Cephable™

The AI Deployment Gap: Why Enterprise AI Investment Outpaces Adoption—And How to Fix It

A White Paper on Workflow Integration, Governance, and Deployment Flexibility

Executive Summary

A White Paper on Workflow Integration, Governance, and Deployment Flexibility

Enterprise AI has achieved something unprecedented: becoming a $37 billion market in just three years,

making it the fastest-scaling software category in history.[1] Conversion rates are double that of traditional SaaS. Individual users adopt AI tools at four times the rate they adopt conventional enterprise software. Budget holders demonstrate remarkably weak price sensitivity, willing to pay premium rates for capability.[2] Yet two-thirds of organizations remain stuck in pilot purgatory, unable to scale AI beyond experimentation.[3]

The conventional wisdom blames this paradox on insufficient model capabilities, unclear ROI, or organizational resistance to change. The data tells a different story. Enterprise AI doesn't have a capability problem—it has three deployment problems:

The Workflow Integration Problem:

AI remains bolted onto existing processes rather than embedded within them. Users must stop their primary work, switch contexts, and interrupt their flow to access AI—creating a friction tax that compounds across dozens of daily interactions. While high-performing organizations redesign workflows at nearly three times the rate of their peers, most companies struggle to make AI feel like a natural extension of work rather than a separate tool.[4]

The Governance Gap:

Shadow AI usage—employees using personal tools like ChatGPT for work tasks—now represents an estimated 27-40% of actual enterprise AI activity.[5] This isn't evidence of employee negligence; it's evidence that sanctioned enterprise tools don't match the convenience and accessibility of consumer AI. Organizations face a choice: either match that convenience within governed systems or watch ungoverned usage continue to grow.

The Deployment Inflexibility Problem:

The infrastructure layer has concentrated investment in cloud-based orchestration for complex multi-agent systems, yet only 16% of production deployments are architecturally complex.[6] Meanwhile, organizations need AI that works reliably across varied environments—cloud, edge, offline, and low-bandwidth contexts—particularly as adoption expands geographically and into operational settings where always-on connectivity cannot be assumed.

This white paper synthesizes findings from five major enterprise AI reports published in 2025: McKinsey's State of AI, OpenAI's Enterprise AI Report, Microsoft's Frontier Firms study, Anthropic's Economic Index, and Menlo Ventures' Generative AI market analysis. Together, these sources provide an unprecedented view into where AI is actually being deployed, what's driving success, and where systematic gaps create opportunity.

The conclusion is clear: the next wave of enterprise AI value won't come from more sophisticated orchestration or more powerful models. It will come from solving these three deployment challenges—making AI truly integrated, properly governed, and flexibly deployed.

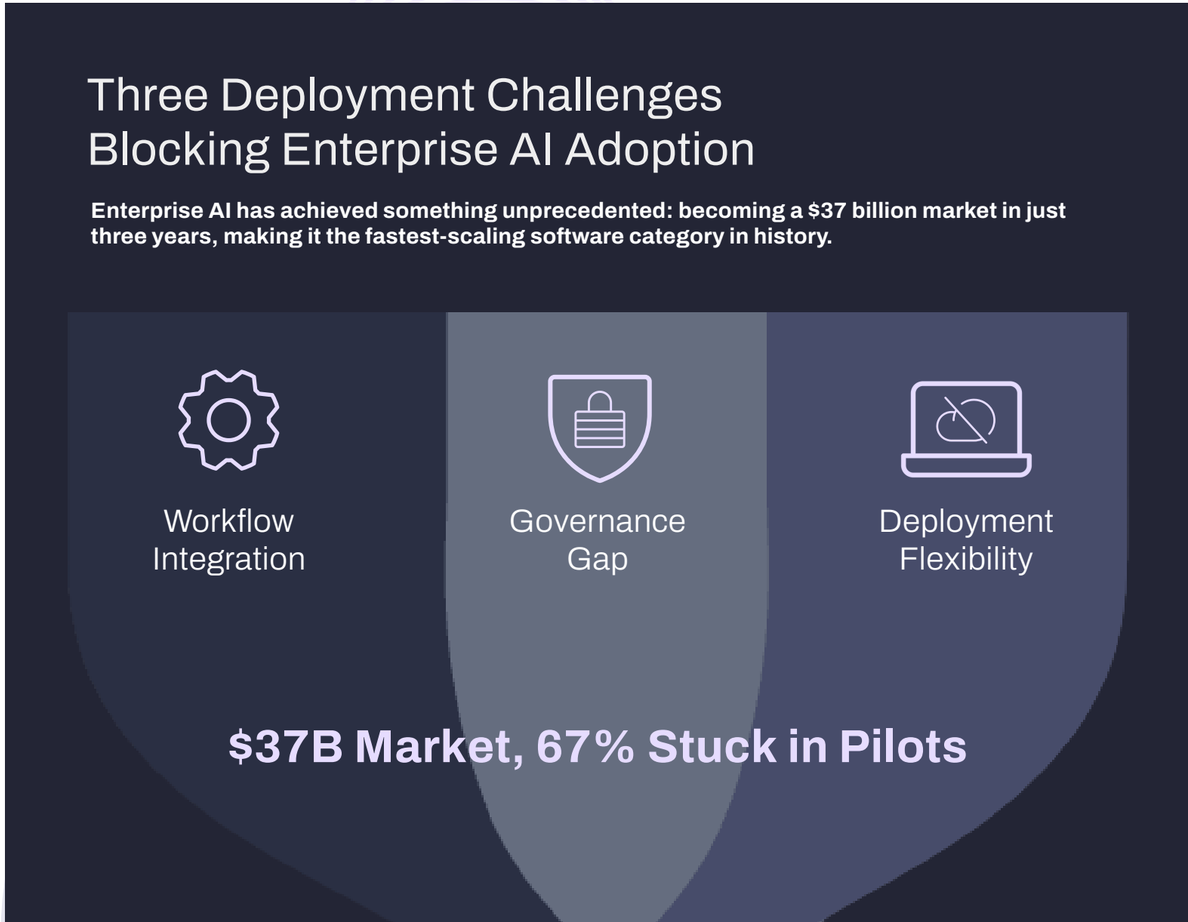
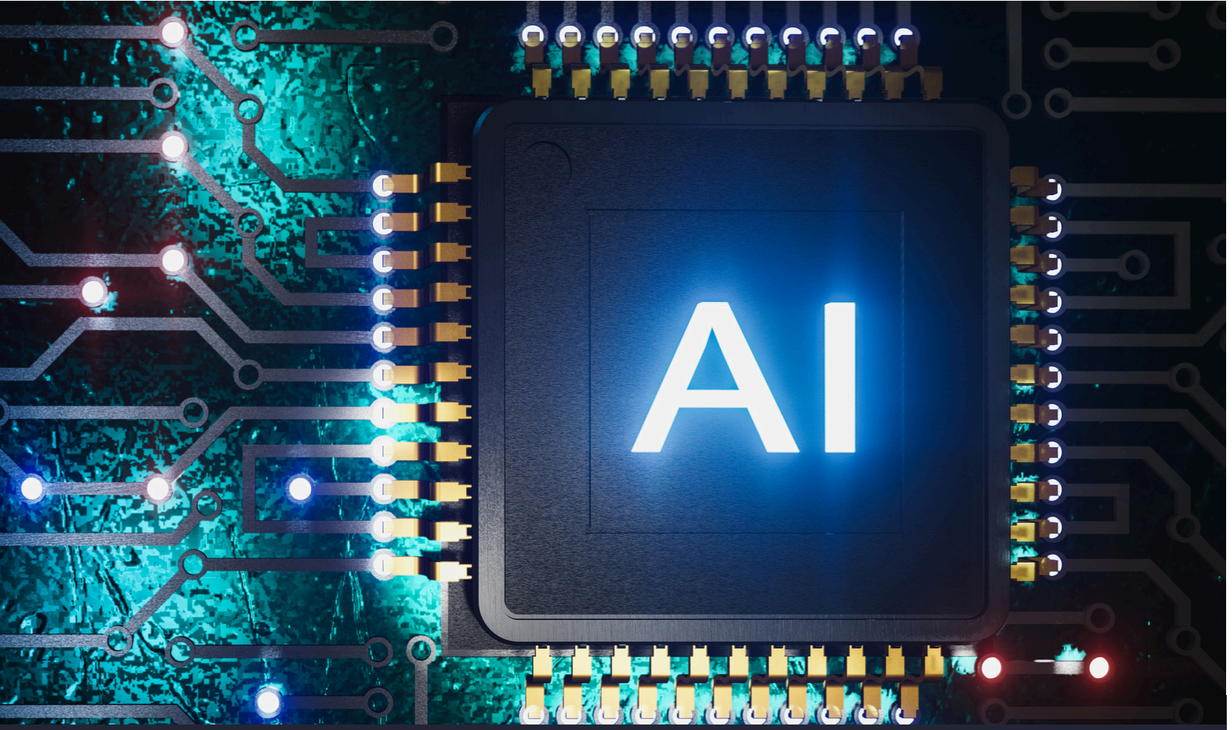


Table of Contents



AI Buyers Convert at Nearly 2x the Rate

AI Conversion Rate: 47%

SaaS Conversion Rate: 25%

The State of Enterprise AI

The Numbers Don't Add Up

By traditional metrics, enterprise AI should be thriving. The market reached **$37 billion in 2025, up from $11.5 billion in 2024—a 3.2x increase dwarfing cloud computing's early trajectory.**[7] At least ten products now generate over $1 billion in annual recurring revenue.[9]

Purchase intent remains extraordinarily high. When organizations evaluate an AI solution, 47% convert to production deployments versus 25% for traditional SaaS.[10] Price sensitivity barely registers—across economically useful tasks, organizations consistently choose higher-cost options when capability justifies it.[11]

The Paradox of Pilot Purgatory

McKinsey found 67% of organizations have not begun scaling AI across the enterprise.[12] This isn't early-stage technology waiting for adoption. Organizations want AI, but something blocks deployment at scale.

The answer lies in how organizations can access it, govern it, and deploy it.

Geography Reveals Adoption Patterns

Geographic distribution of AI adoption offers critical insight. When analyzed per capita, AI usage correlates strongly with GDP—each 1% increase in GDP per capita associates with 0.7% increase in AI usage globally.[14]

Within the United States, Washington D.C. leads with a 3.82x usage index, followed by Utah at 3.78x—both exceeding California's 2.61x.[15]

What distinguishes high-adoption geographies isn't just access to technology—the same models are available globally. It's the presence of complementary organizational infrastructure and deployment capabilities.

The Frontier Firm Divergence

Microsoft's research identified "Frontier Firms"—organizations that successfully scaled AI and are seeing measurable results. These companies report outcomes at 4x the rate of slow adopters.[16] McKinsey's high performers—organizations attributing 5% or more of EBIT to AI—are three times more likely to intend transformative business change.[17]

The gap is widening. OpenAI's data shows frontier workers send six times more AI messages than median workers. Frontier firms generate twice as many messages per seat as median enterprises.[19]

Frontier firms aren't using fundamentally different AI. They're using the same models as everyone else. Their advantage lies in how they've solved the three deployment challenges.

Challenge #1
The Workflow Integration Problem

AI as Interruption, Not Extension

The most fundamental barrier to enterprise AI adoption is the simplest: accessing AI requires interrupting work. Every interaction follows the same costly sequence: recognize need, stop work, switch context, formulate prompt, wait for response, review output, copy relevant portions, switch back, re-establish flow.

For a single interaction, this adds 30-90 seconds of overhead. Enterprise workers don't interact with AI once or twice daily—they interact dozens of times. Across 20-50 daily interactions, this compounds to 10-75 minutes of friction.[20]

Ironically, this friction approaches the magnitude of productivity gains AI delivers. OpenAI reports enterprise users attribute 40-60 minutes of time saved per active day to AI usage.[21] The workflow tax consumes a significant fraction of that gain.

Where Capability Already Exceeds Access

Anthropic's analysis reveals 77% of enterprise API usage exhibits automation patterns—users delegating complete tasks rather than iterating collaboratively.[22] OpenAI tracked directive automation jumping from 27% to 39% in eight months.[23]

As models improve reliability, users delegate more and iterate less. They're not asking AI to help them write code; they're asking AI to write code.

In collaborative workflows, some friction is tolerable. In delegation workflows, friction is deadweight loss. Users want to assign tasks and move on.

Yet current interfaces are optimized for collaboration, not delegation.

The Frequency-Friction Compound Effect

The workflow integration problem doesn't manifest equally across use cases. For infrequent, high-value interactions—comprehensive market analysis or complex legal documents—friction is tolerable because value justifies the interruption.

But much of AI's practical value comes from high-frequency, lower-value interactions: quick factual lookups, brief explanations, simple code completions, draft email responses. These tasks could take 10 seconds with AI but become 60-second interruptions when context-switching overhead is included. Users rationally skip the AI assist.

This creates a paradox: the use cases where AI could provide the most cumulative value—frequent, small assists throughout the day—are precisely where workflow friction is most prohibitive.

Why Workflow Redesign Isn't Sufficient

The standard prescription for AI deployment challenges is workflow redesign.

McKinsey shows high performers redesign workflows at nearly three times the rate of other organizations.[24] Microsoft identifies AI-native processes as a frontier firm characteristic.[25]

This advice is correct but incomplete. Workflow redesign solves critical problems: ensuring AI has the necessary context, making outputs useful rather than requiring manual reformatting, and identifying which tasks can be fully delegated.

What workflow redesign alone cannot solve is the interface problem. Even perfectly redesigned workflows require users to access AI. If accessing AI means stopping their primary task, switching applications, and interrupting flow, friction persists.

GitHub Copilot had every structural advantage: first-mover status, Microsoft's distribution, GitHub's developer relationships, and deep IDE integration. Yet Cursor captured significant market share by solving workflow integration at the interface level—making AI feel ambient rather than requiring constant context-switching.[26]

The pattern generalizes: tools that solve workflow integration at the interface level see dramatically higher adoption than tools requiring workflow redesign alone.

Tools that solve workflow integration at the interface level see dramatically higher adoption than tools requiring workflow redesign alone.

The 90% Use Case Nobody's Optimizing For

The market is building infrastructure for sophisticated, multi-agent systems. Yet only 16% of enterprise deployments and 27% of startup deployments qualify as true agents—systems where an LLM plans actions across multiple steps.[27]

The other 84%? Simple automation: single model calls, basic if-then logic, straightforward task completion. These deployments don't require complex orchestration. They require reliable execution, low latency, and frictionless access.

This is where the workflow integration problem hits hardest. Complex, infrequent tasks justify the friction of accessing specialized tools. Simple, frequent tasks do not. Yet the simple, frequent tasks represent the majority of deployment opportunity.

The opportunity is clear: optimize AI access for the 90% use case. Make simple automation genuinely simple to invoke, use, and integrate.

Challenge #2 – The Governance Gap

The Shadow AI Phenomenon

Here's an uncomfortable truth for IT departments: a substantial portion of enterprise AI usage happens outside sanctioned systems. Research shows that approximately 27% of ChatGPT usage by paying subscribers is work-related, with employees using personal accounts for professional tasks.[28] When this shadow usage is factored in, product-led growth and personal tool adoption may represent close to 40% of actual enterprise AI activity.[29]

This isn't evidence of employee negligence or disregard for security policies. It's evidence that sanctioned enterprise AI tools don't match the convenience, accessibility, and performance of consumer alternatives.

The pattern appears consistently across the data. OpenAI reports that 27% of all AI application spend comes through product-led growth (PLG) motions, where individual users adopt tools directly rather than through enterprise procurement. This is nearly four times the PLG rate in traditional enterprise software (7%).[30] Tools like Cursor reached $200 million in revenue before hiring a single enterprise sales representative, driven entirely by bottom-up developer adoption.[31]

Why does this matter?

Because PLG and shadow AI usage reveal what users actually want: AI that's immediately accessible, requires no procurement friction, works reliably, and integrates naturally into their workflow. When enterprise alternatives require formal requests, approval workflows, and complex deployment processes, employees route around them.

Approximately 27% of ChatGPT usage by paying subscribers is work-related, with employees using personal accounts for professional tasks. Risking data, security, and client information.

The Convenience-Governance Tradeoff

Organizations face an uncomfortable tradeoff. Consumer AI tools optimize for convenience: instant access, continuous model improvements, simple interfaces, zero setup friction. But they come with unclear data policies, minimal audit capabilities, and consumer-grade security.

Enterprise AI tools reverse priorities: robust governance, clear data handling, audit capabilities, contractual SLAs. But they sacrifice convenience through procurement gates, deployment complexity, and feature lag behind consumer alternatives.

The result is a bifurcated market. Knowledge workers use consumer tools for quick tasks and shadow deployments where convenience trumps governance. They use enterprise tools where governance requirements are explicit and enforced.

Why Blocking Isn't a Solution

Some organizations respond to shadow AI by attempting to block consumer tools through network policies or endpoint controls. This approach fails for several reasons.

First, determined users route around technical restrictions. They use personal devices, cellular connections, or browser-based workarounds. Blocking creates friction but rarely achieves comprehensive control.

Second, blocking without providing alternatives pushes productive work onto less efficient methods.

If employees can't access AI for legitimate tasks, they don't stop doing those tasks—they do them slower, less effectively, or find unofficial workarounds that create even larger governance gaps.

Third, the velocity of AI improvement means enterprise alternatives struggle to keep pace with consumer tool capabilities. Blocking the best available tools to enforce use of inferior alternatives frustrates users and reduces productivity.
The fundamental problem isn't that employees are bypassing controls—it's that sanctioned alternatives aren't competitive on the dimensions users care about: accessibility, performance, and workflow integration.

The Path to Governed Convenience

The solution isn't choosing between convenience and governance. It's building systems that deliver both. This requires several capabilities that current enterprise AI largely lacks:

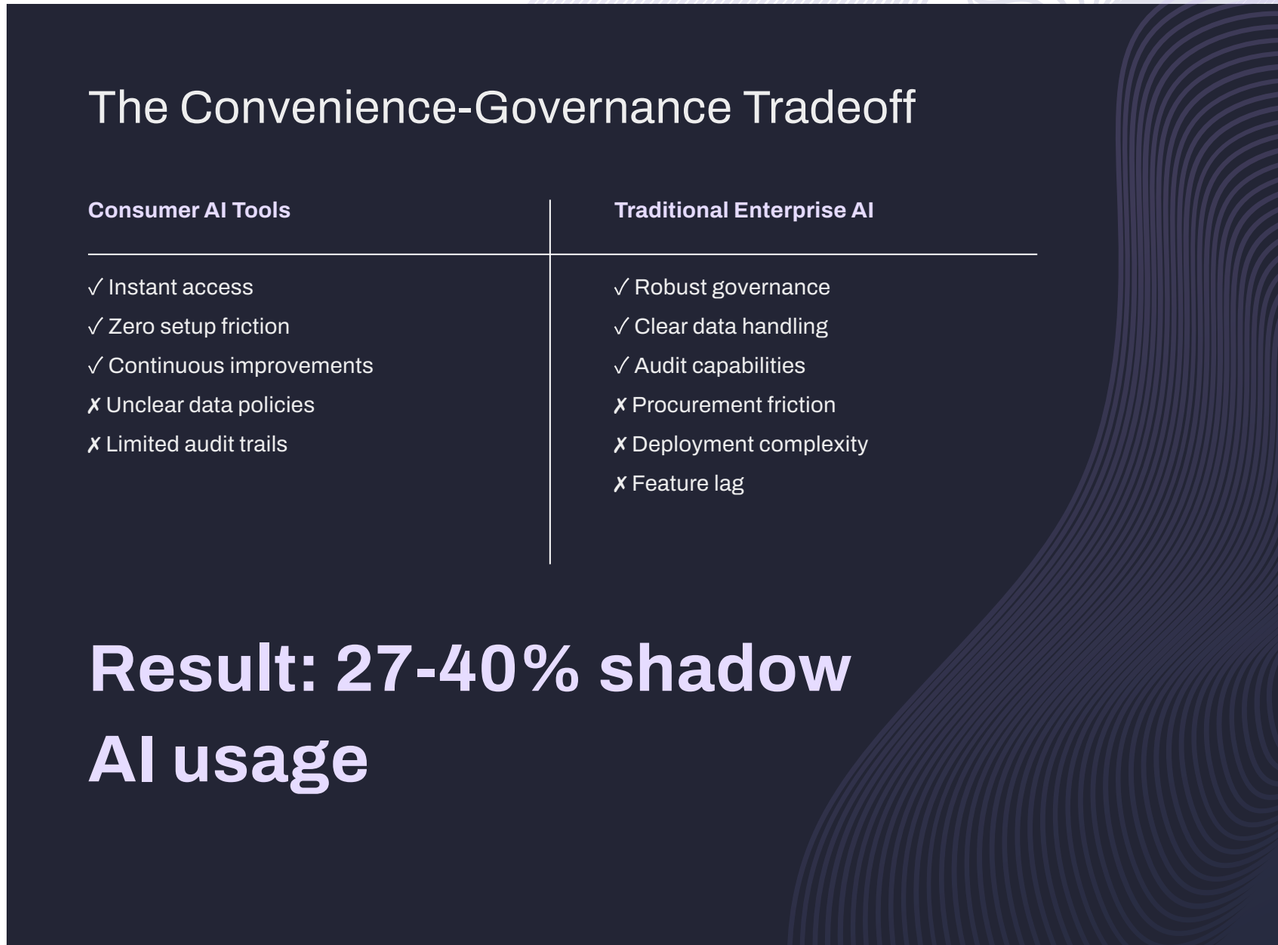
Ambient accessibility: AI that's present in the flow of work, not separate from it. Users shouldn't need to context-switch to access governed AI—it should be as immediately available as consumer tools but with enterprise controls.

Deployment flexibility: Enterprise AI that works across environments—cloud when connectivity is strong, edge when it isn't, offline when necessary. Governance doesn't require centralization; it requires consistent policy enforcement regardless of deployment context.

Frictionless provisioning: Eliminating procurement and deployment barriers without sacrificing control.

Users should get instant access to governed AI the same way they currently get instant access to ungoverned consumer tools.

The organizations solving shadow AI aren't those blocking consumer tools most effectively. They're those making enterprise alternatives competitive on convenience while maintaining governance advantages. The governance gap closes when the governed option is also the best option.



Challenge #3 – The Deployment Inflexibility Problem

The Cloud-Centric Assumption

The AI infrastructure market has made an implicit bet: that enterprise AI will run primarily in the cloud, accessed through APIs, with centralized model serving and orchestration. The spending pattern reflects this—$18 billion in infrastructure investment heavily concentrated in cloud-based systems.[32]

This centralized architecture makes sense for many use cases. Cloud deployment offers scalability, simplified management, rapid model updates, and access to the most powerful available models.

But the assumption that all enterprise AI should be cloud-centric creates three problems:

Geographic constraints: AI adoption correlates strongly with infrastructure quality. Countries and regions with robust, high-bandwidth connectivity show dramatically higher usage.[33] This isn't just access to models—it's the ability to use them with acceptable latency and reliability.

Operational limitations: Many enterprise contexts—manufacturing floors, field operations, remote locations, regulated environments—cannot assume always-on connectivity. Cloud-only deployment excludes these contexts entirely.

Efficiency gaps: Organizations have made recurring investments in endpoint compute—laptops, workstations, mobile devices with increasingly capable processors. Cloud-centric architectures leave this compute capacity idle for AI workloads.

The Edge Computing Signal

The market is beginning to acknowledge deployment flexibility as a requirement. Menlo Ventures' 2026 predictions include: "Models finally move to the edge. Motivated by low-latency requirements and cost reduction, on-device inference becomes table stakes."[34]

This isn't speculative. Mobile manufacturers are shipping dedicated AI accelerators. Apple, Google, and Samsung are building on-device inference into operating systems. Open-weight models are demonstrating acceptable performance on consumer hardware.

The shift toward edge deployment isn't just about cost reduction or privacy benefits—though both matter. It's about deployment flexibility: the ability to deliver AI capabilities reliably across varied operational contexts.

Efficiency and ROI Implications

Consider the economics. Cloud inference pricing varies by model and provider, but organizations pay per token processed—costs that compound across thousands of daily interactions per user. For high-frequency, simple tasks, per-token costs add up quickly.

Meanwhile, organizations have already invested in endpoint compute. Modern laptops ship with capable GPUs or NPUs. Mobile devices include dedicated AI accelerators. These assets sit largely idle for AI workloads.

A hybrid approach—routing simple, high-frequency tasks to edge compute while reserving cloud resources for complex, high-value workloads—could dramatically improve ROI. Users get lower latency for frequent interactions while organizations reduce cloud spend.

9- The Deployment Inflexibility Problem

This isn't arguing for the abandonment of cloud AI. Cloud infrastructure will remain essential for frontier models, complex orchestration, and workloads requiring massive scale. The argument is for flexibility: the ability to route workloads appropriately based on task requirements, connectivity, and economic efficiency.

Deployment Flexibility: AI That Works Everywhere

The enterprise need is for deployment flexibility—AI that works reliably across cloud, edge, and hybrid contexts—not universal cloud dependency.

Cloud: High-performance, centralized, always-connected contexts

Edge: Local processing, lower latency, works offline

Hybrid: Optimal routing: cloud when needed, edge the rest of the time

Hybrid connects the two with policy-based routing—so the system automatically chooses the right execution path for each request while keeping governance, logging, and controls consistent across environments. The result is an AI experience that feels immediate for users, remains manageable for IT, and stays resilient across offices, travel, regulated settings, and variable connectivity.

Privacy and Governance Benefits

Deployment flexibility also addresses governance concerns. For sensitive workloads—processing confidential documents, analyzing proprietary data, handling regulated information—edge processing offers advantages that cloud architectures cannot match.

Data that never leaves the device cannot leak through API calls. Processing that happens locally doesn't create cloud logs requiring retention policies and access controls. Compliance in regulated industries often becomes simpler when data processing happens on controlled endpoints rather than third-party infrastructure.

This doesn't solve the entire governance challenge—edge devices require their own management and security controls. But it provides options. Organizations can make risk-based decisions: cloud processing for general tasks where convenience and capability matter most, edge processing for sensitive workloads where data locality is paramount.
The current architecture offers limited flexibility. Most enterprise AI assumes cloud processing with limited options for local execution. This one-size-fits-all approach forces organizations into suboptimal tradeoffs rather than allowing deployment decisions based on specific requirements.

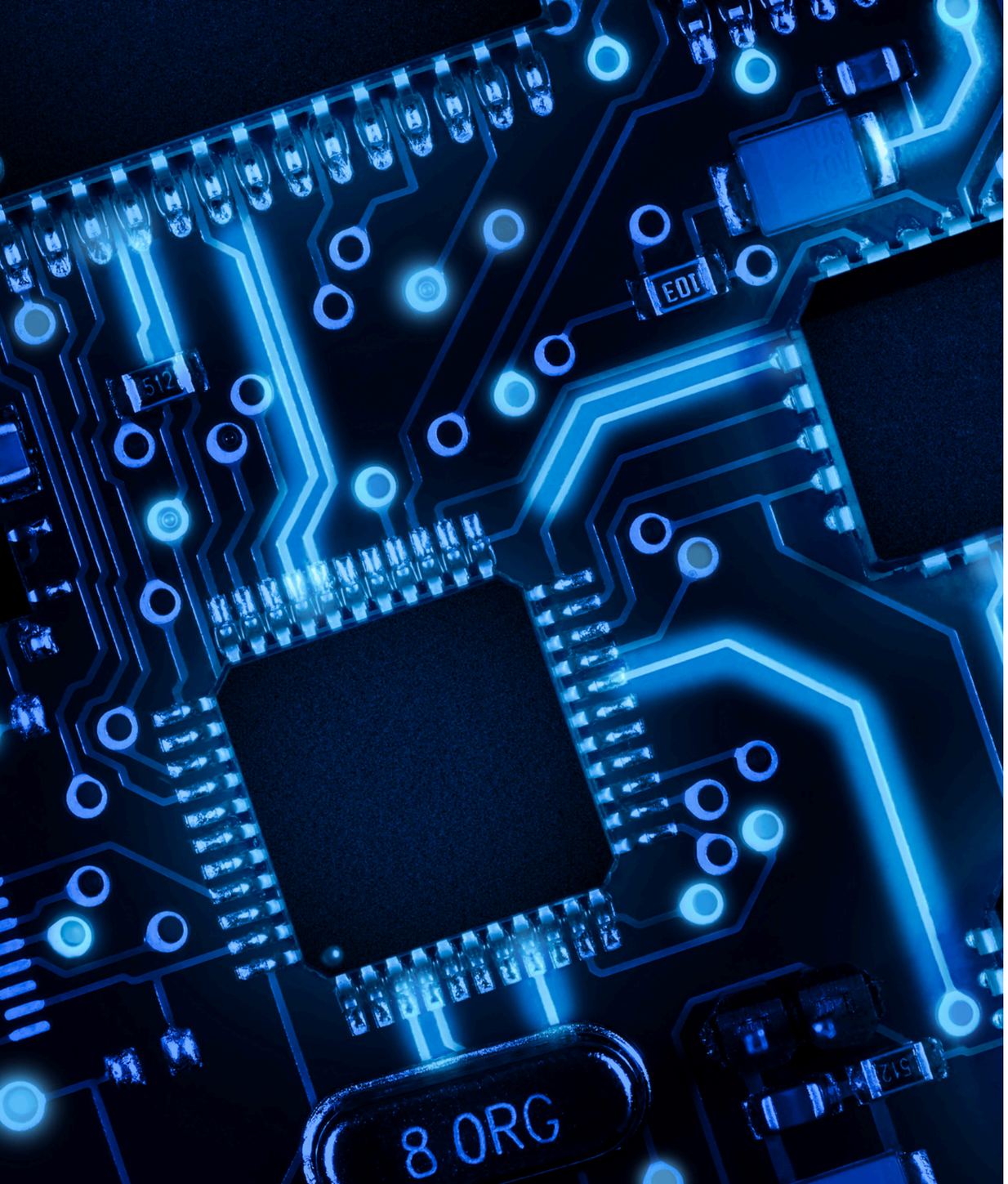
The Infrastructure Timing Mismatch Returns

Here's the core tension: the infrastructure layer has invested $18 billion in cloud-based orchestration optimized for complex, multi-agent systems.

But most production deployments are simple automations that could run efficiently on edge devices. And the enterprise need is for deployment flexibility—AI that works reliably across cloud, edge, and hybrid contexts—not universal cloud dependency.

This creates opportunity. As models continue to improve while becoming more efficient, as edge hardware capabilities increase, and as organizations demand deployment options beyond cloud-only architectures, the market needs solutions that deliver governed, accessible AI across deployment contexts. The question isn't cloud versus edge. It's whether AI infrastructure can evolve beyond cloud-centric assumptions toward genuine deployment flexibility—giving organizations the ability to route workloads optimally based on latency requirements, privacy considerations, connectivity constraints, and cost optimization.

The question isn't cloud versus edge. It's whether AI infrastructure can evolve beyond cloud-centric assumptions toward genuine deployment flexibility.



Where Value Concentrates (And Where It's Blocked)

The Capability Threshold Effect

AI adoption doesn't follow a smooth curve. It follows a step function. For any given domain, usage remains minimal until model capability crosses a "good enough" threshold, at which point adoption explodes regardless of cost or organizational readiness.
Coding provides the clearest example. In 2023, AI coding tools represented a negligible market. In 2024, the category reached $550 million. In 2025, it hit $4 billion—55% of all departmental AI spending and the single largest use case across the entire application layer.[35] This wasn't gradual growth. It was a phase transition triggered when Claude Sonnet 3.5 launched in June 2024, delivering performance that crossed from "interesting" to "economically transformative" for real development workflows.

Anthropic held 54% of the enterprise coding market for eighteen consecutive months not through vendor lock-in or distribution advantages, but through sustained performance leadership on the benchmarks developers actually care about.[36] When Google released Gemini 3 Pro in November 2025, it led most major evaluations—except SWE-bench Verified, where Claude still held the edge. A week later, Anthropic released Claude Opus 4.5, widening the gap again.[37]

The pattern repeats across domains. Healthcare ambient scribes—AI that automatically documents clinical encounters—grew from $250 million to $600 million in a single year, minting two new unicorns in the process.[38] The technology didn't suddenly appear. What appeared was capability sufficient to reduce documentation time by more than 50% while maintaining accuracy standards clinicians would trust.

Where Deployment Challenges Block Value

But capability alone doesn't guarantee adoption. The three deployment challenges—workflow integration, governance gaps, and deployment inflexibility—create friction that prevents organizations from capturing value even when AI capability is sufficient.

Workflow integration blocks high-frequency value. The use cases with the most cumulative impact are those integrated into continuous work: quick lookups, brief explanations, simple automations repeated throughout the day. These are precisely the interactions where context-switching friction is most prohibitive. Organizations end up deploying AI for occasional high-value tasks while missing compounding benefits of frictionless, continuous integration.

Governance gaps block enterprise-wide scaling. Shadow AI usage of 27-40% means a substantial portion of value creation happens outside managed systems.[39] This creates several problems: productivity gains aren't captured in enterprise metrics, best practices don't diffuse across teams, security and compliance risks grow with scale, and IT lacks visibility into what's actually driving value.

Deployment inflexibility blocks operational contexts. Geographic data shows AI adoption correlating strongly with infrastructure quality—each 1% increase in GDP per capita associates with 0.7% increase in AI usage.[40] This reflects not just wealth but connectivity.

Cloud-dependent AI excludes or limits use in manufacturing, field operations, remote locations, and international markets with inconsistent infrastructure. Organizations miss operational value because deployment assumes infrastructure that doesn't exist everywhere.

The Automation Acceleration Signal

Perhaps the most revealing trend appears in how users interact with AI over time. OpenAI tracked directive automation—users delegating complete tasks with minimal back-and-forth—jumping from 27% to 39% of interactions in just eight months. This is the first period where automation usage exceeded augmentation usage.[41]

Anthropic's API data shows an even starker pattern: 77% of enterprise API usage exhibits automation modes, compared to roughly 50% in consumer Claude.ai usage.[42] As models improve reliability and users build trust, they delegate more and iterate less. This shift intensifies the importance of solving deployment challenges. In collaborative workflows where iteration is expected, some friction is tolerable. In delegation workflows where users want to state needs and move on, any friction becomes pure waste.

The trajectory is unambiguous: AI is moving toward delegation and automation, not augmentation and collaboration. Tools optimized for this reality—making delegation frictionless rather than making collaboration sophisticated—will capture disproportionate value.

The Infrastructure-Application Imbalance

The spending distribution reveals a fundamental mismatch. In 2025, infrastructure captured $18 billion while applications captured $19 billion—nearly equal investment.[43]

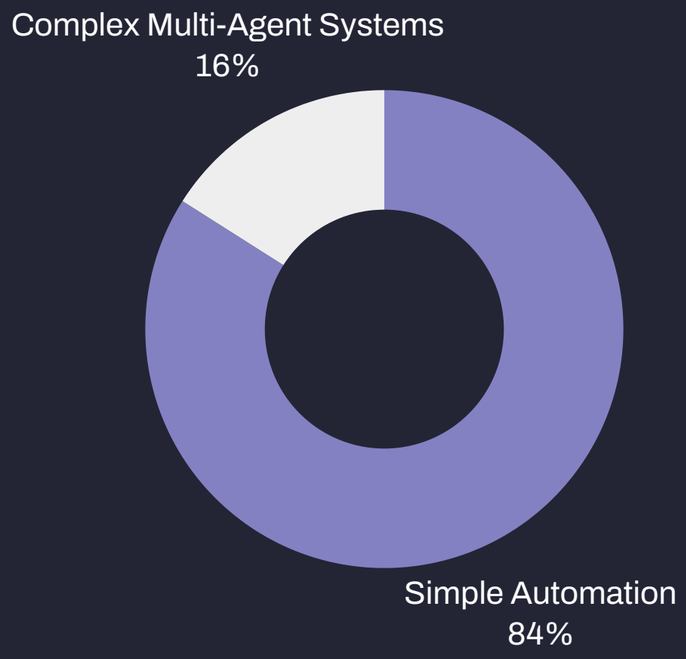
But infrastructure is being built for the 16% of deployments that are architecturally complex, while the 84% majority are simple automations that need different optimization: frictionless access, deployment flexibility, and seamless workflow integration.

This creates opportunity. The next $10 billion in enterprise AI value won't come from more sophisticated backend orchestration serving the 16% complex use case. It will come from solving deployment challenges for the 84% simple automation use case—making high-frequency AI interactions genuinely frictionless.

AI is moving toward delegation and automation, not augmentation and collaboration. Tools optimized for this reality—making delegation frictionless rather than making collaboration sophisticated—will capture disproportionate value.

How AI Is Actually Deployed in Production

Sophisticated workflows justifying complexity



High-frequency, simple tasks requiring frictionless access

Source: Menlo Ventures, 2025

Conclusion: The Path Forward

Three years into the enterprise AI transformation, the market has proven several things conclusively: Demand exists. Conversion rates double traditional software. Organizations demonstrate weak price sensitivity. PLG adoption runs at 4x normal rates. Buyers want AI and they're willing to pay for it. Capability exists. Coding hit $4 billion the moment models crossed the economic threshold. Healthcare, legal, and other verticals are scaling rapidly as domain-specific capability improves. The technology works.

Value exists. Users save 40-60 minutes daily. High performers attribute measurable EBIT impact to AI. Frontier firms report outcomes at 4x the rate of slow adopters. The ROI is real.

What blocks scaled adoption isn't capability, budget, or value recognition. It's three deployment challenges that prevent organizations from capturing the value AI already delivers:

The Three Deployment Challenges

Challenge #1: Workflow Integration
AI remains bolted onto work rather than embedded within it. Every interaction requires context-switching that compounds across dozens of daily uses. The friction tax approaches the magnitude of productivity gains, particularly for high-frequency, simple tasks that represent the majority of value opportunity. Organizations need AI that feels ambient—present in the flow of work, not separate from it.

Challenge #2: The Governance Gap
Shadow AI usage of 27-40% proves that sanctioned enterprise tools don't match consumer convenience. Employees aren't bypassing governance out of negligence—they're routing around tools that sacrifice too much usability for control. Organizations need AI that delivers both: consumer-grade convenience with enterprise-grade governance.

Challenge #3: Deployment Inflexibility
Cloud-centric infrastructure assumes always-on connectivity and accepts latency that limits high-frequency use. This excludes operational contexts, creates geographic barriers, and leaves endpoint compute investments unutilized. Organizations need deployment flexibility—AI that works reliably across cloud, edge, and hybrid contexts based on specific requirements rather than universal assumptions.

What This Means Practically

The data points toward several clear implications:
Stop building for the wrong 16%. The infrastructure layer has over-invested in complex orchestration for multi-agent systems that represent a minority of production deployments. The 84% majority—simple, high-frequency automations—need different optimization.

Optimize for delegation, not collaboration. Automation patterns are accelerating (27% to 39% in eight months). Users want to delegate complete tasks, not iterate collaboratively. Interface design should prioritize frictionless handoff over sophisticated interaction.

Make governed AI competitive on convenience. Shadow AI won't be solved by blocking consumer tools. It requires making enterprise alternatives match consumer convenience while maintaining governance advantages.

The governed option must also be the best option.

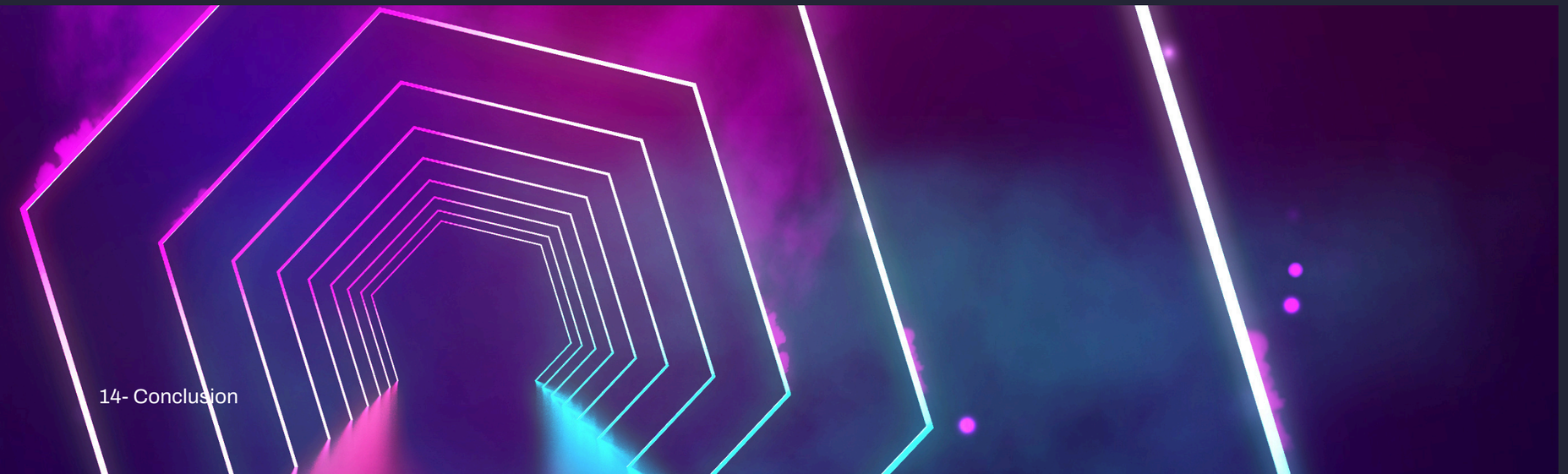
Enable deployment flexibility. As AI expands into operational contexts, international markets, and use cases requiring privacy or offline capability, organizations need options beyond cloud-only architectures. Hybrid approaches that leverage edge compute for simple tasks while reserving cloud for complex workloads optimize both cost and capability.
Solve workflow integration at the interface level. The most successful AI tools—Cursor in coding being the canonical example—don't just offer capability. They eliminate the friction of accessing that capability. Making AI ambient, voice-driven, and present in workflow without context-switching is the next interface evolution.

The Next Wave of Value Creation

The market has spent three years proving AI works. The next phase is making it accessible.
Organizations that solve these three deployment challenges—workflow integration, governance, and deployment flexibility—won't just capture market share. They'll enable a fundamentally different relationship between workers and AI: one where AI feels like a natural extension of capability rather than a separate tool requiring conscious invocation.
This is where the next $10 billion in enterprise AI value lives. Not in more powerful models, not in more sophisticated orchestration, but in eliminating the deployment barriers that prevent organizations from using the AI that already works.

The question is no longer whether AI can transform enterprise work. The question is whether deployment can match capability—making AI truly ambient, properly governed, and flexibly deployed wherever work happens.

For organizations ready to solve these challenges, the opportunity is clear and the time is now.



Ready To Solve The Three Deployment Challenges?

The data is clear: enterprise AI adoption isn't limited by capability, budget, or ROI. It's limited by three deployment challenges:

- Workflow Integration: Making AI ambient rather than requiring context-switching
- Governance: Delivering consumer convenience with enterprise control
- Deployment Flexibility: Working reliably across cloud, edge, and hybrid contexts

Cephable addresses all three challenges with multi-modal, ambient AI that integrates seamlessly into workflow, maintains enterprise governance, and deploys flexibly across environments.

Contact Information

CEO, Alex Dunn: alex@cephable.com
COO, Jason Fields: jfields@cephable.com

Trusted Partners

Cephable™

About The Author

Jason Fields, COO

Jason Fields is the Chief Operating Officer at Cephable, where he leads the operational strategy and cross-functional execution behind the company's on-device AI platform.

Jason's work sits at the intersection of productivity, workflow integration, and enterprise readiness. At Cephable, he helps align product, go-to-market, and client success, while still meeting enterprise requirements for control, security, and accountability.

Before Cephable, Jason served as Chief Strategy Officer at Voicify, where he helped enterprises build and scale voice and AI experiences through the Voicify Experience Platform. That role deepened his expertise in designing AI interfaces that fit real operational constraints, reducing context-switching, and moving from experimentation to durable deployment.

Earlier in his career, Jason was a Founding Member of the Microsoft Customer Engagement Alliance, a cross-organization initiative focused on best practices and practical guidance for modern digital systems. He also served as Senior Vice President of Strategy at Rightpoint, where he led growth initiatives, ran the LA office, and owned the West Coast PnL, supporting organizations navigating complex technology rollouts and change management.

The Process Behind This Paper

As a senior technologist, you are well-acquainted with the rapid pace of advancements in AI within the enterprise sector. The abundance of AI-related content can be overwhelming, making it challenging to stay abreast of the most impactful research.

Recently, I came across a valuable resource on LinkedIn, where Heena Purohit highlighted five pivotal research pieces she deemed essential for 2025. Although I had previously encountered these studies, I had not yet delved into their contents. To efficiently assimilate their insights, I utilized ChatGPT to extract key themes, high-level findings, overlapping data, and correlated insights from each piece.

Subsequently, I employed Copilot, leveraging its extensive contextual knowledge of Cephable, to identify potential benefits and gaps that our software could address. This analysis served as a foundation for drafting a whitepaper specifically tailored to the needs of CIOs and CTOs.

To refine the message and ensure consistency across the aggregated content, I utilized Word, with Cephable's assistance in fine-tuning the tone and presentation.

In summary, by strategically employing four AI tools, each with its unique strengths, I was able to generate comprehensive content in just 2, two-hour sessions, a task that would typically require weeks of meticulous reading, correlation, and authoring. This approach not only streamlines our content creation process but also ensures that we remain at the forefront of AI innovation, ultimately benefiting our organization and its leadership.

About This Research

This white paper synthesizes findings from five major enterprise AI reports published in 2025:

- McKinsey & Company, "The State of AI in 2025: Agents, Innovation, and Transformation"
- OpenAI, "The State of Enterprise AI: 2025 Report"
- Microsoft and IDC, "Bridging the AI Divide: How Frontier Firms Are Transforming Business"
- Anthropic, "The Anthropic Economic Index Report: Uneven Geographic and Enterprise AI Adoption"
- Menlo Ventures, "2025: The State of Generative AI in the Enterprise"

Combined, these reports represent survey data from over 5,000 enterprise decision-makers, analysis of billions of AI interactions, and market sizing across the full AI technology stack.

---

[1] Menlo Ventures, "2025: The State of Generative AI in the Enterprise," December 2025, market sizing analysis.
[2] OpenAI, "The State of Enterprise AI: 2025 Report," showing 47% conversion rates and Anthropic Economic Index Report documenting price elasticity of -0.29.
[3] McKinsey & Company, "The State of AI in 2025: Agents, Innovation, and Transformation," November 2025, Global Survey findings.
[4] McKinsey & Company, "The State of AI in 2025," workflow redesign practices among high performers.
[5] Menlo Ventures, "2025: The State of Generative AI in the Enterprise," shadow AI analysis citing Chatterji et al., "How People Use ChatGPT," NBER Working Paper, September 2025.
[6] Menlo Ventures, "2025: The State of Generative AI in the Enterprise," analysis of AI architectures in production.
[7] Menlo Ventures, "2025: The State of Generative AI in the Enterprise," comparing $11.5B (2024) to $37B (2025).
[8] Ibid., application layer analysis showing $19B of $37B total spend.
[9] Ibid., market landscape analysis.
[10] OpenAI, "The State of Enterprise AI: 2025 Report," conversion rate comparison.
[11] Anthropic, "The Anthropic Economic Index Report," September 2025, API cost analysis section.
[12] McKinsey & Company, "The State of AI in 2025," Global Survey results on scaling status.
[13] Anthropic, "The Anthropic Economic Index Report," geographic analysis with Anthropic AI Usage Index (AUI) methodology.
[14] Ibid., correlation analysis between GDP per capita and AI usage.
[15] Ibid., United States state-level analysis.
[16] Microsoft and IDC, "Bridging the AI Divide: How Frontier Firms Are Transforming Business," November 2025, frontier firm analysis.
[17] McKinsey & Company, "The State of AI in 2025," high performer characteristics.
[18] Ibid., AI agent scaling patterns among high performers.
[19] OpenAI, "The State of Enterprise AI: 2025 Report," frontier worker and firm usage patterns.
[20] Analysis based on typical AI interaction patterns and context-switching research in knowledge work.
[21] OpenAI, "The State of Enterprise AI: 2025 Report," productivity metrics from enterprise user survey.
[22] Anthropic, "The Anthropic Economic Index Report," API collaboration mode analysis.
[23] OpenAI, "The State of Enterprise AI: 2025 Report," collaboration mode evolution from late 2024 to August 2025.
[24] McKinsey & Company, "The State of AI in 2025," workflow redesign practices among high performers.
[25] Microsoft and IDC, "Bridging the AI Divide: How Frontier Firms Are Transforming Business," organizational transformation characteristics.
[26] Menlo Ventures, "2025: The State of Generative AI in the Enterprise," coding market analysis and competitive dynamics.
[27] Menlo Ventures, "2025: The State of Generative AI in the Enterprise," AI architectures in production analysis.
[28] Chatterji et al., "How People Use ChatGPT," NBER Working Paper No. 34255, September 2025.
[29] Menlo Ventures, "2025: The State of Generative AI in the Enterprise," shadow AI and PLG analysis.
[30] Ibid., PLG adoption rates in AI vs. traditional SaaS.
[31] Ibid., Cursor case study.
[32] Menlo Ventures, "2025: The State of Generative AI in the Enterprise," infrastructure layer spending analysis.
[33] Anthropic, "The Anthropic Economic Index Report," geographic adoption patterns and infrastructure correlation.
[34] Menlo Ventures, "2025: The State of Generative AI in the Enterprise," 2026 predictions section.
[35] Menlo Ventures, "2025: The State of Generative AI in the Enterprise," departmental AI spend breakdown.
[36] Ibid., LLM market share analysis.
[37] Ibid., coding market dynamics and model performance timeline.
[38] Menlo Ventures, "2025: The State of AI in Healthcare," October 2025, ambient scribe market analysis.
[39] Menlo Ventures, "2025: The State of Generative AI in the Enterprise," shadow AI usage analysis.
[40] Anthropic, "The Anthropic Economic Index Report," correlation analysis between GDP per capita and AI usage.
[41] OpenAI, "The State of Enterprise AI: 2025 Report," collaboration mode evolution.
[42] Anthropic, "The Anthropic Economic Index Report," API vs. Claude.ai collaboration patterns.
[43] Menlo Ventures, "2025: The State of Generative AI in the Enterprise," infrastructure vs. application spend breakdown.