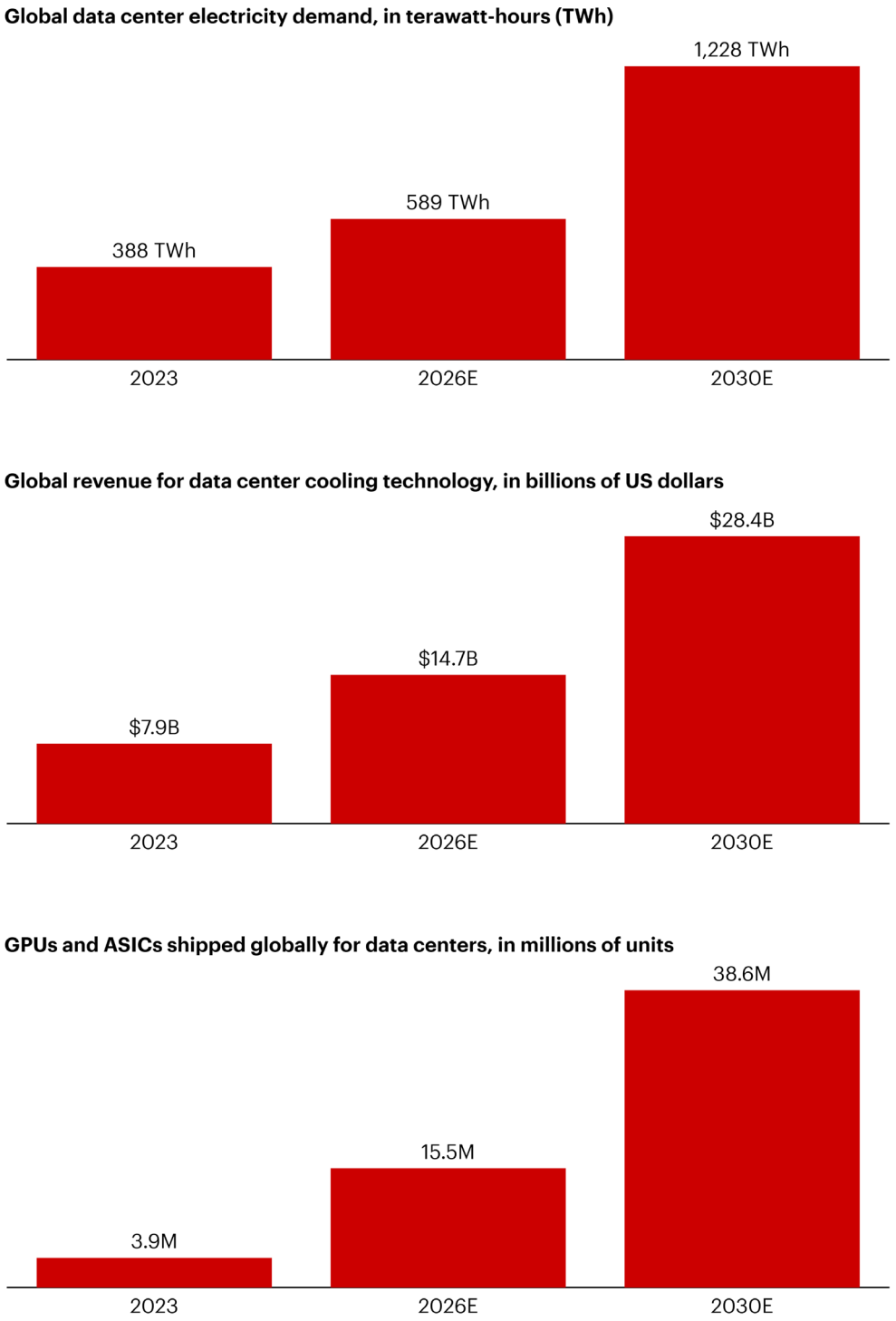
Five years ago, a 50-megawatt (MW) facility was considered large. Now, hyperscalers are planning to build data center campuses 100 times bigger: 5 gigawatts (GW) or more, costing $150 billion to $200 billion each.

AI compute demand is already outrunning supply, and the shortage is intensifying. To overcome it, the tech industry is planning investments at unprecedented scale and speed. In a base case, Bain's Data Center Model projects approximately $5 trillion to $6.5 trillion of spending to build nearly 150 GW of new compute capacity by 2030 *(see Figure 1)*, almost tripling global capacity in just five years. Most of the growth will come from the US, but compute capacity will grow by double digits across regions.

What many business leaders are missing is that the AI compute shortage is unlikely to resolve on its own. The usual assumption that supply chains respond to demand and that they'll eventually balance doesn't hold here.

Power, hardware, and skilled labor are simultaneously constrained *(see Figure 2)*, and public approvals are increasingly difficult to secure. Lead times are long enough that decisions made today determine capacity several years from now.

Figure 2: Data center expansion will accelerate demand for resources across the supply chain



Note: Cooling market data excludes China
Sources: Bain Data Center Model, July 2026; Bain AI Chips Model, July 2026; TD Cowen; Bain analysis

- **Power:** Grid connection is delaying and freezing projects. Adding grid capacity takes four-plus years, and critical equipment inventory (e.g., transformers, etc.) is scarce. Onsite, behind-the-meter power relieves some of the bottleneck, but it adds regulatory exposure, longer equipment lead times, and more project complexity.

- **Semiconductors:** GPUs, memory, optical networking, and other components are being locked in years ahead. Operators without long-term supply agreements are effectively out of the queue.

- **Labor:** Skilled trades are the human version of the same problem, with too few electrical workers, mechanical tradespeople, and cooling specialists. Meeting demand means recruiting, training, and retaining talent far above historical rates.

- **Regulations and public opinion:** Complex permitting rules across local, state, and national levels add time and uncertainty, even when primary inputs are secured. Citizens and politicians are increasingly pushing back regarding water use, energy prices, and noise. Local opposition blocked or delayed at least 75 projects worth $130 billion in the first quarter of this year alone, nearly matching the $156 billion worth of projects disrupted in all of 2025. In July, New York became the first US state to pause construction of the largest data centers; 14 others have introduced similar legislation. Countries including Ireland, the Netherlands, and Denmark have restricted or paused new grid connections, at least temporarily.

Structural solutions are required

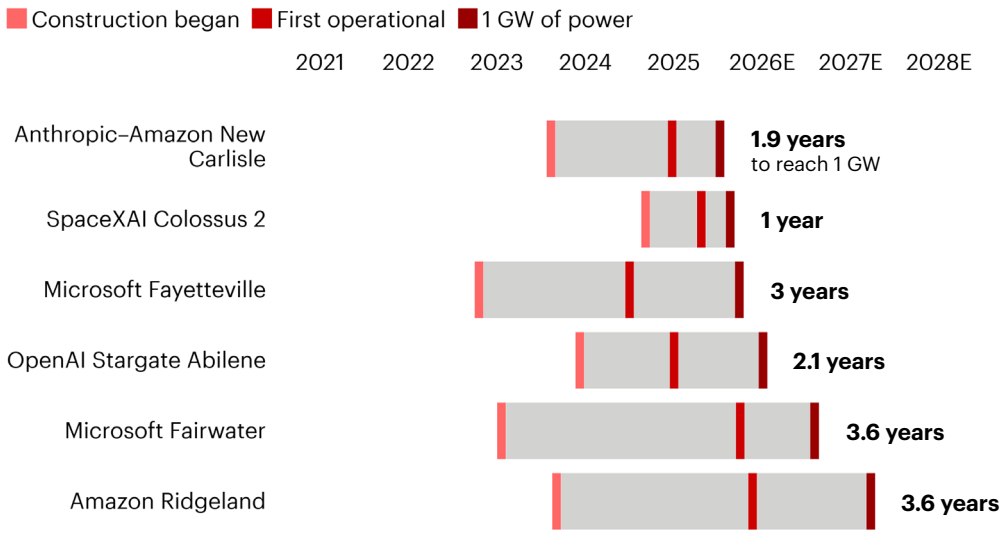
Individual projects are finding workarounds, but they can't be replicated at scale, and some are controversial. Colossus (from xAI, now part of SpaceXAI) was built in under a year by repurposing a factory and leasing dozens of gas turbines without federal clean air permits, drawing a lawsuit. Constructing a 1 GW data center typically takes one to four years, even on sites with existing power *(see Figure 3)*, and it can take longer when new generation and transmission infrastructure is needed.

Meeting the global ambition instead will require system-level fixes outside the usual free-market playbook. A grand challenge on this scale demands grand solutions, and executives planning on business as usual will get blindsided. The forces below are deliberately provocative and mostly long term, and some will create new challenges, even as they mitigate others. But overcoming the shortage will take at least one of the following forces: consolidation, new power at scale, orbital data centers, government investment, and social license.

Consolidation: Consolidation could occur as capital requirements grow so large that only a handful of companies survive. Hyperscalers' data center capex could top $1 trillion by 2028.

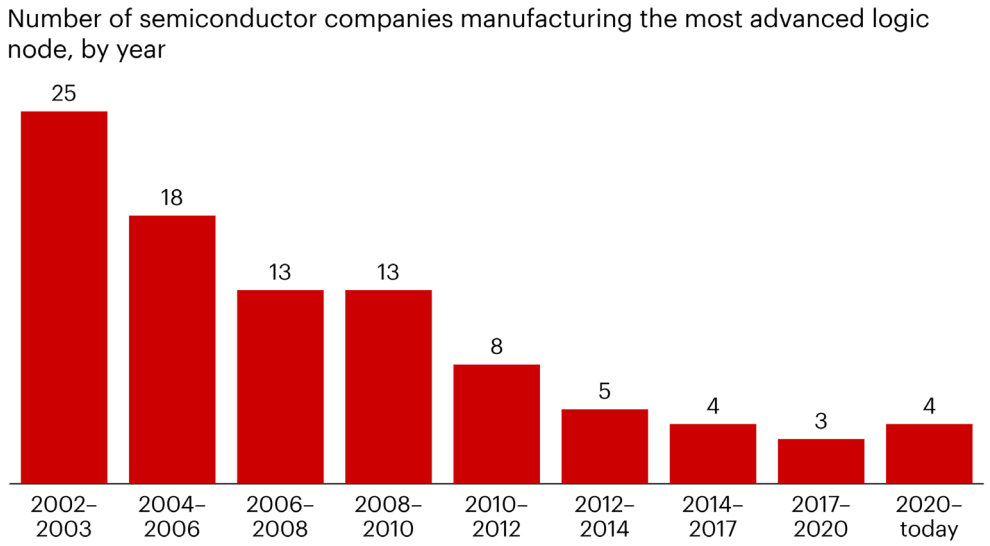
Semiconductors provide a useful corollary. Staying at the leading edge grew so costly (tens of billions of dollars per generation) that missing one generation made catching up effectively impossible. Companies exited, were acquired, or retreated to older nodes *(see Figure 4)*.

Figure 3: Building a 1 GW data center typically takes one to four years, even on sites with existing power



Source: Epoch AI

Figure 4: Most semiconductor manufacturers have exited the market for the most advanced logic node, leaving TSMC, Samsung, Intel, and Rapidus



Note: Does not include companies that entered the node after the logic node leadership period
Source: Bain analysis

Facing unsustainable costs, data center providers may also turn to once-unthinkable solutions: shared power procurement, colocated servers, and jointly funded grid upgrades.

New power at scale: Power is one of the most challenging, multidimensional bottlenecks, and no single source solves it. The realistic path combines near-term options that could come online this decade with long-term ones that are scheduled to pay off in the 2030s.

Near term, behind-the-meter generation (mostly natural gas today) is one of the few ways to add large blocks of power on data center timelines. Pairing it with carbon capture, utilization, and storage could make it cleaner, but that's unproven at scale.

Tech companies are also turning to renewables, storage, and virtual power plants (VPPs) that use software to aggregate and redirect power in real time from distributed sources such as batteries and smart thermostats. Google, for instance, is funding 1.6 GW of renewables (mostly wind) and 300 MW of storage alongside a new Minnesota data center. North America's VPP market grew by 13.7% last year, to 37.5 GW, per Wood Mackenzie. Novel VPP agreements are emerging, such as the Google-Voltus deal that would aggregate up to 100 MW annually over three years in the PJM grid.

Power is one of the most challenging, multidimensional bottlenecks, and no single source solves it. The realistic path combines near-term options that could come online this decade with long-term ones that are scheduled to pay off in the 2030s.

Nuclear is another clean source that could eventually help, but new, large-scale capacity realistically won't arrive until the mid-2030s. Recommissioning existing facilities (as at the Palisades and Three Mile Island power plants in the US) and building new traditional nuclear reactors would require national coordination of schedules, supply chains, the workforce, regulations, and financing.

Small modular reactors are a credible piece of the puzzle, but likely can't move fast enough or scale alone. Google (via Kairos Power) and Amazon (via X-energy) are among those betting on this technology.

Orbital data centers: Closing the gap may require technologies that don't fully exist yet. Orbital data centers could theoretically address land, cooling, and power constraints. Similar to nuclear, they're long-horizon bets, not near-term relief. SpaceXAI's post-IPO valuation reflects investor conviction that space-based compute will eventually matter, but timelines and economics remain unproven.

Government investment: Sovereign entities are already investing directly in AI infrastructure and chips, including in the US, South Korea, UAE, Saudi Arabia, and EU.

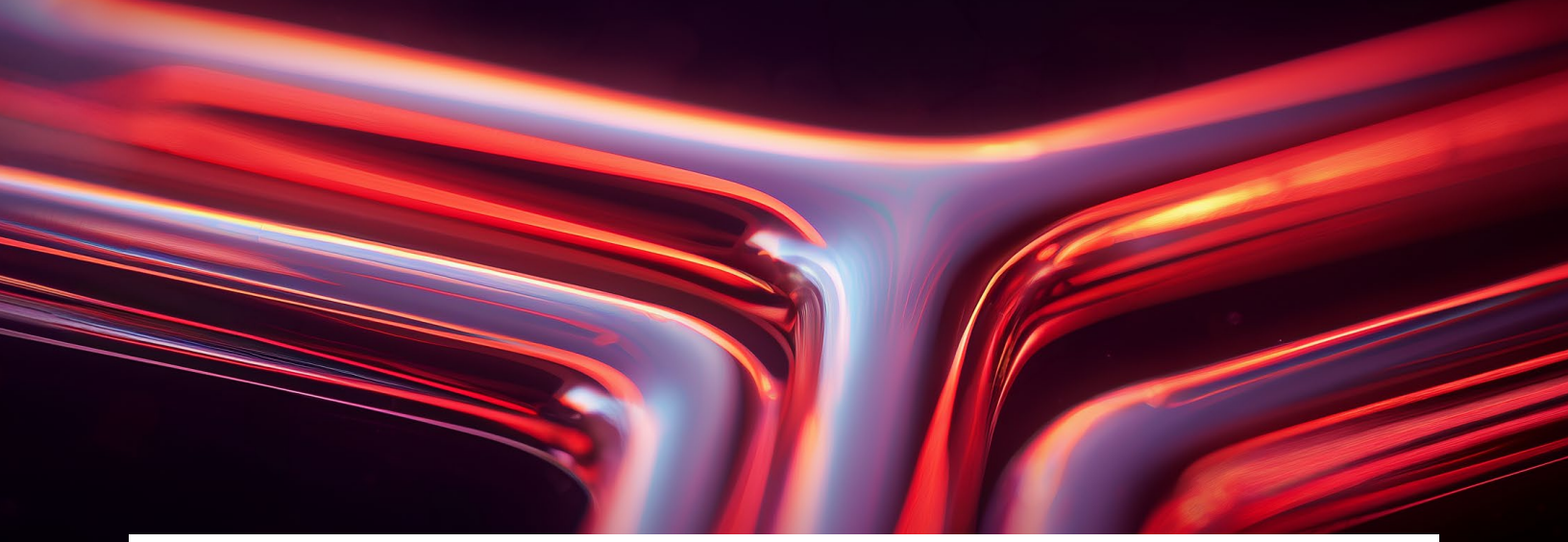
Governments may move beyond subsidies toward a direct role, such as coordinator, financier, or (in some countries) even owner. For example, China's government is investing $295 billion in AI data center construction and $740 billion in grid upgrades to pipe renewable power from its resource-rich west to data centers. But government interventions can be double edged. A coordinated, Manhattan Project–style push might accelerate builds. But the same interventions that can unlock power and land can introduce control, oversight, and delays, and companies and investors will need to plan around that risk.

Social license: This will be as decisive as build capability. Communities and governments will increasingly demand binding agreements on data use, environmental impact, and governance before granting permits.

Frameworks are emerging. In the US, the city of Lancaster, Pennsylvania, signed a contract with Chirisa Technology Parks and Machine Investment Group pursuant to an agreement to codevelop a data center for CoreWeave that caps its water use and sets clean energy, noise, emissions, and local hiring requirements.

Implications for executives

- **Enterprise technology leaders:** Treat compute access as both a supply chain risk and a strategic asset. Those who wait for the market to sort itself out will get locked out.
- **AI providers, hyperscalers, and data center developers:** Because social license and power access are critical to each project's success, building a larger portfolio of carefully chosen and managed sites is now imperative.
- **Regulators:** The window to set coherent rules is open, but it won't stay that way for long. Absent clear frameworks, government involvement won't shrink; it would likely grow but on a crisis timeline, when leverage is gone.
- **Investors:** Capital needs for data infrastructure will remain high, and with a growing field of firms building and operating these assets, assessing counterparty risk and end-customer exposure is essential. Bottlenecks in power, semiconductors, and other inputs carry large capital needs of their own, opening additional entry points for investors. And as sovereign infrastructure becomes a bigger part of national strategies, partnerships offer both a way in and geographic diversification.
- **Power providers (utilities and independent power producers):** Partnering earlier with AI providers, hyperscalers, and developers is critical to both shape power requirements and jointly craft creative solutions to meet system, community, and affordability needs.



STRATEGIC BATTLEGROUNDS

The $100-Billion SaaS Opportunity Hiding in Cross-System Labor

Agentic AI can automate much of the coordination work among systems, creating a large new market for software companies.

By David Crawford, Chris McLaughlin, and Greg Fiore

At a Glance

- Agentic AI's big opportunity isn't replacing software as a service (SaaS); it's converting labor costs into software spending by automating coordination work.
- Bain estimates the potential market could be $100 billion in the US, and more than 90% remains uncaptured.
- Not all user workflows are equally automatable: Six factors determine where agents can realistically take over.
- Winners assess data assets, identify valuable workflows, close competitive gaps, and use every deployment to capture data that makes the system smarter.

Generative and agentic AI are reshaping software as a service (SaaS), compressing some workflows, enhancing others, and in certain cases cannibalizing them entirely. (See our recent brief "Will Agentic AI Disrupt SaaS?" for our take on the offensive and defensive priorities SaaS leaders should set.)

But market disruption is only half the story. The more actionable question for tech executives and investors is: Where does agentic AI create entirely new markets for software, and how do you capture them?

The answer, it turns out, lies not in the systems of record themselves but in the expensive human labor that connects them.

Converting labor costs into software spending

Every enterprise runs on workflows that span systems: enterprise resource planning (ERP), customer relationship management (CRM), billing, support ticketing, and vendor management. Between those systems sits an enormous amount of what we call cross-system coordination work—employees pulling budget data from an ERP, checking inventory in a spreadsheet, interpreting free-text responses, and making judgment calls about when to escalate. This coordination work is labor intensive, expensive, and—critically—unable to be automated by rules-based software or traditional robotic process automation (RPA).

The reason is straightforward. Rules-based automation breaks when it encounters ambiguity, context dependence, or information scattered across multiple systems. An RPA bot can move data from field A to field B, but it can't reconcile budget, inventory, and project context across an ERP and three spreadsheets. It can't interpret a vendor's ambiguous email and decide whether to respond, escalate, or wait. It can't chase an approver whose delay is driven by an organizational signal rather than a process failure.

Agentic AI changes this. Agents can interpret information from disparate sources, reason about context, coordinate decisions across systems, and execute end to end with policy guardrails rather than hard-coded rules *(see Figure 1)*. This unlocks automation of the most expensive, human-mediated work in the enterprise, and it represents the largest and most tractable new potential for software's total addressable market (TAM).

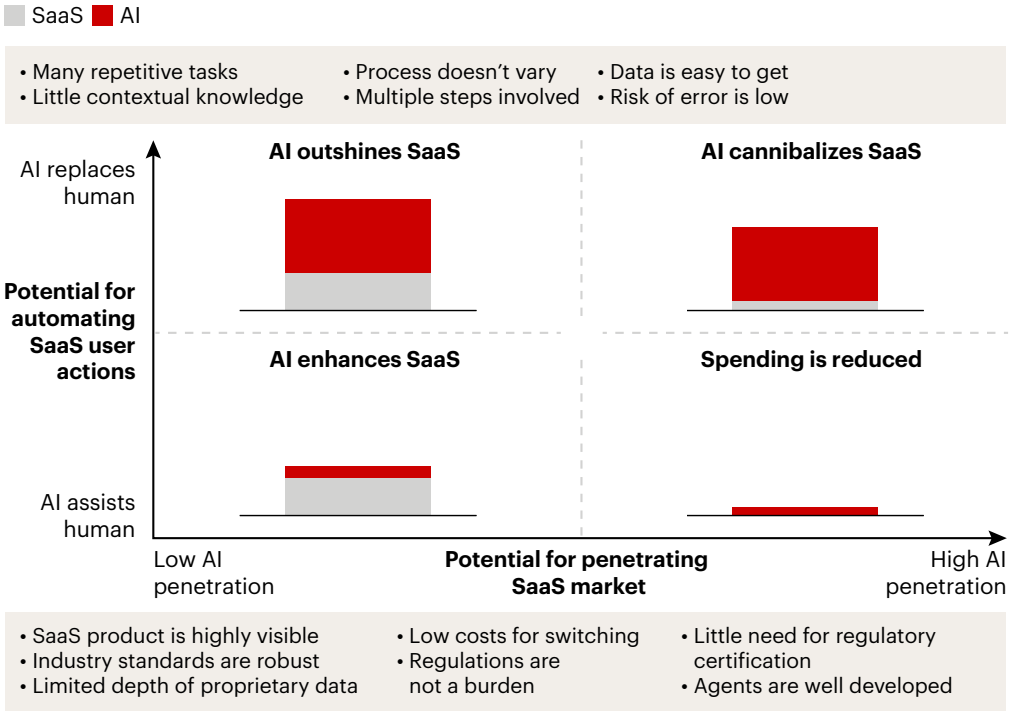
In some cases, AI will also extend the value of existing SaaS by automating workflows such as data presentation and summary, record storage, and comparison of current and prior reports from the same system.

For two decades, SaaS companies built competitive moats around systems of record. Own the CRM and you own the customer relationship. Own the ERP and you own the financial backbone. Depth in a single domain was the winning strategy.

Agentic AI reopens this contest. The new competitive advantage is what we call cross-workflow decision context—the ability to see, interpret, and act across workflows that traverse multiple systems *(see Figure 2)*. The highest-value automation opportunities are concentrated precisely where no single system of record owns the outcome—where decision context spans, for example, ERP, CRM, billing, and support, and where historical decision-making patterns are fragmented across all of them.

This shift is already playing out in the market. Sierra built its platform to resolve customer issues autonomously across enterprise systems rather than within a single ticketing tool. Glean coordinates employee requests across multiple functions rather than indexing a single knowledge base. GitHub's Copilot expanded from source control into AI-powered developer productivity and security automation—

Figure 1: Assessing AI's ability to automate workflows can determine the future of software-as-a-service (SaaS) products



Source: Bain & Company

a move enabled by its observational data across code repositories, not by owning any single developer workflow. In each case, breadth of visibility matters more than depth in any one domain.

The implications for pricing are equally profound. When agents deliver end-to-end outcomes—a resolved customer issue, a processed invoice, a qualified lead—the natural pricing unit shifts from seats and logins to outcomes and usage. Companies like AppLovin have already made this transition, monetizing based on advertising performance outcomes rather than platform access. Cursor offers tiered seat-based pricing, but the value proposition is fundamentally about developer output, not tool access.

Bain's research estimates this new market could be worth $100 billion in the US *(see Figure 3)*. Vendors are already capturing about $4 billion to $6 billion, but more than 90% of the opportunity remains untapped. AI-native entrants are scaling fast. Cursor has surpassed $2 billion in annual recurring revenue (ARR), doubling in a single quarter. Sierra has crossed $150 million. Harvey has passed $190 million. Glean reached $200 million. And fast-mobilizing incumbents like Salesforce (with Agentforce), ServiceNow (which acquired Moveworks), and Workday are embedding agentic capabilities into their existing platforms.

Figure 2: Cross-workflow observability will overtake system of record ownership as the key advantage

Old paradigm	New paradigm	Examples
Own the system of record	Own the workflow decision context	**Harvey:** Builds defensibility around legal reasoning and workflow decisions
Depth in one function	Breadth across functions	**Glean:** Coordinates employee requests across multiple functions
Per-seat pricing	Outcome-based pricing	**AppLovin:** Monetizes based on advertising performance outcomes
Human-in-the-loop by default	Agent-in-the-loop by default	**Sierra:** Autonomously resolves tickets across CRM, billing, and support; humans handle escalations

Source: Bain & Company

The opportunity is not evenly distributed across functions. Sales represents the largest single slice at roughly $20 billion, due more to the large number of sales employees rather than an exceptionally high potential for automation. Cost of goods sold (COGS) and operations contribute about $26 billion, reflecting the enormous operational workforce, where even modest automation rates translate into large dollar figures. R&D and engineering, customer support, and finance each represent $6 billion to $12 billion in addressable opportunity. These functions combine relatively high automation potential (40% to 60% for customer support and R&D) with meaningful employee bases and salary levels.

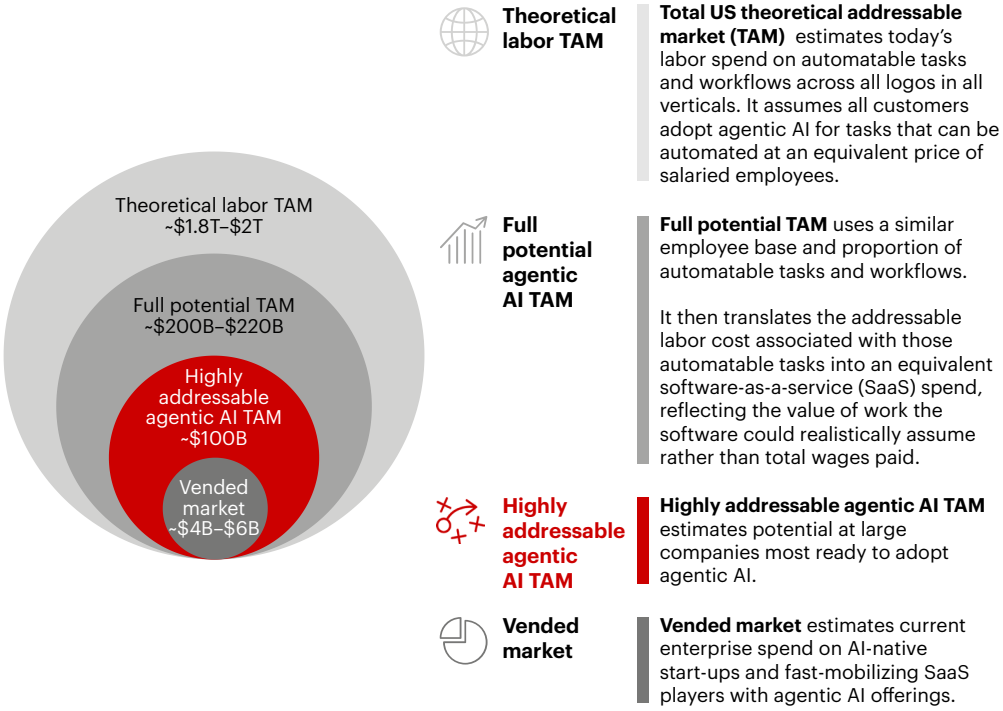
Looking beyond the US, the opportunity in Canada, Europe, Australia, and New Zealand is about the same, doubling the total opportunity to about $200 billion.

Assessing potential for workflow automation

A common mistake in AI strategy is treating automation potential as a binary—either a workflow can be automated or it can't. In practice, six factors determine how much of any given workflow an agent can realistically handle.

- **Output verifiability** measures whether an agent's work product can be validated automatically or by humans. Code that compiles and passes tests, invoices that reconcile, and support tickets that resolve all generate clear verification signals. Financial planning scenarios, by contrast, depend on subjective judgment and are therefore harder to verify.

Figure 3: The highly addressable market for new applications of agentic AI in the US could be about $100 billion



Note: US spending only
Sources: US Census; Bain analysis

- **Consequence of failure** captures the regulatory, safety, or financial liability involved. When failures are costly or irreversible—as with tax filings, security incident response, or legal compliance—enterprises require tighter human oversight, even when agents are technically capable. Agents increasingly operate in a shadow mode alongside human workers in these domains, demonstrating accuracy before they're granted autonomy.

- **Digitized knowledge availability** assesses whether agents have well-labeled, machine-readable inputs at runtime, including the tacit knowledge that typically lives only in people's heads. Rich, structured data in ticketing systems and code repositories makes customer support and engineering highly automatable. But even in data-rich environments, automation stalls when the context for a decision hasn't been documented: a veteran buyer's vendor preferences, an engineer's institutional memory of past failures, or the unwritten exception logic that governs how a policy is actually applied. In practice, the availability of this digitized contextual knowledge is almost always the binding constraint on automation.

- **Integration and orchestration complexity** captures how many systems, application programming interfaces (APIs), and handoffs an agent must coordinate to complete a workflow end to end. Single-

system workflows are straightforward; workflows spanning five or six systems of record with different authentication, data models, and exception-handling logic are harder to orchestrate reliably.

- **Process variability** distinguishes straight-through flows from those with many exceptions. Agents are increasingly capable of handling branching logic and common exceptions, but workflows with lots of novel, unpredictable edge cases still force escalation back to humans.
- **Physical world dependency** captures whether a workflow requires actions that computers physically cannot perform or that organizations do not yet trust them to perform unsupervised, such as in-person inspections, physical signatures, or safety-critical interventions. Where these dependencies exist, a hybrid human-agent model is likely to continue, regardless of how capable AI becomes.

Applying these six factors across enterprise functions reveals a wide range of automation potential.

- Customer support and R&D/engineering sit at the high end, with roughly 40% to 60% of workflow tasks automatable. Both benefit from structured data, standardized processes, and well-defined outputs.
- Finance and HR fall in the 35% to 45% range, with pockets of very high automation (accounts payable, payroll) alongside judgment-heavy work (financial planning and analysis, employee relations).
- Sales and IT sit at 30% to 40%, constrained by relationship nuance, deal-by-deal variation, and the unpredictability of security incidents.
- Legal offers an interesting contrast: Contract review and compliance workflows are highly repeatable, but the consequences of errors are severe, keeping overall automation in the 20% to 30% range despite strong structural fit.

Two paths to grow the market

SaaS companies have two main paths for market growth: automating core or adjacent workflows.

Automating core workflows is the most straightforward option because it's where SaaS companies have the deepest knowledge, strongest systems integration, and greatest buyer trust. While there's some risk of cannibalizing seat-based revenue by eliminating seats, the overall economics can be better, since many customers are willing to pay more for getting the whole job done, rather than paying for software that helps an employee do it.

The second path is automating adjacent workflows that are not offered by the company today. Agentic AI can unlock the latent value of underlying data assets and customer penetration in those adjacent workflows. This second path is harder because these patterns and opportunities are often nonobvious and require bottom-up, detailed mapping of workflows that are economically important to customers.

Figure 4: Observable assets unlock nonobvious adjacencies

Observable assets	Adjacent workflows enabled	Why uniquely positioned?
Decision patterns with outcomes	Decision automation	Knowledge of what "good" looks like
Work product repository	Work product generation	Training data and style patterns
Process traces	Process execution	Data around workflow steps and exception logic
Interaction transcripts	Interaction automation	Know-how of what works in conversations
Reference/lookup patterns	Proactive recommendation	Knowledge of what people need before the search begins

Source: Bain & Company

GitHub illustrates this pattern clearly. Its core business is developer collaboration and source control: Git repositories, pull requests, code reviews, and continuous integration/continuous deployment (CI/CD) workflows. But the observational data generated by that core business—decision patterns with outcomes, work product repositories, process traces, interaction transcripts, and reference and lookup patterns—enabled GitHub to expand into AI-powered developer productivity (Copilot) and security automation (code scanning and dependency analysis). Neither of these adjacent workflows was part of GitHub's original product, but no competitor could replicate the data position that made them possible.

Most companies won't get to choose whether they are in the first category. Their core workflows are either getting automated or they aren't—and if they are, the only question is whether you are the one automating them or a competitor is. But companies can choose how aggressively to leverage differentiated assets to pursue the second path: automating adjacent workflows that agentic AI has turned into newly accessible market share. The most valuable opportunities often emerge from nonobvious adjacencies—workflows that technology firms don't serve directly, but that their decision context data uniquely enables *(see Figure 4)*.

A playbook for capturing market share

Phase one: Assess the upside. Start by identifying the high-value workflows of your end customers that are now automatable with agentic AI. Apply the six automatability factors at the subprocess level—not the function level. For example, accounts payable and accounts receivable have very different automation profiles than financial planning and analysis or investor relations. Then size the market potential by

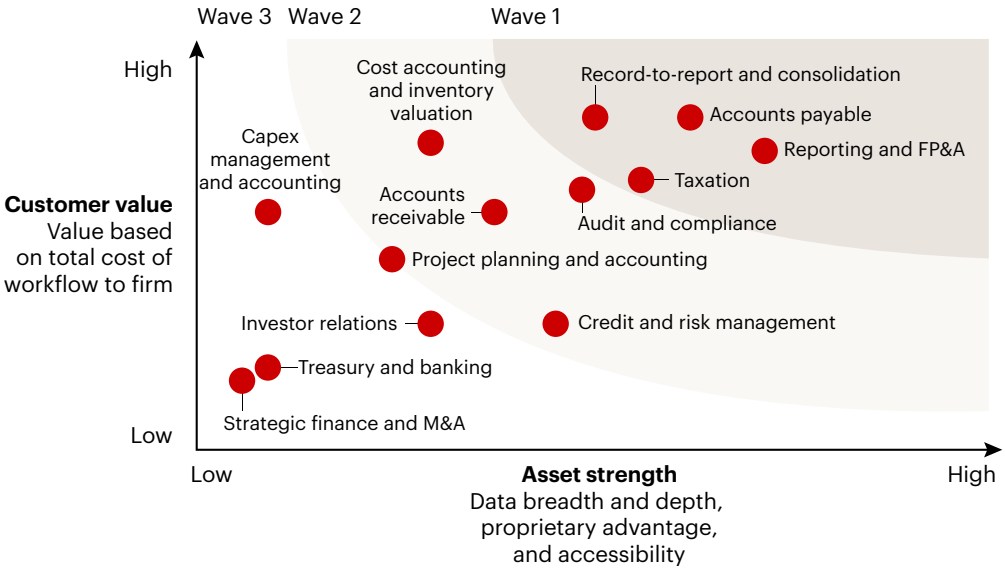
comparing the cost of deploying agents with the labor costs they replace or augment. Focus where the economics are most compelling.

Phase two: Decide where to play. Assess your data assets to understand how comprehensive, detailed, outcome-linked, differentiated, and usable they are for automation. Next, identify the workflows where those assets can create the greatest differentiated value, including adjacent opportunities that may not be immediately obvious. Finally, develop a clear, end-to-end view of how those workflows actually operate. Pay particular attention to the informal steps, handoffs, and exceptions, since much of the real automation opportunity lies in how work gets coordinated beyond core systems *(see Figure 5)*.

Phase three: Execute with three coordinated moves. First, close critical capability gaps with decisions to build, buy, or partner to secure the technology, data, and coverage needed for differentiated automation. AppLovin built its Axon platform in-house to maintain control over ad optimization data. ServiceNow acquired Moveworks to add an AI layer that made its workflow engine more discoverable. Salesforce partnered with Workday to extend Agentforce across HR and finance workflows without building a competing system.

Second, set up the organization to execute. Attract AI engineering talent, build cloud-native architecture for multi-agent orchestration, and secure capital for model training and inference. Align incentives by shifting pricing from seats to outcomes, restructuring sales compensation to reward AI-driven results, and removing structural bias toward legacy products.

Figure 5: Automation will play out in waves, beginning with the highest-value activities first



Note: FP&A is financial planning and analysis
Source: Bain & Company

Third, redesign your data and product foundations to be agent ready. Create agent-native data models—schemas and handoffs designed for machine execution rather than human interpretation. Enable agents to act end to end, capturing decisions and outcomes to build longitudinal data. Use that data to improve performance with every run. The goal is a durable moat: accumulated execution data that grows more valuable over time and becomes harder for competitors to replicate.

Take action

Technology transitions are often met with hype as well as resistance, and are framed in negative terms such as disruption, dislocation, or competitive threat. But once they're understood and incorporated into company strategies, they often unleash waves of innovation and opportunity for incumbents and new entrants alike.

The strategic imperative isn't to protect the legacy SaaS model from disruption; it's to capture the larger opportunity of converting labor costs into software spending.

Still, the opportunity requires action. The strategic imperative isn't to protect the legacy SaaS model from disruption; it's to capture the larger opportunity of converting labor costs into software spending. But the window for SaaS leaders to define their position in the agentic AI market is measured in quarters, not years. AI-native startups are scaling at unprecedented speed: Cursor went from $100 million to $2 billion in annual recurring revenue in 14 months. Enterprise buyers are already deploying agents in high-volume, high-cost workflows. And every month that passes generates more execution data for early movers to compound their advantage. The companies that move first to map priority workflows, build proprietary data assets, and scale automation faster will define the next era of enterprise software.

Disruption was inevitable. The real question now is: Who captures the new markets?

STRATEGIC BATTLEGROUNDS

Hardware Strikes Back in the AI Era

For the first time in years, hardware and semiconductors are growing faster than software. The shift is structural, not cyclical.

By Michael Schallehn, Peter Hanbury, and David Crawford

At a Glance

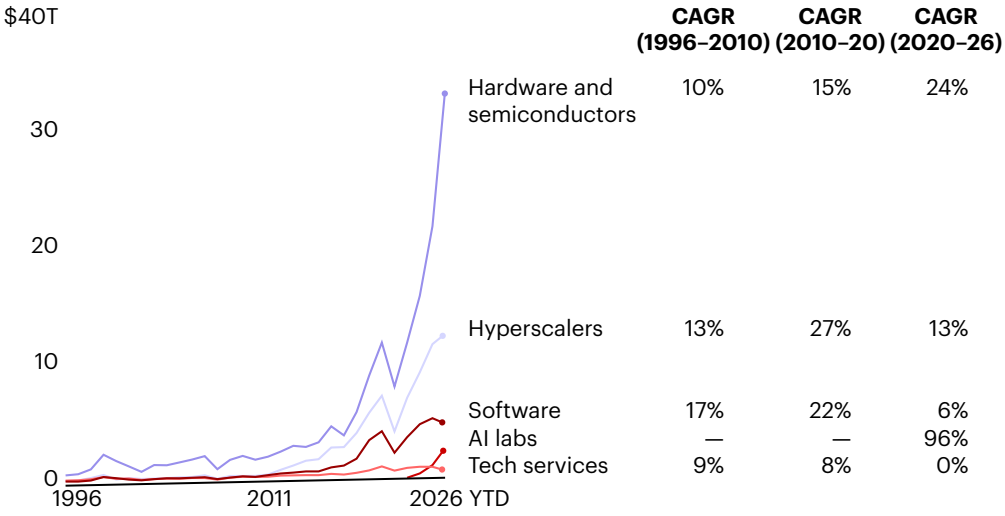
- In a reversal, hardware and semiconductor stocks grew at a 24% compound annual rate from 2020 to 2026 vs. 6% for software.
- Insatiable AI compute demand is creating value in bottleneck technologies such as high-bandwidth memory and advanced packaging.
- Custom silicon has become the fastest-growing type of data center compute chips.
- Silicon architecture, supplier diversification, and geopolitical exposure matter more than ever.

For nearly two decades, software was tech's fastest-growing sector while semiconductors and hardware operated in its shadow—admired for engineering, undervalued by markets. AI has reversed that.

Hardware and semiconductor valuations surged over the past several years while software valuations compressed—the so-called SaaSpocalypse *(see Figure 1)*.

Figure 1: Hardware market valuations have grown at a much faster pace than other tech sectors in recent years

Market cap, in trillions of US dollars



Notes: Includes publicly traded companies except for AI labs category; AI labs category includes OpenAI, Anthropic, xAI, and others, and their market valuations are based on private investors' valuations except for xAI, which merged with SpaceX in February 2026 at a $250 billion valuation and has been included in 2026 data; all market cap data is from December 31 of respective year; S&P Global BMI data is from May 28, 2026; software category includes IT sector companies and select non-IT firms with heavy tech-driven models; values are rounded
Sources: S&P; Pitchbook; Bain analysis

An AI infrastructure flywheel reversed the hierarchy. Hyperscaler capex, venture-backed neo-cloud build-out, and frontier-model development are fueling step-change demand for compute. That's accelerating innovation across the hardware and semiconductor stack and creating new pockets of value in bottleneck technologies, such as high-bandwidth memory (HBM), leading-edge silicon and advanced packaging, optical networking, and custom AI silicon.

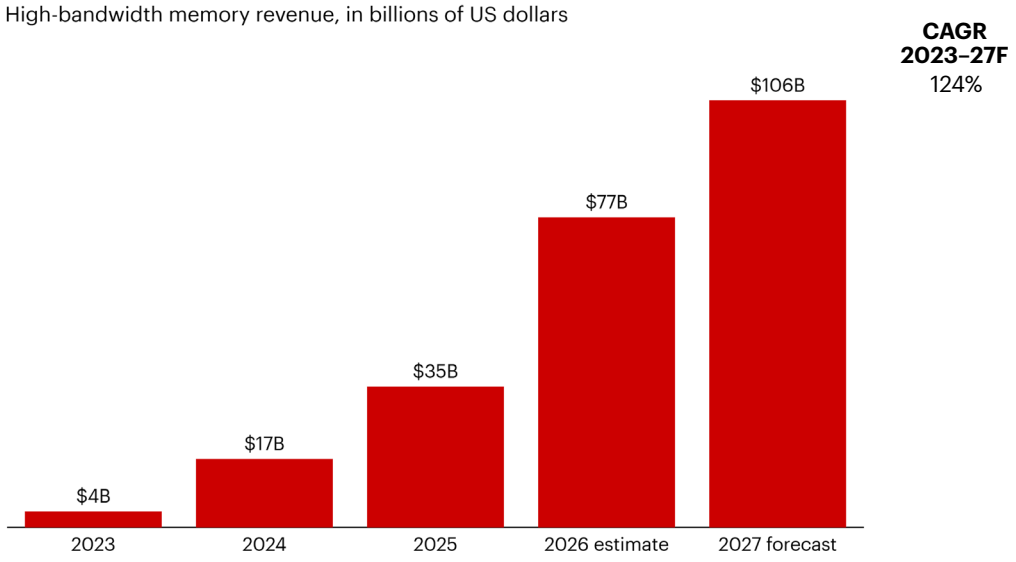
We'll focus on three of the fastest-growing segments.

High-bandwidth memory: Samsung, SK Hynix, and Micron have become critical development partners in the AI data center boom, pushing HBM revenue up sharply since 2023 *(see Figure 2)*.

To improve performance, dynamic random-access memory (DRAM)-based stacks are tightly integrated with a logic die at the base. That codevelopment means vendors can no longer swap HBM suppliers as easily as with commodity DRAM, shifting power dynamics.

High, stable margins are new to a famously cyclical memory market. SK Hynix and Micron posted record gross margins of 75% to 85% in each of the past two quarters. The three major DRAM players have concentrated most new capacity on HBM, leaving limited investment for conventional DRAM and NAND. That's worsening shortages and raising the prices of smartphones and PCs.

Figure 2: Sales of high-bandwidth memory chips have exploded as a result of the AI data center boom



Note: Values are rounded
Sources: Bain analysis; Bank of America; New Street Research

Advanced packaging: As AI fuels near-insatiable demand for leading-edge compute, Moore's Law no longer bends the cost curve as it once did. Advanced packaging picks up some of the slack, delivering system-level performance and power gains.

Advanced packaging has become a competitive battleground of its own, and the market is growing rapidly *(see Figure 3)*. Mature and emerging technologies will allow chipmakers to integrate heterogeneous dies that combine logic, memory, and input/output with performance approaching that of monolithic designs. The winners will define the next generation of compute density.

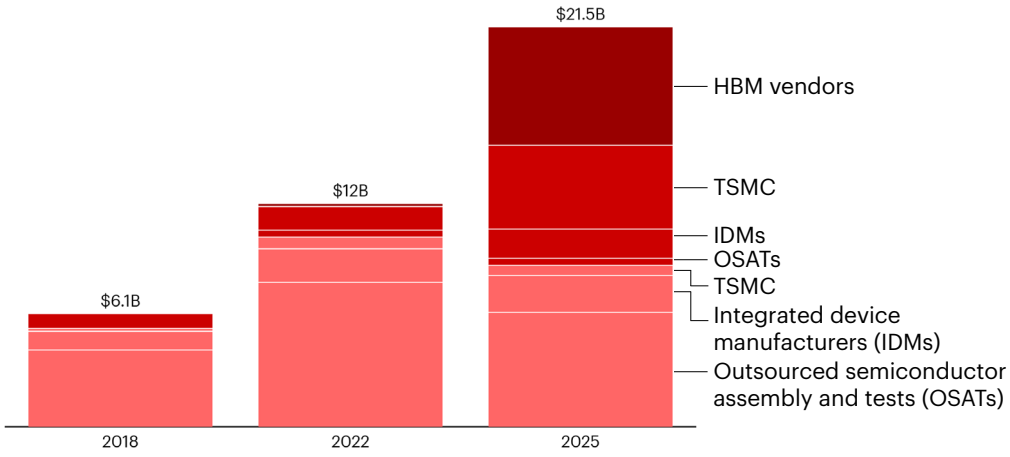
Custom silicon: Special-purpose, high-performance accelerators are moving from niche to mainstream as hyperscalers and AI-native firms design chips tuned to their workloads. The proliferation of these designs has fueled the rapid growth of Broadcom's ASIC unit while smaller companies such as Groq and Cerebras have carved out defensible niches that are now beginning to scale. The result: ASICs are the fastest-growing segment of data center compute chips *(see Figure 4)*.

Four shifts turned custom silicon from an uneconomic bet into the default strategy for anyone competing at scale.

The most important is the emergence of grand-scale, homogeneous workloads (training, inference, agentic) that now run well past the volume needed to amortize a custom design. Each favors different silicon architectures. Training's sustained floating-point operations per second, throughput orientation,

Figure 3: Advanced packaging profits are rapidly growing, with TSMC and high-bandwidth memory vendors capturing much of the new value

Semiconductor packaging gross profit, in billions of US dollars



Notes: Advanced packaging types include high-bandwidth memory, 2.5D, 3D, and fan-out; traditional packaging includes all other types; values are rounded
Sources: Company reports; Yole; Gartner; UBS; Bank of America; Morgan Stanley; Bain analysis

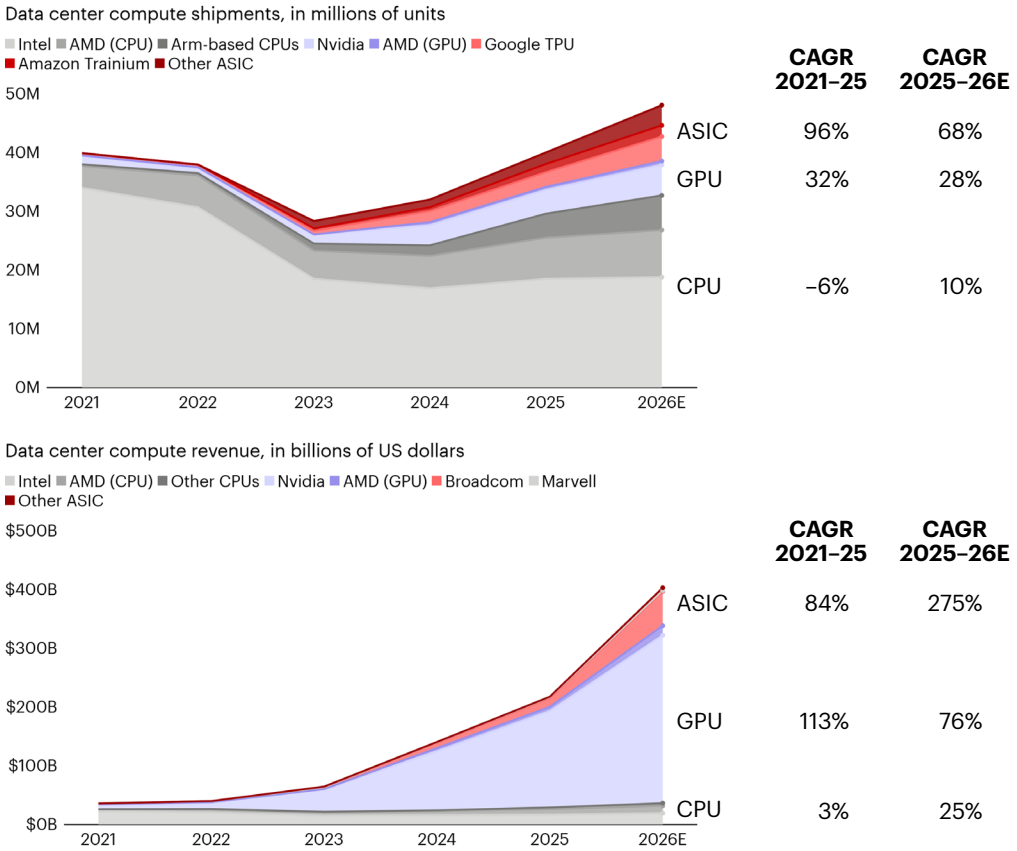
and lower-latency sensitivity require big, dense compute; large HBM; and fast scale-up/scale-out capabilities delivered only by GPUs, wafer-scale processors, or purpose-built chips. Inference's decode phase must reread the model's active weights and key-value (KV) cache from memory for every token, rewarding designs that keep everything in fast, on-chip memory to sidestep HBM constraints. Agentic resembles inference but accumulates context across multistep actions and leaves accelerators idle while waiting on tool calls. That favors less expensive places to park expanding context and faster ways to retrieve it, such as KV cache offloading, memory tiering, and software that toggles between flash and DRAM.

The second shift: The cost of buying turned punishing, with merchant GPU gross margins as high as 75%. Third, the make side got less expensive through the innovation of TSMC, ASIC design houses, and electronic design automation vendors *(see Figure 5)*. This ecosystem has served up chiplets, advanced packaging, foundry capacity, and contract ASIC design services, letting buyers field custom parts without standing up a full silicon organization. Fourth, the payoff has never been higher. Silicon can account for most of a data center's cost and power consumption, so every efficiency gain compounds.

Semiconductors as strategic assets

The pandemic-induced chip shortage made silicon supply a board-level priority, and now the AI boom has made revamping supply chains more urgent. For example, TSMC said it intends to build nine phases of new fabs this year, more than double its recent annual rate, while also increasing by 80% its

Figure 4: Custom chips are grabbing a larger share of the data center compute market



Note: Values are rounded
Sources: TD Cowen Datacenter Silicon Model, April 2026; Epoch AI; Bain analysis

manufacturing capacity of chip on wafer on substrate and system on integrated chips. Even so, the company has said it won't be able to keep up with customer demand. As leading-edge capacity tightens, single-source exposure brings pricing and supply risks from natural disasters, geopolitical disruption, and export controls.

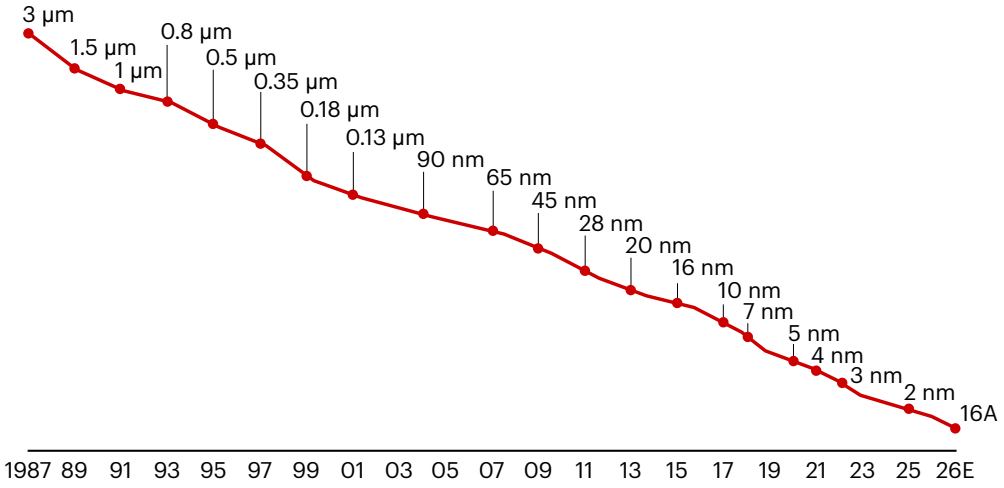
Companies are securing multiple supply locations, not just multiple vendors. Foundries are responding in kind, particularly in logic chips *(see Figure 6)*. This diversification answers geopolitical tension around Taiwan, still the industry's single-largest supply source and vulnerability.

Key implications

For tech and industrial leaders, hardware strategy is no longer solely the procurement team's remit. It demands C-suite attention. Here are five of the most important takeaways for executives.

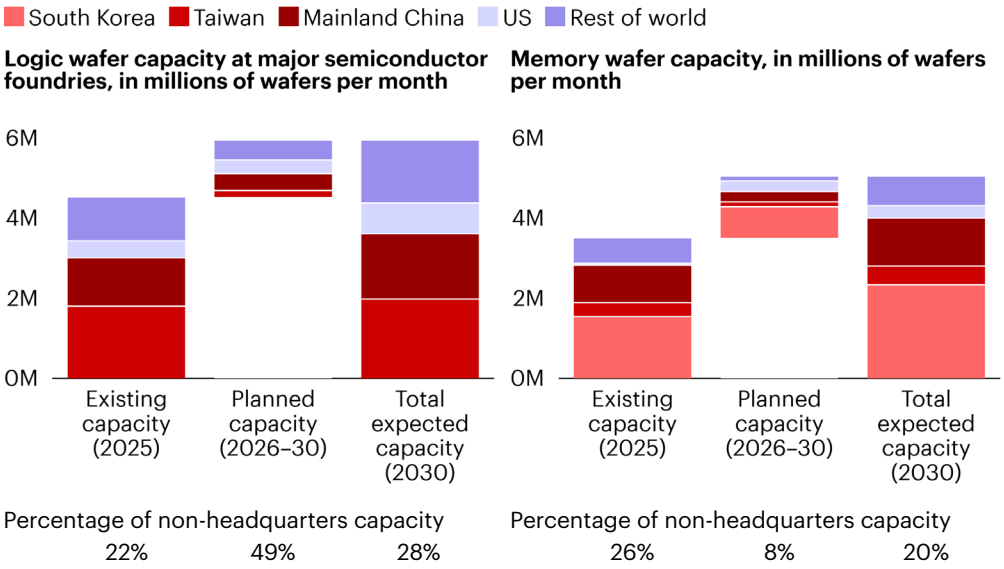
Figure 5: TSMC has advanced process technology over the past four decades

Process node (log scale), in microns (μm), nanometers (nm), and angstrom (A)



Sources: Bain analysis; TSMC

Figure 6: Logic foundries are diversifying their global manufacturing footprint while memory companies are doubling down on their headquarter countries



Notes: Memory wafer capacity includes high-bandwidth memory, DDR, and NAND; rest of world includes Japan, Singapore, France, Germany, South Korea (logic only), Ireland, Israel, and Italy; foundry capacity includes all nodes but doesn't include capacity at non-foundries (e.g., integrated device manufacturers); values are rounded
Sources: Bain Foundry Capacity Database, June 2026; Bain Memory Fab Database, May 2026

Product strategy matters more than ever. Every technology company now faces critical decisions about their silicon architecture and whether to build, buy, or partner. GPUs, CPUs, and ASICs each support inference and agentic workloads in different combinations, each with trade-offs among latency, memory, and other performance variables. Hyperscalers are seeking any edge, increasingly by designing (or even selling) their own chips. For example, Anthropic buys chips from Amazon and Google.

The old model—one vendor innovates and sells to everyone else—is no longer sustainable. A wave of verticalization is likely as leaders determine that they can't cede control of their silicon. The industry is unlikely to revert to the horizontal PC-and-server model in which a handful of companies owned the intellectual property and hardware makers competed on assembly and marketing. A vertical structure is more difficult to navigate, making the build-buy-partner choice matter more. Elon Musk's reported Terafab concept is the most extreme version of the bet, combining the logic, memory, and packaging under one owned roof.

Customers increasingly want to work directly with silicon makers. For example, automakers are starting to bypass the Tier 1 integrators that once sat between them and chip vendors. They want to tune hardware at a low level, ship features faster, and tighten security. Expect the same pull in other industries, along with rising demand for engineers who can work directly with silicon vendors.

Silicon value creation will become less concentrated. As custom work grows, value should spread out from Nvidia. ASICs may soon ship in greater volume than Nvidia's GPUs, even if Nvidia keeps the larger revenue share (as Figure 4 on page 36 shows). Much of the new value will accrue to the companies behind the hyperscalers' chips. A larger set of companies could capitalize, including semi-custom and legacy hardware providers. Design tool vendors that can speed custom silicon design also stand to gain.

Sourcing and supplier management now sit near the top of the C-suite agenda. Procurement has become strategy. How buyers treat their suppliers in good times matters when shortages hit. That means investing in supplier capacity, long-term agreements, equity investments in the supply chain, and multi-vendor, multi-geography sourcing. Companies that treat procurement and supplier relationships as an afterthought will pay for it sooner than later.

Governments will increasingly scrutinize silicon investments for domestic production. Now that they treat leading-edge silicon as strategic infrastructure, governments will push for domestic or allied production through subsidies, export controls, and local content rules.

Supply chain risk continues to grow with East-West tension. This has forced companies to reassess where they source and sell and whether they need distinct China and Western strategies. TSMC is working to get in front of this reality by committing $260 billion to build fabs in Arizona. For the West, the need for large-scale leading-edge production remains critical.

Recent conflicts in the Gulf region highlight a broader infrastructure risk: High-capex projects such as data centers and semiconductor fabs must account not only for resources and capital access but also exposure to military disruption, transport choke points, and regional security volatility.

STRATEGIC BATTLEGROUNDS

Tech Services: Cracking the Code on AI-led Growth

AI has repriced the entire tech services sector.

By Megha Chawla, Mark Brinda, Sushant Khandelwal, and Sandeep Nayak

At a Glance

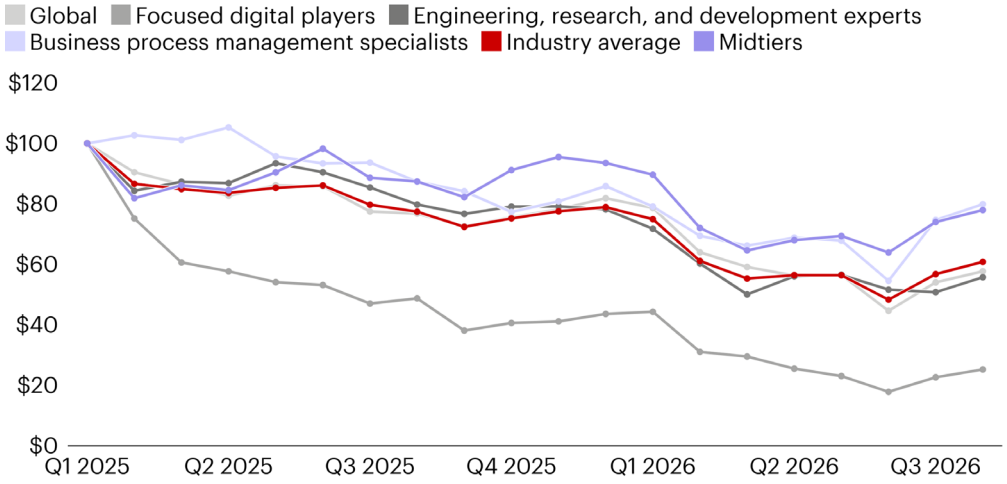
- AI is reshaping the economics of tech services, with investors and clients already expecting major productivity gains and lower delivery costs.
- While AI can deliver significant productivity improvements, few firms have successfully scaled AI delivery or built repeatable commercial models.
- Leaders will focus on differentiated market segments, AI-first delivery models, new pricing and talent strategies, and turning AI into measurable business outcomes for clients.

The market is pricing in a dramatic AI-driven reset for tech services companies. Investors increasingly expect AI will fundamentally compress the labor required to deliver services, enabling the same output with a fraction of the headcount. The drop in share prices reflects investor concerns that these companies will not adapt quickly enough to operating models that depend on fewer employees *(see Figure 1)*.

Customers are already factoring in productivity gains before they arrive, demanding as much as 20% efficiency improvements at renewal negotiations and awarding new work on assumptions of at least 30% (and in some cases as much as 50%) lower delivery costs. With many contracts spanning three to five years, service providers are committing to productivity they have yet to achieve.

Figure 1: Tech services company share prices have declined sharply since first quarter 2025

Share price trend of listed tech services companies, in US dollars (indexed to January 2025)



Sources: S&P Capital IQ; Bain analysis

AI is already improving productivity in tech services, which is expected to be between 15% to 45% across various service lines, but no one has cracked AI delivery at scale *(see Figure 2)*. Many providers have built AI platforms to deliver their offerings, but few have built a repeatable commercial model. Even when well developed, bringing these platforms to market remains a challenge: Frontline teams lack the confidence and capabilities necessary to deliver AI transformations for customers, making it difficult to convert offerings into funded programs.

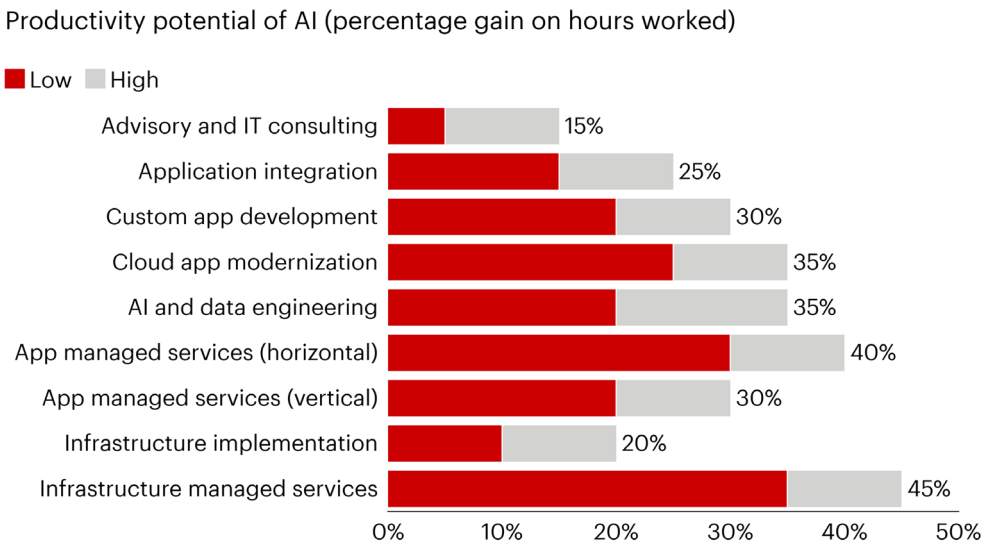
The market is still growing despite this readiness challenge: Firms such as Coforge, EXL, and Persistent delivered growth as high as 30% in 2025. But the market still cannot distinguish future winners and losers; that opportunity remains open.

Sustaining double-digit growth

Delivering double-digit growth in this environment requires more than an AI offering; the firms pulling ahead are changing how they compete and how they operate. While no company has perfected the model, a clear set of winning moves is emerging.

- **Win in specific segments, not across the board.** Generic AI capabilities are now table stakes. Leaders are focusing their investments on high-priority micro-verticals and defining specific battlegrounds at the intersection of spending priorities, micro-vertical, and geography. They then build offerings for that precise context. That depth makes a firm the obvious choice rather than one of five that can do something similar. A sustained Net Promoter Score® of 50 or higher is a strong

Figure 2: Productivity potential of AI varies across service lines



Source: Bain & Company

signal that the differentiation is real. Coforge, for example, grew 29% over its last fiscal year by combining deep domain expertise with more than 150 AI engagements focused on four industries: banking and financial services, insurance, travel, and healthcare.

- **Become a reinvention partner, not just a delivery partner.** Incumbent providers already possess an advantage that new entrants struggle to replicate: intimate knowledge of their customers' applications, data flows, integrations, business rules, exceptions, and technical debt. The firms that win will use those insights to identify and unlock value rather than waiting for a defined scope of work.

- **Build a go-to-market engine that sells AI transformation.** Traditional sales motions can produce interesting conversations about AI while gaining limited commercial traction. Leading firms are building frontline teams that combine knowledge about the customer and workflow with AI and technology capabilities to develop a winning sales proposition.

- **Make platform differentiation real.** Platform delivery no longer differentiates providers: Every firm now points to an orchestration layer, guardrails, and an agent library. In some industries, customers require service providers to join their platform rather than introducing another. True differentiation is more about depth and quality: the breadth and maturity of the agent library, how well it maps to real offerings, and whether it holds up at enterprise grade beyond the demo.

- **Evolve the commercial model.** Time and materials pricing is structurally misaligned with a platform delivery model. It rewards inputs, not outcomes. There is no single winning model yet, but outcome-based and gain-sharing structures are on the rise because they align better with incentives. The goal of any model must be to deliver real savings to the client and better margins for the provider.

- **Win with partners, don't just manage them.** Winners will need to rewire their partner model to be much broader than before, with partnerships that fit customers' varied AI journeys and priorities. The next era of partnerships with hyperscalers and frontier models will require both parties to work together to generate new opportunities rather than just fulfilling existing demand.

- **Transform talent and leadership.** AI is changing the talent model in many ways, including amplifying the productivity of entry-level work. New roles such as AI- and technology-proficient forward deployed engineers will become the delivery unit of record. Leadership needs to sell transformation outcomes, not technical capacity, which implies a fundamentally different talent strategy: career paths based on capabilities, pricing based on a collection of skills rather than roles, and hiring profiles that prioritize creative thinking and domain expertise over technical credentials.

Turning AI into competitive advantage

The market has already decided that AI will reshape the economics of tech services, but it hasn't determined the winners yet; that window is still open. Winning providers won't be those with just the best technology; they'll be the ones that can differentiate themselves with an AI-first operating model that delivers new business value for customers. Leaders are beginning to emerge, but the window to be one of them is narrowing. Several priorities stand out.

The market has already decided that AI will reshape the economics of tech services, but it hasn't determined the winners yet; that window is still open.

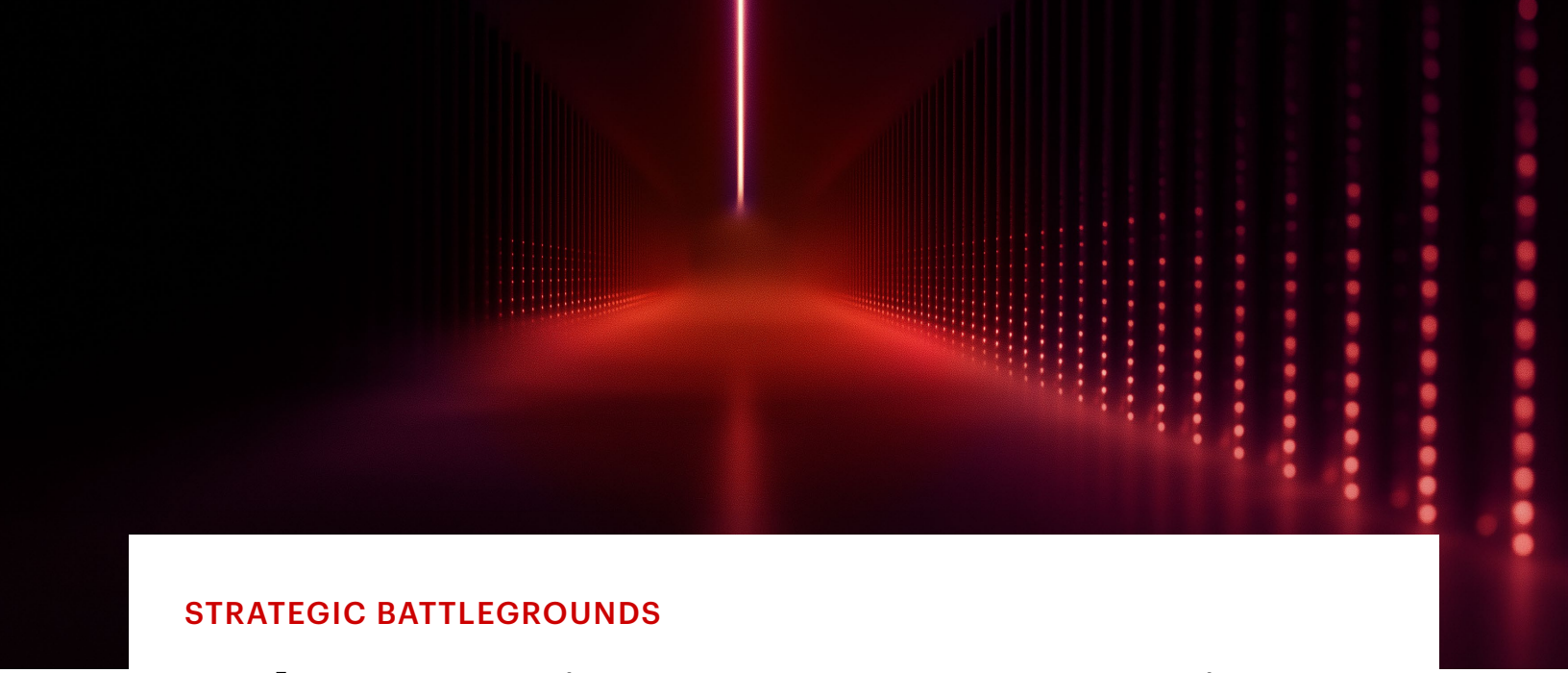
- **Focus where you can win.** Winning firms aren't spreading investment across every opportunity; they're concentrating resources on a few high-potential markets and building the capabilities and expertise to thrive in each. Good candidates identify their ideal client, develop proprietary offerings, and build a delivery model priced on outcomes.

- **Design the toolchain and the client experience together.** Most providers have built a few AI features, but few have created the full technical system that actually delivers results and a clear path for clients that moves them from their current way of working to the new one. The client's journey matters just as much as the technical tools that support it.

- **Reinvention means rebuilding delivery.** Leaders are redesigning how work gets done, including team structures, global delivery models, and pricing models, from the ground up; they're not wrapping AI tools around an existing time and materials model. The benchmark has moved, and productivity is already expected to improve by 20% to 35% compared with the current baseline. Firms still optimizing around the current model are solving the wrong problem.

- **Build a repeatable commercial model.** Focus on one high-potential opportunity, build it properly, and show that it delivers client savings and margin improvements. Use that as the template. Firms that try to scale up before proving the formula end up with neither.

AI is creating entirely new service categories, and the providers that establish credibility in these new markets are poised to lead them. Incumbents already have trusted client relationships, giving them a head start. The firms that move now have an opportunity not just to adapt to the next era of tech services but to define it.



STRATEGIC BATTLEGROUNDS

Cybersecurity's New AI Imperative: Attacking the Backlog of Vulnerability Alerts

How leading companies are responding to the deluge of warnings from AI-powered defenses.

By Syed Ali, Frank Ford, Matthew Leybold, Alexandra Juegelt, and Dana Aulanier

At a Glance

- As the rise of AI-powered hacking grabs the headlines, cybersecurity leaders are quietly starting to define best practices for deploying AI as a defensive tool.
- Leaders are initially strengthening their remediation operating model to handle the surge in vulnerability findings created by AI-powered threat scans.
- But this is just a first step toward a broader short-term response and a structural overhaul of capabilities in areas such as AI agent discovery, threat detection, and governance.

Most businesses will know that AI has made cyberattacks faster and more powerful. High-profile incidents, notably involving frontier AI model tests, have illustrated the pressure being put on organizational defenses that are designed to operate at the speed of human decisions, not the quicksilver rapidity of AI automation.

However, with emerging best practices in AI-powered defense gaining fewer headlines, executive teams might be unaware of the early progress that some companies have made in addressing the new threats, notably in vulnerability discovery.

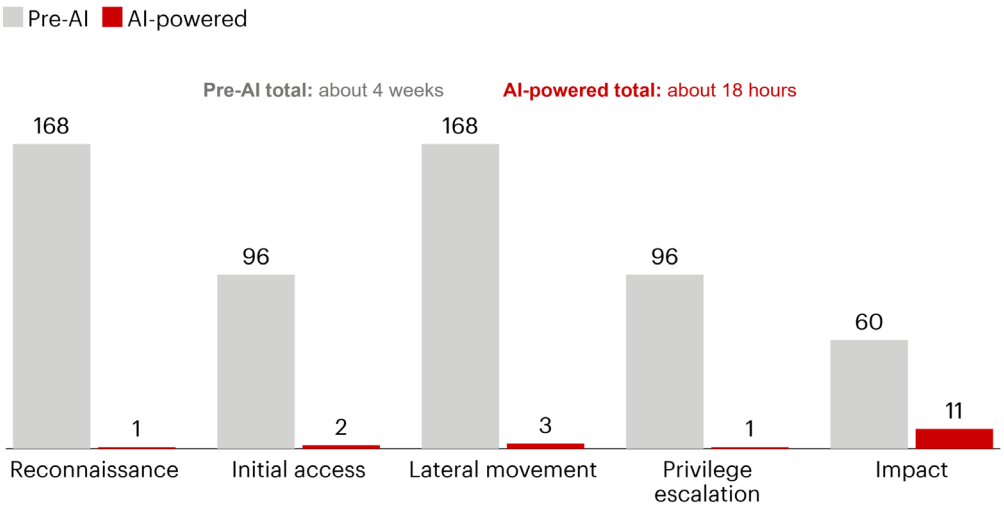
The danger at the heart of this evolution has not been exaggerated. Malign actors without coding expertise can now hack with the sophistication of a nation-state. Skilled hackers that used to spend weeks on a campaign now only need hours *(see Figure 1)*. Whereas hackers used to down tools to sleep, AI enables 24/7 fraud. Future attacks might not even need human intent amid signs that AI agents will sometimes act illicitly to fulfill a legitimate mandate.

Companies are feeling the impact. For one thing, the volume of phishing attempts has rocketed. Those attacks are now more convincing, too, through increased personalization of content and use of QR codes as a lure. AI-powered voice cloning and real-time video impersonation are also now occupational hazards for many businesses.

The adoption of AI agents has expanded the attack surface, too. Many companies would struggle to say how many agents they have, who owns them, and what they are permitted to do—let alone take the crucial step of installing a kill switch in each agent. Poor agentic housekeeping is a glaring weakness with an understandable cause: The lack of a silver-bullet solution from vendors is leading many firms to hesitate when they should be proactively creating a pragmatic and flexible solution through a build/buy/partner approach.

Figure 1: AI has massively compressed the time required by a typical cyberattack from an estimated 4 weeks to about 18 hours

Typical hours per attack stage



Source: Bain analysis

For their part, chief information security officers (CISOs) at large enterprises understand that AI-related risks require an overhaul of capabilities going well beyond vulnerability management to span AI workload and agentic discovery, hardening, monitoring, threat detection and response, governance, risk, and compliance. However, the scale of the vulnerability management challenge is currently all-consuming.

The CISOs we surveyed also highlighted supply chain risk as an issue. Companies must understand and control how vendors deploy AI through the whole life of a contract, including midcycle changes and fourth-party exposure. But with vendors shipping changes to models and software daily, oversight systems based on infrequent questionnaires aren't coping.

Companies instead need to start assessing third-party risk continuously, particularly with the vendors that are most integrated with their operations. Firms must tighten contracts to hear about and penalize breaches quicker, demanding preapproval of new AI deployment by vendors. At best, this is a work in progress for many.

How leaders are adapting cybersecurity to the AI era

Of course, AI is not just a tool for attackers. Some big companies are already deploying frontier AI at enterprise scale to strengthen defenses and accelerate their responses to machine speed. But that's creating its own challenge, with vulnerability alerts increasing by as much as eightfold. These surges can rapidly exceed a firm's ability to validate, prioritize, and fix weaknesses.

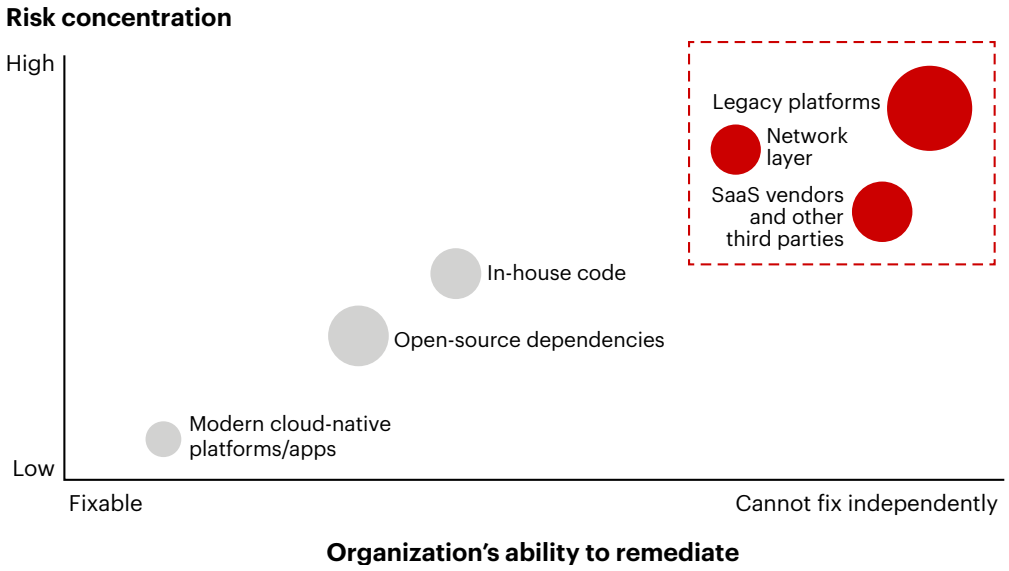
Some big companies are already deploying frontier AI at enterprise scale to strengthen defenses and accelerate their responses to machine speed.

While AI-enabled vulnerability management is still at an early stage, leading companies are making progress. Rather than analyzing vulnerabilities in a vacuum, these leaders are taking a more holistic view, emphasizing exploitability, reachability, systemic dependency, and blast radius.

They are strengthening their remediation operating model, too. Leaders have increased remediation budgets, typically by a double-digit percentage, while redirecting as much as 20% to 25% of their cybersecurity human resources from other work to remediate alerts from AI-powered scans (using third-party partners to maintain normal operations). One leader even tripled the size of its remediation team.

Other cybersecurity leaders are now pushing their DevOps teams to spend more time finding and mitigating vulnerabilities. For instance, the CISO at one large enterprise said they were now taking a

Figure 2: The most concentrated and hard-to-fix AI risks include those posed by legacy platforms and software-as-a-service vendors



Note: Bubble size reflects the number of chief information security officers who identified this as their most material unresolved exposure
Source: Bain analysis

harder line on enforcing an internal guideline that asks developers to spend around 30% of their time on proactive vulnerability management. Some companies have made board reporting more frequent. Accelerating critical patches is another focus.

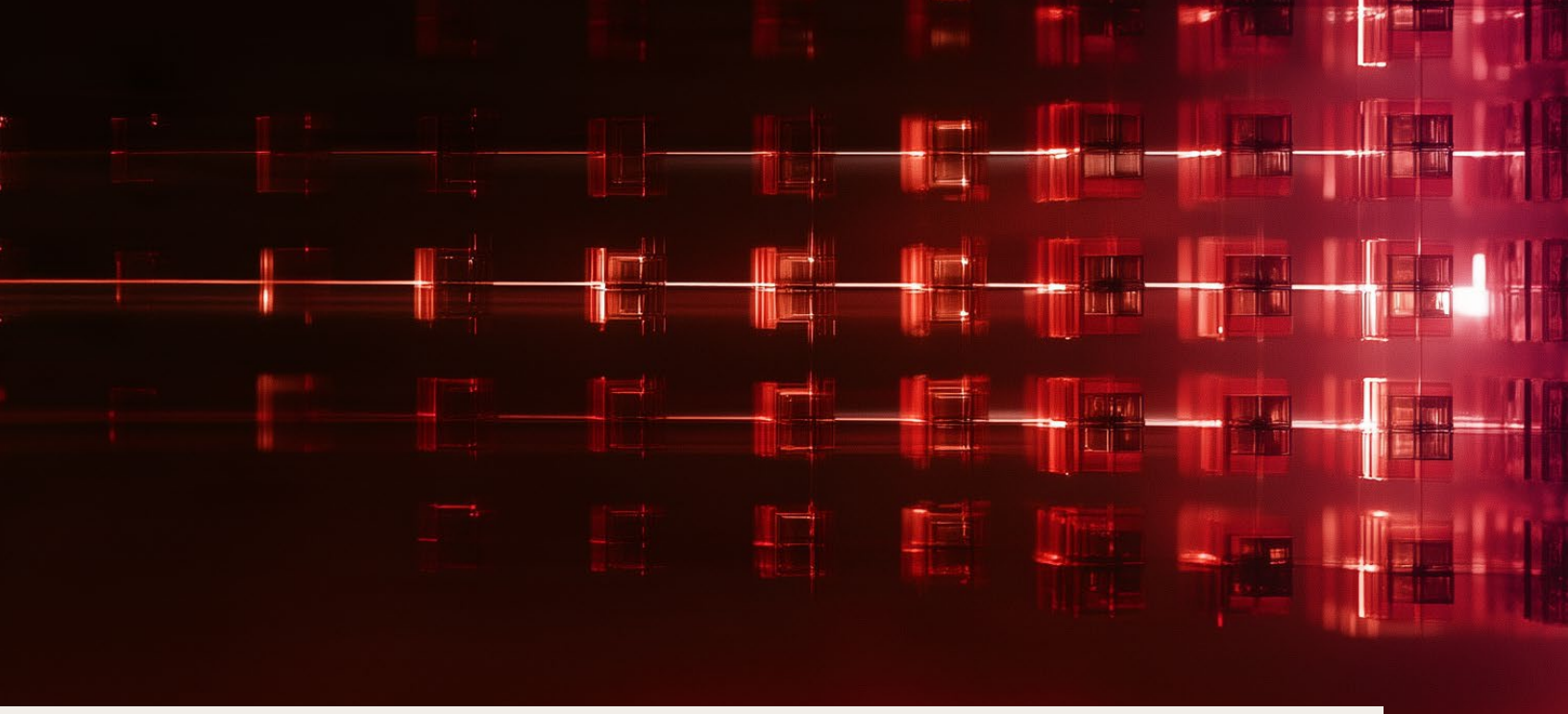
We also see a pattern emerging in the structural exposures that AI-powered cybersecurity leaders are prioritizing, with legacy platforms, network layers, and critical software-as-a-service (SaaS) vendors singled out for action ahead of other risks in areas such as in-house code and open-source dependencies *(see Figure 2)*.

For companies still formulating their AI-powered response to the recent turbocharging of cyberattacks, we see six key imperatives emerging from early-stage best practices.

- **Scan the full estate, not just the crown jewels.** More than half an organization's vulnerabilities are found outside critical applications. Legacy platforms, APIs, authentication systems, and dependencies should be scanned continuously, too.
- **Start building remediation capacity at the same time as scanning with AI for vulnerabilities at scale.** Prepare for the surge in alerts by automating simpler fixes (with a human in the loop) and using third-party support.

- **Treat technical debt as a live cyber exposure.** Where patching is not viable, companies need to be bold and strategic about retiring vulnerable technology rather than relying on rolling stopgaps.
- **Hold the chief information officer and chief technical officer accountable for the vulnerability debt.** And follow the cyber team's guidelines on DevOps resourcing to hit the backlog hard.
- **Give boards operational metrics, not abstract awareness briefings.** Vulnerability reports must be granular, actionable, and at least monthly (or even weekly). Direct board access for the CISO will help.
- **Start early, and budget as a multiyear investment.** Every month of delay gives rivals a head start. Similarly, executive teams should avoid underinvesting in talent while overinvesting in tooling.

While the news may have been dominated by the cybersecurity risks posed by frontier models, most companies should still be able to reap a defense dividend from AI advances, one that at least reduces the asymmetrical advantage recently gained by attackers.



Operational Transformations

The AI-Native Enterprise: Absorption Is the New Advantage 50

The Half-Finished Redesign:
How AI Reshapes Software Organizations 58

Managing Token Spending without Choking Off Opportunity . . .62

The Missing Architecture for
Agentic Software Development 67

How AI Is Changing Data Monetization 73

OPERATIONAL TRANSFORMATIONS

The AI-Native Enterprise: Absorption Is the New Advantage

As AI accelerates, the constraint is shifting from what the technology can do to how quickly companies can put it to work.

By David Crawford, Anne Hoecker, Jue Wang, and Chris McLaughlin

At a Glance

- Most enterprises are still deploying AI tools in narrow use cases, but leaders who treat AI as a full business transformation are seeing 10% to 25% EBITDA growth.
- The new competitive variable is absorption speed, the pace at which companies can put AI to work, and vendors are adopting forward-deployed engineering models to help enterprises break the bottleneck.
- Enterprises seek to operate with a portfolio of model options, using the lowest price and best-suited option for each use case. In the process, some profits may migrate from models to harnesses.
- The most prominent barriers to AI adoption today are organizational, not technical. Companies that redesign workflows are more likely to generate new value from their AI investment than those that deploy tools to individual employees.

By 2024, enterprise AI leaders had cracked the code on how to run AI transformations. They stopped treating AI as an IT experiment and instead ran business transformations with C-suite sponsorship and programmatic discipline. These leaders were rewarded with 10% to 25% EBITDA growth.

Figure 1: Leaders are reporting significant improvements from AI in well-architected transformations

New products and features	Customer service and contact center	Sales and marketing	Software and product development	Back office and operations
3x–4x increase in e-commerce revenue powered by AI features	**10x–100x** increase in response speed	**60%–70%** fewer people needed to produce each piece of content	**10x** faster prototyping	**30%** increase in procurement savings targets
15% revenue boost from AI-powered personalized recommendations	**More than 40%** reduction in contact center cost per interaction	**10%–30%** higher revenue from better conversion rates	**50%** faster on major refactoring tasks	**20%–30%** reduction in finance recurring cost base
	10%–30% reduction in call volume	**2x–3x** improvement in conversion rates from AI segmentation	**30%–50%** overall developer productivity boost	**50%** reduction in support headcount costs

Source: Bain & Company

However, our experience suggests that as much as 90% of enterprises today remain focused on tool deployment and narrow use cases, with little target or change-management discipline. This results in micro productivity but is mostly just "pave the goat trails" with little impact on revenue or earnings. These enterprises will eventually follow the taillights of the leading 10% in the coming years or risk being left behind.

The lesson from these transformations has been consistent: The most important work lies in process redesign and modernizing the data and application environment. Companies we work with have found that for every dollar spent on technology, four are spent on people and process. And because those investments are required regardless of how the technology evolves, waiting for the dust to settle is a losing strategy. The real advantage will come from making the organization transparent to and understood by machines and employees: its systems, standards, workflows, exceptions, customer knowledge, decision rights, and institutional judgment, without making all of that information indiscriminately accessible.

The evidence is difficult to ignore. Across customer service, sales and marketing, software development, operations, and the back office, well-designed transformations are delivering double-digit efficiency gains *(see Figure 1)*.

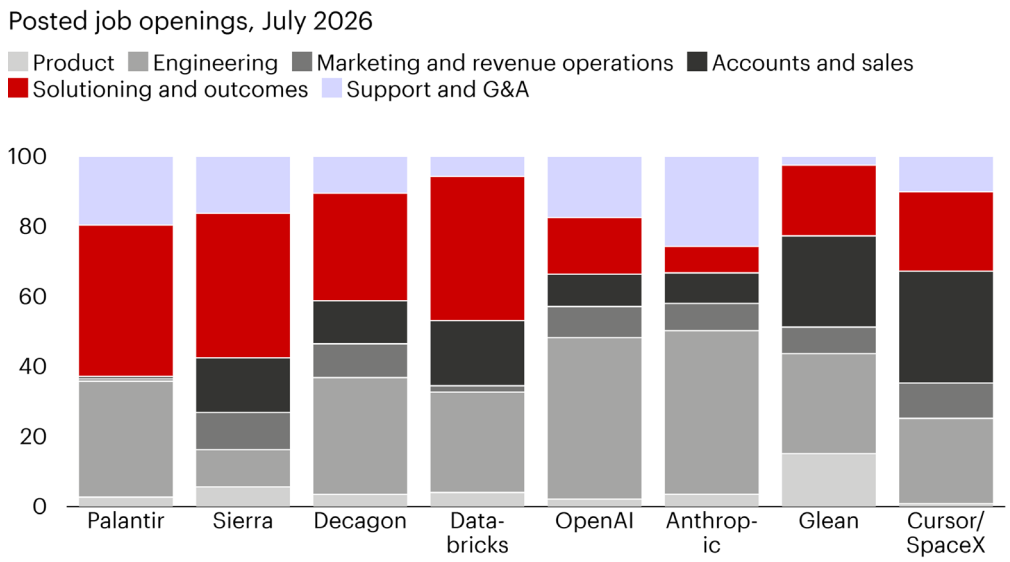
Absorption enablement is a new competitive battleground

While enterprises race to copy the winning business transformation playbook, the major labs and platforms are speeding up their release cycles faster than most enterprises can absorb them. That gap has turned enterprise absorption itself into a competitive variable. Recognizing this, vendors are racing to break the barrier, and that includes investing in the forward-deployed engineer (FDE) model leveraged by Palantir *(see Figure 2)*. Anthropic has committed roughly $1.5 billion of joint ventures to accelerate enterprise and midmarket adoption. OpenAI's DeployCo vehicle marshals more than $4 billion. Microsoft has committed $2.5 billion to set up Microsoft Frontier Company, a new operating business to embed engineering experts at customers' locations. Amazon Web Services is investing $1 billion to set up an FDE unit, and Google Cloud is investing $750 million in its new Gemini Enterprise transformation program, all with the goal of helping customers get measurable business outcomes.

The stack and the profit pool continue to evolve

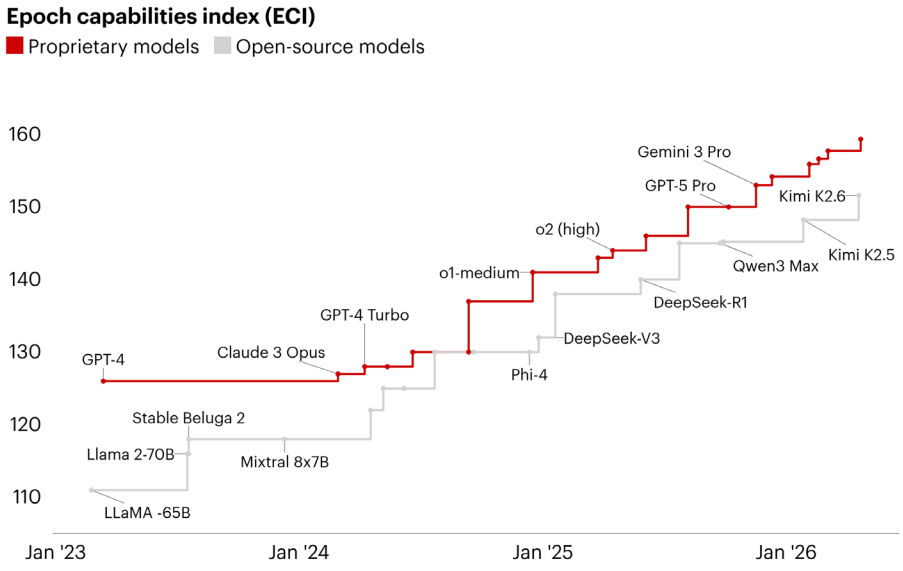
As use cases scale, enterprises move beyond their first wave of innovation and begin to optimize for greater efficacy and autonomy, continuous improvement, lower token costs, and stronger data security. At the same time, open-weight models have narrowed the performance gap with closed models for many enterprise use cases, expanding choice for enterprises and intensifying competition between models *(see Figure 3)*.

Figure 2: Firms are hiring roles designed to build and deliver AI outcomes



Notes: Revenue operations typically includes sales, marketing, and customer service; solutioning may include technical presales, forward-deployed engineers, and other delivery roles; G&A includes general and administrative roles
Source: Bain analysis

Figure 3: Open-source models are narrowing the capability gap with proprietary models



Note: Epoch Capabilities Index (ECI) is a composite metric aggregating results from multiple AI benchmarks into a single measure of overall capability
Source: Epoch AI

Some observers see this growing competition and falling token prices as indicators that the modal layer is commoditizing and its profits pools disappearing.

At Bain, we don't come to the same sweeping conclusion about commoditization, though we clearly see price pressure on mature model tokens and the growth of value in other layers of the stack (infrastructure, agent platforms, context, applications) for several reasons.

First, the boundary of the model will continue to evolve as capabilities in today's harness are incorporated into the model and run natively.

Second, leading model performance will remain valuable where intelligence matters most. Frontier models will tackle increasingly difficult problems that remain unsolved by humans, including disease treatments, fusion, and quantum gravity.

Third, in multistep agentic workflows, even modest differences in model accuracy can compound dramatically. For example, a 97% accurate model would deliver just 36% accuracy in a 34-step process vs. 71% accuracy for a model that is 99% accurate. That seemingly trivial 2% advantage can make the difference between a viable process and an unacceptable error rate.

We see the more likely scenario as a continuum of frontier and mature models. New and unproven use cases will initially favor frontier models, then likely be migrated to lower-cost alternatives as the use

cases mature. Frontier providers may consolidate as they push toward more fundamental challenges, capturing value where superior intelligence matters most. Meanwhile, a rich ecosystem of lower-cost, specialized, and optimized models will efficiently serve more common tasks, providing the services and support that enterprise customers need.

We may be headed into a segmentation between frontier models and trailing models or an expanding definition of models rather than a classic commoditization pattern.

In short, the industry structure is far from settled, and competitors are increasingly competing across multiple fronts. We may be headed into a segmentation between frontier models and trailing models or an expanding definition of models rather than a classic commoditization pattern.

Vendors are racing to sell the harness

As this debate rages, vendors aren't waiting; they're building out the application and infrastructure layers. Palantir's Alex Karp acknowledged the shift on an August 2026 earnings call when he said, "You own the weights, you own the alpha, you own everything.... We have a product that allows you to switch out models." Microsoft's Satya Nadella conveyed a similar sentiment on a July 2026 earnings call, saying that for customers, "[T]he goal is to have the firm be in control of their own destiny.... And the models are an input.... That means any given model at any given time is swappable."

These application layers act as harnesses that connect AI to the enterprise and turn model intelligence into business outcomes. Agents require context, data, memory, tools, permissions, routing, orchestration, evaluation, and learning loops. The harness determines how effectively an enterprise can put AI to work, operating over a pool of models and routing each request to the best-suited model that can meet the threshold for quality and correctness. How well an enterprise assembles this harness determines how fast it can absorb new AI capability and compound its gains.

HubSpot illustrates the potential. Its AI abstraction layer allows it to evaluate and swap models without rebuilding the underlying architecture while its proprietary customer context helps agents outperform generic models. The results are tangible: HubSpot's agents now resolve more than 70% of support tickets while its Prospecting Agent books nearly twice as many meetings as it did the previous year. The same architecture also helps HubSpot absorb new AI capabilities faster, accelerating product development.

No single vendor yet owns every layer, and open protocols such as Model Context Protocol and Agent2Agent give enterprises the flexibility to compose across them, which means the harness is becoming a strategic choice, not a vendor lock-in.

Enterprise AI leaders are compounding their gains with AI-native workflows

While vendors race to enable enterprise absorption, enterprise leaders are compounding their gains with another wave of business redesign. Having cleaned up their data and redesigned their first workflows, they are now building processes that resemble what an AI-native company would design from a blank sheet, unencumbered by legacy org charts.

While vendors race to enable enterprise absorption, enterprise leaders are compounding their gains with another wave of business redesign.

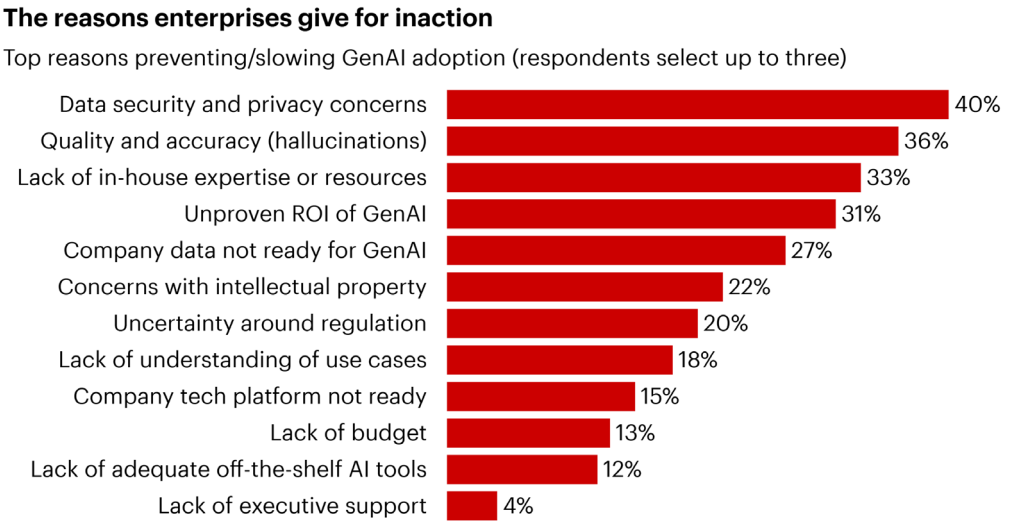
Consider Klarna's campaign production function: By deploying AI agents across image generation and creative development, the company compressed what had previously taken six weeks down to seven days while simultaneously increasing output volume. Weekly updates to images aligned with faster-paced retail events and created a better customer experience. The work moved from getting ready to getting better.

Some companies are redesigning their product launch processes to be AI native. Instead of automating the existing roles across design, marketing, and sales, they condense the process around three senior leaders and a suite of agents that specialize in capabilities such as communication, competitive analysis, channel analysis, and sell-through analysis. Among the teams that were engaged in multi-month preparations for launch, the emphasis now moves to post-launch iteration based on continuous evaluation and improvement. The three leaders presiding over the product launch process spend less time on mundane analytics and troubleshooting and more time reviewing markdown files and iterating to improve the product launch's effectiveness.

Executives blame the tech, but organizational barriers matter

Leading enterprises have charted a clear roadmap and set of best practices for deploying AI in ways that pay back the investment and deliver competitive advantage. Even so, most enterprises remain in the pilot phase of AI absorption. Asked what holds them back, they point mostly to technical and talent worries: data security, hallucinations, unproven ROI *(see Figure 4)*. Our work with companies suggests that the real reasons for slow progress are broader and better represented by those found in a 2025 study by the Massachusetts Institute of Technology. Organizational barriers such as challenging change management, tentative executive sponsorship, poor user experience, and a simple unwillingness to adopt new tools.

Figure 4: Enterprises blame technology, but the real barriers are organizational



Source: Bain GenAI Survey 2026

The next era of AI is when companies begin turning the technology into a true competitive advantage. It's defined by blank-slate redesigns of processes and organizations, and it's powered by continuously improving hybrid human-agent workflows (see the Bain Brief "Proprietary Intelligence: How to Win with AI").

For CEOs looking to accelerate their AI programs, the starting point is clarity on the stakes and the ambition. Leaders need a rigorous view of where AI can create the most value, where it could reshape competitive advantage, and where inaction could leave the business vulnerable. From there, they can set ambitions bold enough to force genuine redesign rather than incremental improvement, concentrate leadership attention on the few domains in which AI-native ways of working could create breakaway advantage, and establish clear ownership for delivering results. The goal is not just to do more with AI but to make deliberate choices about where AI can fundamentally change the trajectory of the business.

To help executive teams move in the right direction, several imperatives stand out.

- **Master AI absorption as a competitive advantage by committing to the discipline of an AI-enabled workflow transformation.** AI capabilities are being released faster than any organization can naturally absorb them. This is difficult work, but the playbook is well understood. The companies compounding their gains aren't waiting for a stable platform; they are actively redesigning workflows and organizations in parallel with the technology's evolution. Every quarter spent in pilot mode is a quarter your competitors are spending on redesign.

- **Don't let the harness debate stall you.** Where the LLM harness boundary settles has more to do with vendors' interests than yours. It will keep moving, and no viable vendor will build a business on stealing customer IP.
- **Design your AI to be a learning system.** Organizations that design their workflows to capture and recirculate operational data automatically, without waiting for human review cycles, create a compounding advantage that pure efficiency gains cannot replicate. The goal is not a faster process; it is a process that gets faster on its own.

The companies pulling ahead are not the ones with the best models; they are the ones that have learned to absorb and act faster than their competitors. Waiting for the dust to settle is the one strategy guaranteed to ensure it never does.

OPERATIONAL TRANSFORMATIONS

The Half-Finished Redesign: How AI Reshapes Software Organizations

Most companies have changed the easy things and left the hard ones untouched.

By Jonathan Frick, Purna Doddapaneni, and Arun Ganti

At a Glance

- AI is reshaping software organizations with smaller teams, evolving roles, and a growing emphasis on judgment over coding.
- Most companies have redesigned team structures but have yet to change decision making and accountability, leaving major bottlenecks in place.
- The winners will be organizations that pair AI with faster decisions, stronger talent development, and operating models built for continuous change.

Everyone expected AI to reshape software organizations by enabling smaller teams, shifting people from doing work to directing agents, flattening organizational structures, and pushing decisions to the edge (see the Bain Brief "How Will AI Change Software Organizations?"). Those changes are underway, but what has been less expected is the speed, scale, and unevenness. Companies have redesigned team structure, redefined roles, and invested faster than during previous operating model shifts. But most companies have addressed only part of the challenge. Without continued effort to reorganize work and how decisions are made, bottlenecks will continue to hold companies back.

The organization has already changed

The pace of change is already evident: Bain's recent survey of software companies found significant shifts in team structures, hiring priorities, and ways of working over the past two years.

Fewer engineering organizations report traditional pyramid structures, down from 66% to 29%. Fluency with AI tools is now the leading capability sought out in engineers; cited by 62% of survey respondents, it is three times higher than raw coding skill. Judgment and problem solving also rank highly, reflecting the growing importance of directing agents.

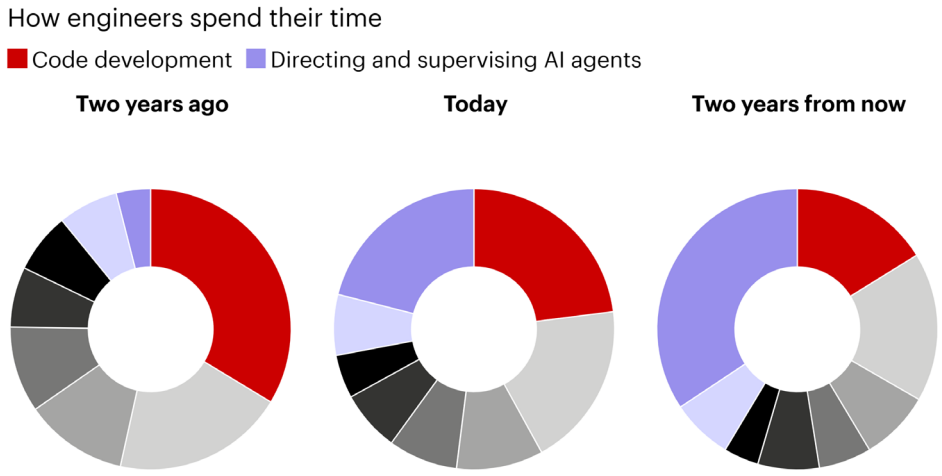
The nature of software development is changing rapidly, too. Developers expect to spend a third of their time directing AI agents within two years, up from almost none two years ago *(see Figure 1)*. Job boundaries are blurring, and new hybrid roles are emerging as developers take on testing and deployment, product managers build prototypes and contribute to marketing, and designers write front-end code *(see Figure 2)*.

Teams have shrunk to pods of three to five people, with routine execution tasks fading away to be replaced by judgment and management skills often found in more senior roles.

The frontier shows the end state

Native AI companies offer some clues about where this leads since they never had to unwind a legacy organization. Their advantage didn't start with AI but with good fundamentals: strong test automation, clean requirements, disciplined site reliability, and a culture in which the team that builds something also

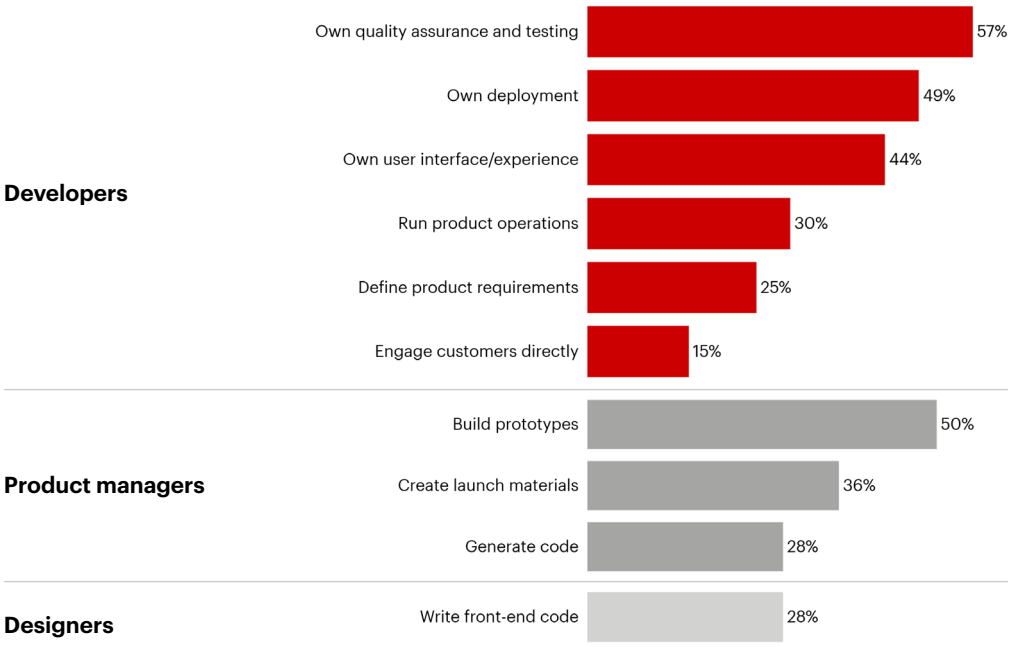
Figure 1: Developers are spending less time coding, more time directing agents



Source: Bain AI Org Assessment Survey, 2026 (n=155)

Figure 2: Developers are absorbing the most adjacent work

Percentage of companies that say employees in these roles are taking on these adjacent activities "often" or "always"



Source: Bain AI Org Assessment Survey, 2026 (n=155)

runs it. Only then did they add agents on top. Companies that skip the basics and reach straight for agentic AI merely exacerbate the problems they already have. But when real change is successful, the numbers are impressive. For example, AI-native marketing organizations generate several times the revenue per employee of traditional software companies by organizing work into small pods, with an embedded engineer and agents doing work that once required entire departments.

Of course, knowing the destination isn't the same as knowing how to reach it. For most companies, the journey starts with a few provocative questions.

What shape does the organization settle into? The pyramid is flattening, but the successor is uncertain. The shape depends on what the company is solving for. Some amplify capacity while leaving size and roles largely intact. Some strip out management layers to hasten decision making. Some cut pod size sharply as agents absorb much of the work, leaving a more senior team behind. Others rebuild entirely around small, multidisciplinary teams of generalists. No single answer suits all.

Where does the next generation of senior talent come from? The base of the pyramid is disappearing faster than anyone expected. Junior engineers used to make up one-third of teams on average, and now they're less than a fifth while senior ranks swell. The routine work now performed by AI is the same work

on which junior engineers once learned. Companies need more senior judgment than ever, yet they are quietly dismantling the path that produced it. Competitive advantage will go to companies that can compress years of judgment into a fraction of the time by teaching junior engineers how to direct agents. A few already hire this way, recruiting for the ability to talk to customers, solve unfamiliar problems, and build, betting on judgment over pedigree.

Who decides now that the chart is flat? More than any other, this is the change that companies have skipped, possibly because this is less about tooling and more about redistributing power. Seventy percent still make decisions centrally, and that applies across different levels of AI maturity (see the Bain Brief "Redesigning Tech Company Operating Models for an AI-Accelerated World"). It's also the change that matters most because speeding up engineering is wasted if the rest of the organization can't act on the output. Code ships faster than ever, but the approvals, handoffs, and escalations around it haven't moved. So, the gains stall before they reach a customer. The fix isn't another reorganization but rather in distributing decision rights, defining where humans decide, where agents act, and who owns the outcome. One AI-native company lets the engineers building a feature launch its marketing themselves rather than routing it through a central team so that go-to-market moves as fast as the code. Structure changed first. Accountability has to follow, or the flatter organization simply relocates old bottlenecks.

Most companies have completed the visible redesign: Teams are smaller, roles are changing, and AI is rewriting software development. The next phase will determine how work gets done.

What leaders should do now

Most companies have completed the visible redesign: Teams are smaller, roles are changing, and AI is rewriting software development. The next phase will determine how work gets done. Competitive advantage will come less from adopting AI than from building an organization that can fully capitalize on it, with faster decisions, clear accountability, and a talent model that values human judgment over routine execution. Senior executives may be pleased with the progress they've achieved, but the harder and more valuable half of the redesign is still a work in progress.

OPERATIONAL TRANSFORMATIONS

Managing Token Spending without Choking Off Opportunity

Most of the cost risk comes from a handful of workflows; govern those tightly, and get out of the way almost everywhere else.

By Jue Wang, Anne Hoecker, Alec Koh, Chris McLaughlin, and Vito Trinchera

At a Glance

- As AI adoption grows, companies often respond to rising costs with user-level limits, but these controls can curb valuable experimentation and adoption.
- Companies should manage AI spending by workflow, moderating low-value use cases, unleashing high-value ones, and ring-fencing risky cases.
- Measuring cost and value by workflow helps companies manage spending more effectively, channeling investment to where it creates the most value.

The first time a company receives a startling AI bill, the instinct is almost always the same: Put limits on usage. Many companies set up per-person budgets and implement approval processes to clamp down on costs before they spiral out of control.

The reaction seems to make sense, but it solves the wrong problem. Companies can become so concerned about controlling costs that they restrict the very behaviors they're trying to encourage. Developers ration AI use, sales teams stop experimenting, and leaders worry more about the bill than the value.

Our view is that the bigger problem right now in most organizations isn't runaway token spending, but too much spending in the wrong places. Most companies should be spending tokens across more of their organization to radically change their business. Across organizations, underuse is the bigger problem. One global technology company set a soft cap of $500 per month for each developer to help encourage AI use; the average cost landed around $200. At another tech company, engineering leaders were metered so cautiously that they left their allocated $40-per-person monthly budget unspent, never touching the buffer pool sitting behind it. Leadership eventually told them to stop metering and go spend.

But this restraint is often overshadowed by a few big spenders within a small number of AI workflows that are not spread evenly across thousands of users. A lack of transparency limits the ability to measure returns and channel spending to the places where it will generate the most value. For one widely deployed knowledge-assistant tool, nearly all tokens are cached input tokens (embedded and reused in API requests) that a company can't easily see or forecast.

Managing token spending effectively requires precision, not blanket controls.

Manage workflows, not users

The right question isn't "how much should each employee spend?" It's "which workflows deserve more AI, and which deserve less?"

Framing the question in such a way leads to three distinct approaches, each with its proper place in the company *(see Figure 1)*.

Figure 1: Three distinct approaches to token spending play out differently across functions

	Moderate	Unleash	Ring-fence
Marketing	Generates structured briefs as well as brand and compliance guideline checks	Sets campaign strategy, developing customer and market insights	Runs open-ended campaign-optimization loops
Sales	Sharpens client emails	Creates the killer pitch via positioning and pricing	Works on account and territory optimization
Software R&D	Summarizes code and documentation	Performs day-to-day development in a governed harness	Enables full codebase migrations and refactors as well as overnight optimization loops

Source: Bain & Company

- **Moderate:** Some activities don't require frontier AI models or even an agent. Routine email drafting, document summaries, or straightforward information retrieval can often be handled by cheaper models or without an LLM at all. Open-weight models can reduce costs for such routine tasks since they can be self-hosted or run at significantly lower per-token rates than frontier APIs. The goal is to reduce unnecessary spending without sacrificing quality. But it's difficult to enforce routing to cheaper options when the frontier model is one slash command away.

- **Unleash:** Most day-to-day work falls into this category. Software development, sales preparation, customer service, and many structured business processes generate significant value from AI while costs remain predictable. Here, organizations should encourage adoption rather than constrain it. The goal shouldn't be minimizing token spending but maximizing productivity.

- **Ring-fence:** A small number of workflows require much tighter oversight. They share traits: Open-ended optimization, autonomous research loops, massive security scans, and persistent AI agents can consume large numbers of tokens without clear stopping points. These should be tested in controlled environments, instrumented carefully, and equipped with safeguards before being deployed at scale.

The principle is simple: Govern the risky tail tightly, and get out of the way almost everywhere else.

Measure cost by workflow

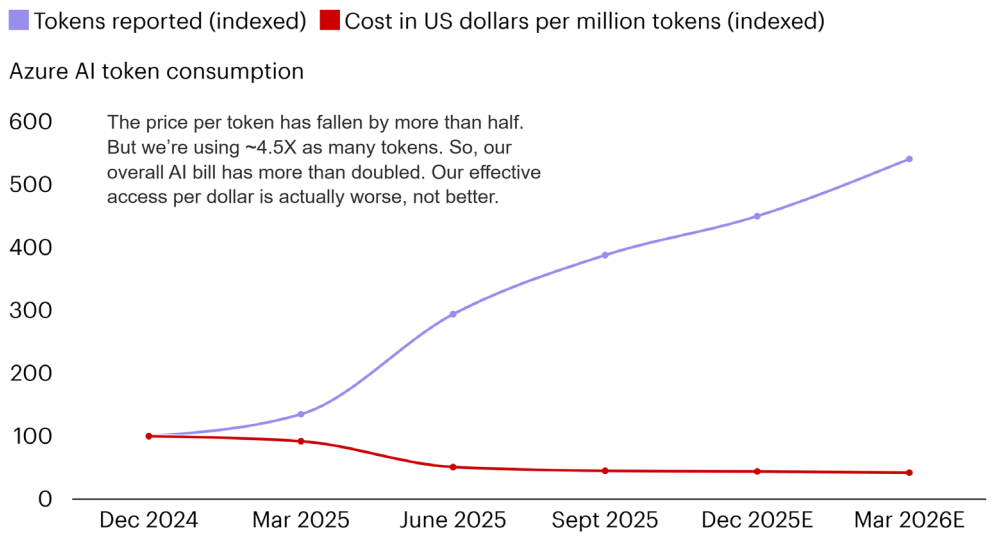
This approach only works if organizations understand what each workflow actually costs and what it's worth. Most companies track AI spending at the user or business unit level, but that doesn't show where value is being created. The more useful question is "what does it cost to complete a proposal, resolve a customer issue, review a contract, or generate production-ready code?"

The principle is simple: Govern the risky tail tightly, and get out of the way almost everywhere else.

First versions of agentic workflows often cost far more than expected because prompts, orchestration, and model choices haven't been optimized yet. In our experience, a new workflow might cost 10 times more than it ultimately should. Once teams understand the economics, they can often remove most of that excess through better architecture, smarter model selection, and more deterministic workflow design.

As token prices declined, companies expected their costs to follow. But most companies typically upgrade to the latest frontier models, agents consume more tokens as they tackle increasingly sophisticated tasks, and adoption expands into more workflows. As a result, lower per-token costs are offset by higher usage,

Figure 2: Average cost per token is falling, but token use is rising faster



Sources: Wells Fargo Securities, October 2025; Bain analysis

leaving overall cost per task stubbornly high *(see Figure 2)*. The expectation is that these economics will eventually improve, but the anticipated cost inflection point has yet to materialize (for more, read the Bain Brief "How Token Economics Will Change Opex").

Even so, the right strategy isn't just to use cheaper models; it's to reserve expensive reasoning for the parts of a workflow that genuinely require it while making the rest more structured and predictable.

Govern selectively

This also changes how organizations should govern AI spending. Most workflows belong with the business. Engineering leaders should manage engineering workflows, and sales leaders should manage sales workflows. They're best positioned to understand where AI creates value, and they can encourage adoption without unnecessary bureaucracy (for more, read the Bain Brief "Mobilizing the Organization in the Agent Economy").

Only the highest-risk workflows require central oversight. A small governance team can establish standards for testing, monitoring, and cost controls while ensuring that open-ended agentic systems include appropriate safeguards before they are deployed broadly. Where possible, those controls should be built into the core AI platforms themselves, making safe, cost-effective behavior the default rather than relying on users to follow policy.

The same principle applies to spending limits. Low caps often create exactly the wrong behavior, discouraging experimentation and slowing adoption. High soft limits, combined with clear visibility into

unusual consumption, encourage teams to use AI where it creates value while still protecting against genuinely excessive costs.

Prepare for the next wave

Most organizations are still in the early stages of AI adoption. As autonomous agents become more capable, AI consumption is likely to increase dramatically. Persistent agents, long-running optimization tasks, and increasingly sophisticated workflows will all drive significantly higher token usage than today's interactive assistants.

Most organizations are still in the early stages of AI adoption. As autonomous agents become more capable, AI consumption is likely to increase dramatically.

That makes it even more important to establish the right management approach now. Organizations that rely on blunt cost controls will struggle to capture AI's full value. Those that understand the economics of their workflows will be able to expand adoption confidently while managing the relatively small number of use cases in which costs can escalate quickly.

OPERATIONAL TRANSFORMATIONS

The Missing Architecture for Agentic Software Development

Most companies have the AI tools, but few have built the engineering systems that make those tools effective.

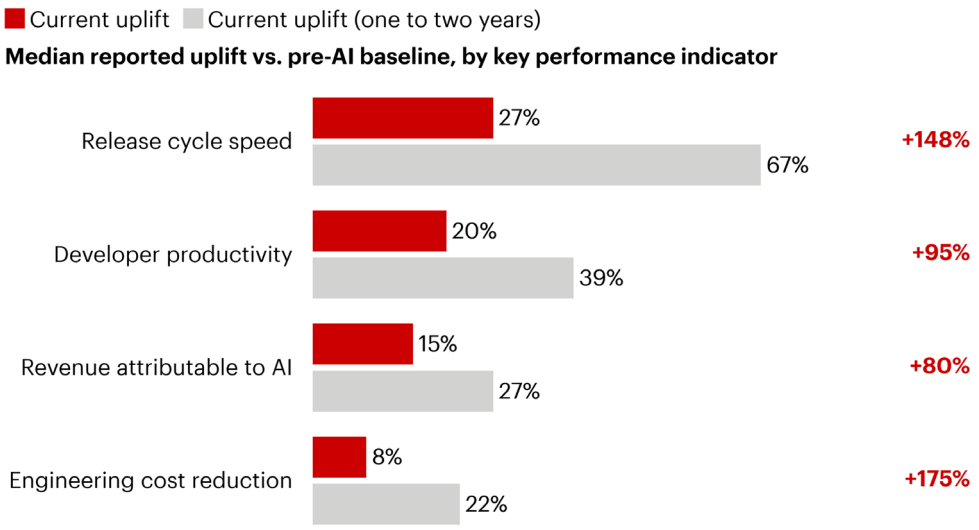
By Purna Doddapaneni, Chris McLaughlin, Stanley Liu, and Bill Radzevych

At a Glance

- Deploying the latest frontier models is no longer enough: AI leadership results from developing an architecture that enables AI to deliver against expectations.
- AI has sped up coding, but that has shifted bottlenecks to review, governance, and software delivery.
- The companies capturing the largest gains make engineering knowledge machine-readable for agents, automate quality verification, and treat the software development life cycle as a continuously improved product.
- As AI compresses the cost and time for software creation, competitive advantage will belong to organizations that redesign the systems around AI.

Over the past three years, the binding constraint on AI-assisted software development was model capability, but that has changed. Frontier models can now reason, generate code, and execute increasingly complex development tasks. The limiting factor has shifted from the model to the connective architecture around it. Most companies haven't built that architecture yet.

Figure 1: The gap between current gains and future expectations is widening



Notes: Median values reported; uplift percentage shows expected increase over current gains
Source: Bain Tech and Engineering Survey, April 2026 (n=293)

Naturally, expectations have risen: In Bain's latest Tech and Engineering Survey of 293 senior technology leaders, respondents anticipate a 148% improvement in release-cycle speed and a 95% uplift in software developer productivity within the next one to two years. That's much higher than the productivity gains most are capturing today: between 20% and 27% across key metrics. But the gap isn't due to technical limitations *(see Figure 1)*.

The organizations that are pulling ahead are distinguished less by the models they use than by the engineering systems those models operate within. Most software organizations have access to good AI tools now; very few have built engineering environments that can deliver against expectations. This architecture has three parts: knowledge that agents can consume, a harness that enforces quality deterministically, and the discipline to treat the software development life cycle (SDLC) itself as a product.

Where the bottleneck has moved

The pattern is now well established: AI accelerates individual coding tasks, but the bottleneck shifts to everything around the code. Our survey found that developers complete roughly 21% more tasks but review time rises approximately 91%, highlighting human review as the critical constraint. Developers are increasingly acting as orchestrators rather than producers, managing 47% more concurrent workstreams as AI generates work faster than surrounding processes can absorb it.

This is the shifting bottleneck paradox we first described in 2024 (see the Bain Brief "Beyond Code Generation: More Efficient Software Development"). Companies that speed up code generation without redesigning the surrounding process don't compound their gains; they redistribute their pain. Every bottleneck AI removes exposes the next one, and the system reverts to the throughput of its slowest human checkpoint.

Once AI compresses coding time, the focus for improving software delivery shifts from maximizing developer productivity to minimizing the friction across the entire development life cycle.

The three parts of the missing architecture

If the SDLC has become the new unit of optimization, then its architecture has to be built deliberately; it won't emerge on its own. The organizations realizing the largest gains share three characteristics: They have structured the knowledge on which AI agents depend, they engineer quality and trust into the workflow, and they manage the SDLC as a continuously improving product.

The organizations realizing the largest gains share three characteristics: They have structured the knowledge on which AI agents depend, they engineer quality and trust into the workflow, and they manage the SDLC as a continuously improving product.

Make engineering knowledge accessible to agents. Humans can infer, guess, and make reasonable assumptions, but agents need explicit organizational knowledge.

So how do successful AI-native development organizations capture the most relevant knowledge of the engineering team and put it into forms that agents can understand and use? Think about how a senior engineer brings a new hire up to speed: explaining why the code base is structured the way it is, which conventions to follow, when to make a judgment call, and when to escalate. Today, that knowledge mostly lives in people's heads and surfaces through conversation and code review.

Making engineering knowledge explicit and machine-readable means capturing it in structured specifications that serve as a single source of truth; in documentation that codifies conventions, decisions, and escalation paths; and in well-defined interfaces among planning, building, and testing. Within this knowledge architecture, agents operate with clear inputs and verifiable outputs. It's like onboarding agents the way one would onboard a junior engineer, except it's done once, in writing, and every agent inherits it.

Stripe took this approach when it built "minions," agents with well-defined tasks, such as writing unit tests, fixing linter warnings, and migrating code to a new API version. Each agent has a spec defining its objective, scope, relevant context, verification method, and constraints, with 1,300 AI-authored pull requests merged per week across its engineering organization, all subject to the same human review process as any other change. Stripe accomplished this not by making its agents more powerful but by making the knowledge surrounding each task more precise. The specification, not the model, is the leverage.

Automate confidence. When an AI agent can generate code in seconds, the question is no longer "how fast can we build?" It becomes "how fast can we verify?" And the answer can't be "we'll have humans check everything." As AI output scales, manual inspection quickly becomes the new bottleneck.

Leading organizations solve this by automating confidence. Rather than relying on humans to catch every issue, they build a deterministic quality layer or harness that validates agent-generated work against codified standards before anything reaches production. Human judgment doesn't disappear, but it shifts to the exceptions, edge cases, and higher-risk decisions that automation can't resolve.

This harness has a risk-scoring engine that evaluates the blast radius of each proposed change. Policy-as-code guardrails enforce security, compliance, and architectural standards automatically. Evaluation frameworks test, validate, and iterate until outputs meet predefined acceptance criteria. Audit trails capture every agent action, decision, and escalation for traceability. The result is a system that can safely absorb AI-generated work at a pace no human review process could sustain.

Companies implement this principle in various ways, from traditional stage-gated workflows with automated quality gates to risk-based models that route higher-risk changes for human approval. The right choice depends on regulatory environments, risk tolerance, and product complexity. The core principle is the same across models: Automate routine verification so that humans can focus on decisions that require judgment.

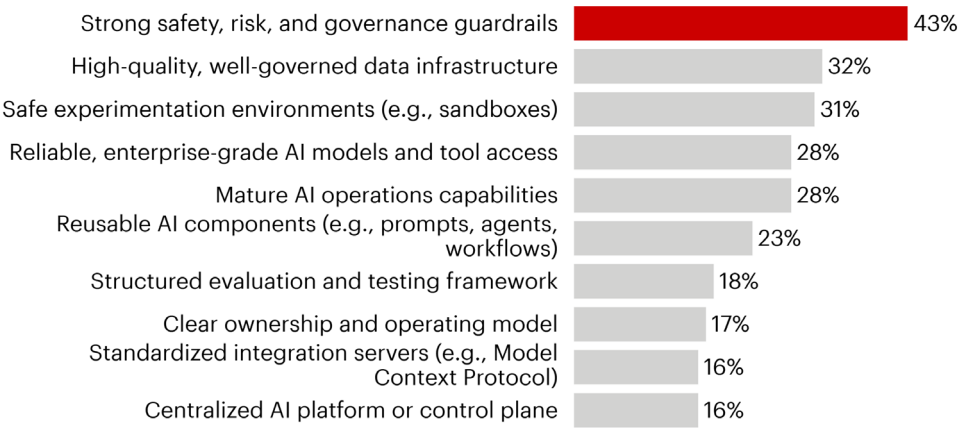
Our survey found that 41% of companies expect to deploy a risk-tiered model while only 6% envision fully autonomous development. More importantly, respondents consistently identified governance, safety, and risk controls as the most important enablers for scaling agentic AI, more than model quality, data infrastructure, or platform investment *(see Figure 2)*. As model capabilities continue to converge, competitive advantage will come less from generating more code and more from generating confidence in that code.

Run the SDLC as a product. Most companies still treat their development process as a fixed backdrop, the stage on which engineering performs. Leaders treat it as a product to be continuously designed, measured, and improved.

Treating the SDLC as a product changes how work flows, how teams are organized, and how performance is measured. Traditional handoffs among product, engineering, quality assurance, and deployment break down when an agent can take a requirement from spec to working code that passes tests in a single workflow. Product and engineering increasingly operate as an integrated system. In our survey,

Figure 2: Guardrails and governance, not tooling, are the top enablers for scaling agentic AI

"Which is most important to have in place when scaling agentic AI?" (percentage of responses)



Notes: Multi-input question; respondents could select multiple options; top 10 of 15 options shown
Source: Bain Tech and Engineering Survey, April 2026 (n=293)

respondents anticipate smaller teams, with nearly a fivefold increase in agents per team and a continued shift of human work up the stack, from execution to intent definition and quality governance.

Measurement must evolve, too. Metrics designed for human coding, such as those by Google's DevOps Research and Assessment (DORA), remain important, but they are no longer enough. In an agentic world, velocity alone can be misleading: Teams can ship faster while introducing more defects, more technical debt, and more risk. Organizations need metrics that distinguish human and agent contributions, track end-to-end system performance, and elevate risk and control to first-class scorecard dimensions alongside speed and quality.

Organizations that treat SDLC as a product give it everything successful products have: an owner, a roadmap, customer feedback, continuous iteration, and clear measures of success. They continuously improve the system that produces software, not just the software itself. As AI accelerates execution, that system increasingly becomes the source of competitive advantage.

The next redesign begins now

The organizations pulling ahead aren't simply deploying better AI models; they're redesigning the system within which those models operate. At Amazon, a mandate to embed AI into every development system has enabled up to 70% of code reviews to be handled by AI. Customer validation cycles that previously took weeks now take days. At Craft Docs, a 20-person engineering team adopted a shared agentic

platform with governance controls that helped reduce process friction, advancing from 15 to 20 issues per week to more than 100. Both companies invested in underlying foundations: structured knowledge that agents can reliably consume, automated quality controls that establish trust without relying on human review, and an engineering operating model designed for continuous improvement.

For technology leaders, this marks an important shift. The first wave of generative AI was largely about tools; the next is about operating models. As AI compresses the cost and time of software creation, competitive advantage moves to the systems that surround it: how knowledge is captured, how quality is verified, how decisions are made, and how the SDLC itself is managed. Organizations that continue to layer AI onto legacy ways of working will realize productivity gains, but they're unlikely to achieve the transformational improvements they expect.

Organizations that continue to layer AI onto legacy ways of working will realize productivity gains, but they're unlikely to achieve the transformational improvements they expect.

There's still time to make this transition. While nearly three-quarters of technology leaders now view AI as their most important lever for engineering productivity, relatively few consider themselves AI-first today. The organizations that close that gap won't be those with privileged access to better models; they will be those that build the capabilities to turn AI into velocity, reliably and safely and at scale. In the years ahead, the defining question won't be what model to deploy but whether you've built the architecture in which AI can thrive.

OPERATIONAL TRANSFORMATIONS

How AI Is Changing Data Monetization

As AI agents reshape enterprise software, companies must take a fresh look at what data to expose, how to provide access, and how to price it.

By Eric Maltiel, Josh Sandberg, Justin Murphy, and Mukarram Bhaiji

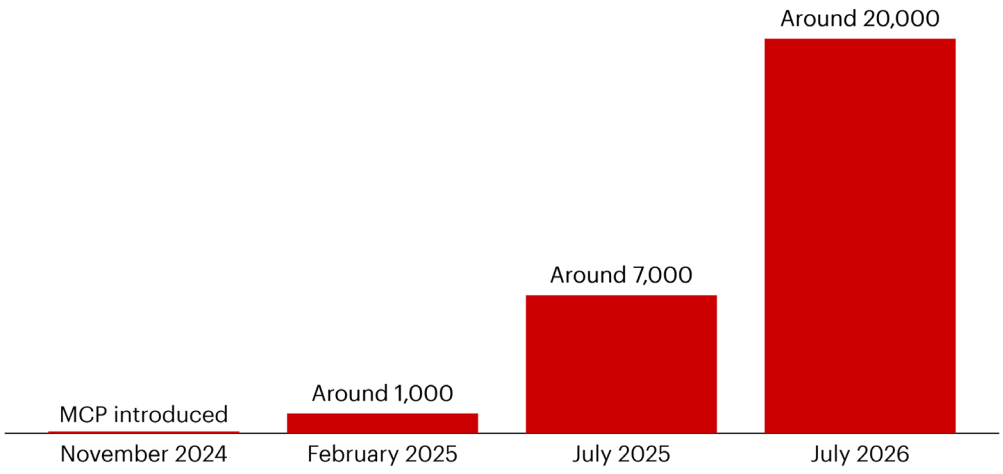
At a Glance

- Model Context Protocol (MCP) is emerging as a new standard that allows AI agents to access enterprise software and data, prompting companies to make some critical decisions.
- Deciding what data to expose means weighing the risk of opening up valuable data against the risk that customers and AI agents bypass your platform in favor of more accessible alternatives.
- Deciding whether to use MCP means balancing easier integration into agent workflows against the potential for customers to accomplish more without ever engaging with your interface.
- Deciding how to monetize comes last, with pricing designed to capture the value of agent access without cannibalizing higher-value revenue elsewhere in the portfolio.

Model Context Protocol (MCP), an emerging standard that lets AI agents securely connect to enterprise software and data, has gone from a niche protocol to a de facto standard at remarkable speed, growing from its late 2024 launch to more than 20,000 servers less than two years later *(see Figure 1)*.

Figure 1: Model Context Protocol server growth signals its rise as the new default for agentic access to data

Model Context Protocol (MCP) servers published and available for developers



Sources: MCP.so (a server registry); Glama.ai; PulseMCP

Vendors are understandably asking how to charge for AI agents that access their data and workflows through MCP. But pricing should be the third question, not the first. Before deciding how to monetize, software companies need to first answer two more fundamental questions:

- What data should we expose programmatically?
- Should that access be provided through MCP?

Only then should they decide how to monetize it.

Those first two decisions determine a company's strategic position in the agent era. Pricing should rest on this foundation.

Decide what data to expose

This is the hardest decision because exposing data is often a one-way street. Once customers and partners build their systems around this data, removing it becomes very difficult. Companies should weigh the benefits against the risks for each category of data rather than treating openness as all-or-nothing.

Benefits:

- meet growing customer expectations for access to the provider's data;
- become embedded inside customer workflows, increasing stickiness; and
- create new monetization opportunities where customers are willing to pay.

Risks:

- disintermediation as customers and competitors bypass the user interface;
- greater data vulnerability as agents gain read/write access;
- higher infrastructure and telemetry costs; and
- regulatory issues around data sharing.

Not all data should be treated equally. Customer records, proprietary analytics, benchmark data, schemas, credentials, and workflow metadata each have different strategic value and risk profiles.

Not all data should be treated equally. Customer records, proprietary analytics, benchmark data, schemas, credentials, and workflow metadata each have different strategic value and risk profiles.

A key strategic decision is determining where your competitive advantage lies. If it's proprietary data, then protecting access may strengthen your position. If your advantage is workflow orchestration and governance, broader access may actually reinforce your platform. The contrast between LinkedIn and ServiceNow illustrates this distinction. LinkedIn treats its professional graph as the core asset and tightly controls programmatic access; ServiceNow treats workflow execution as the asset, exposing governed workflows broadly through MCP while monetizing usage and retaining operational data.

Decide whether that access needs an MCP

Once you've decided what to expose, the next question is how to expose it.

Traditional APIs are designed for developers who write custom integrations against specific end points. MCP serves a different purpose: It provides a standard way for AI agents to discover and use an application's data and actions without requiring a bespoke integration for each vendor. As MCP adoption grows, it lowers the cost of connecting enterprise software into agentic workflows.

That creates a strategic trade-off. Offering MCP access makes it easier for customers' agents to work with your platform, boosting adoption but also making it easier for users to accomplish work without ever opening your interface. Declining to offer MCP reduces that risk but raises another: Agents may simply route around your platform in favor of vendors that are easier to integrate.

The question, then, isn't whether MCP is risky; it's which risk is greater for your business: enabling disintermediation or being excluded from emerging agent ecosystems.

One way to mitigate this trade-off is to pair MCP access with investment in your own first-party agents—that is, agents you build on your own platform, such as Atlassian has done with Rovo. This allows customers to work through your agent experience while still participating in broader agent ecosystems. In effect, you're placing two bets: You're making it easy for customers to use MCP as that ecosystem grows while also building your own agent experience to compete for those same users.

Decide how to monetize

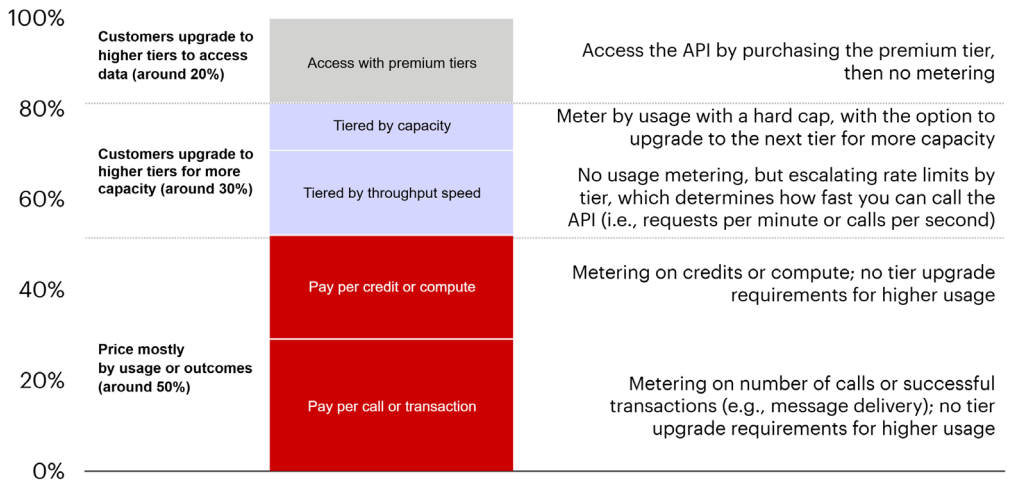
Only after deciding what to expose and whether to offer MCP access should vendors turn to pricing. Here, the market is still experimenting rather than converging. Roughly half of vendors monetize programmatic access through usage- or outcome-based pricing, about one-third charge for additional capacity or higher rate limits, and roughly one-fifth monetize access itself. Most apply similar pricing principles to MCP and traditional APIs, although many, particularly AI-native companies, differentiate pricing by the type of action performed *(see Figure 2)*.

The right pricing model depends less on the prevailing market practice than on how agent access fits within the broader commercial model. Low-cost MCP access can inadvertently cannibalize higher-value seat licenses if customers shift work from users to agents, making it essential to evaluate pricing across the entire product portfolio rather than in isolation.

Vendors must also decide who should pay. Charging end customers keeps the commercial relationship direct; charging developers or ecosystem partners can reduce customer friction but may also weaken ownership of the customer relationship over time.

Figure 2: Five API pricing models account for how most software vendors charge for access

Percentage of software vendors by API pricing model



Source: Bain analysis

Finally, pricing changes should be introduced deliberately. Once developers have built against an interface, tightening access or introducing new charges has repeatedly triggered customer backlash and, in some cases, regulatory scrutiny. Where changes are necessary, a phased transition is generally more effective than an abrupt shift.

Durability of decisions

Pricing will continue to evolve as the market matures, but the strategic choices made around data access are far more durable. The data that gets exposed, the interfaces through which it is accessed, and the role your organization plays in agent workflows will shape competitive positioning even as pricing models change.

Choosing the pricing method is the easy part. The harder decision is what you let out the door and whether doing so strengthens your moat or erodes it.

Bold ideas. Bold teams. Extraordinary results.

Bain & Company is a global consultancy that helps the world's most ambitious change makers define the future.

Bain & Company works with leaders worldwide to solve their toughest challenges and deliver enduring results. Since 1973, we've partnered with clients, including private equity and portfolio companies to build the capabilities they need to stay ahead of change and help them redefine their industries. We measure our success by our clients' success, and we proudly hold the highest levels of client advocacy in our field.

Bain is consistently recognized globally as one of the best places to work. We operate as one global team, uniting strategists, industry and functional experts, technologists, and advisors with a vibrant ecosystem of technology partners.

BAIN & COMPANY