softserve

THE ECONOMICS OF AGENTIC AI: COST DYNAMICS AND NEW VALUE STREAMS

Agentic AI represents a fundamental shift from static content generation to autonomous systems that reason, plan, and act toward defined business goals. Unlike earlier Generative AI systems, agentic AI actively participates in enterprise operations, executing complex, multi-step workflows while adapting in real time.

The economic impact of this technology adoption is already measurable. Organizations implementing these systems are seeing clear benefits, including measurable gains in productivity, significant cost savings, and noticeable improvements in customer experience. Many leaders anticipate that agentic AI will transform the workplace in ways that surpass previous technological shifts and are prioritizing investments to accelerate its deployment and scale.

However, realizing the full economic potential of agentic AI requires understanding its unique cost structure and implementing the right economic model. This three-part report examines the value streams, cost dynamics, and strategic considerations essential for maximizing return on investment in multi-agent systems.

An AI-led economic shift

The economic implications of this shift extend far beyond traditional AI implementations. Where previous systems required direct human input for each interaction, agentic AI systems can operate continuously, making decisions and taking actions that directly impact business outcomes.

This autonomy creates new value streams, with the Journal of Advances in Artificial Intelligence reporting that, when compared to traditional AI, multi-agent systems boast

34.2%
reduction in task completion time

7.7%
increase in accuracy

13.6%
improvement in resource utilization

However, agentic AI also introduces novel cost considerations that organizations must understand to succeed.

The transition from passive AI tools to active AI agents changes the fundamental economics of artificial intelligence in the enterprise. Rather than measuring value per query or interaction, organizations must now evaluate business value per autonomous action, considering the full spectrum of computational costs, infrastructure requirements, and operational complexities that come with deploying systems capable of independent reasoning and decision-making.

Understanding these economics becomes critical as organizations scale their AI implementations. The difference between a successful agentic AI deployment and one that fails to deliver expected returns often comes down to how well leadership understands and manages the unique cost dynamics of multi-agent AI systems.

Find new value streams with tangible ROI

Agentic AI and multi-agent systems have created entirely new categories of business value that were previously impossible or impractical to capture. These value streams extend beyond traditional automation to include areas where autonomous reasoning and decision-making can create competitive advantages, including:

Automation of knowledge discovery accelerates research and insight generation across industries. Rather than requiring human analysts to manually sift through data sources, agentic systems autonomously identify patterns, correlate information across multiple domains, and generate actionable insights. This capability allows organizations to respond more quickly to market changes and identify opportunities that might otherwise be missed.

A retail bank automated credit-risk memos by using agentic AI to extract data from multiple sources, generate confidence scores, and propose follow-up questions.

This method reduced turnaround time by **30%**

Increased analyst productivity by **60%**

Customer engagement benefits from AI agents that maintain context across complex, multi-step interactions while personalizing responses based on individual customer needs and preferences. These systems can handle escalating complexity in customer requests, often resolving issues that would previously require human intervention.

150,000

Number of agentic AI systems a large beauty brand used to analyze dermatologist-annotated images and customer data to increase engagement through personalized recommendations.

Software development has seen particularly strong adoption, with agents handling everything from code generation to debugging and deployment processes. The autonomous nature of these systems allows them to iterate on solutions, test multiple approaches, and optimize code quality without constant developer oversight.

Notably, NVIDIA's AI Blueprints help developers build agents that autonomously analyze documents, summarize video content, and orchestrate multi-agent workflows.

Measurable productivity gains

Early implementations across multiple sectors demonstrate the tangible business impact of agentic AI adoption. Notably, the implementation of multi-agent systems has already had an impact on the following sectors:

Software engineering

Software engineering has experienced significant productivity improvements through agent-driven optimization. A study of GitHub data found that about one-third of pull requests in its repositories are now created by bots and automation tools, representing

49% year-over-year increase

This shift indicates that agentic tools are handling an increasing portion of routine development tasks, freeing human developers to focus on higher-value architectural and strategic work.

Healthcare

Healthcare applications show promising results in both diagnostic accuracy and operational efficiency. The Permanente Medical Group saw

30% reductions in physician documentation time during early agentic deployments

This improvement allows healthcare providers to spend more time on patient care while maintaining comprehensive medical records.

Financial services

Financial services has seen material operational savings as agents automate triage and investigation processes. Banks report that consumer servicing costs have decreased by roughly

30% as AI tools, including agentic workflows in fraud detection and process operations, scale across different lines of business

The ability of these systems to handle routine inquiries and flag suspicious activities has reduced the human workload while maintaining or improving service quality.

Retail

Retail operations benefit from agents that personalize customer experiences and optimize supply chain decisions. Deployed agent programs in retail environments have demonstrated the ability to handle up to

80% of customer service issues

90% faster response times have been achieved

This automation has contributed to higher customer satisfaction while reducing operational overhead. Additionally, agentic systems have helped retailers reduce excess inventory and waste through better demand forecasting and supply chain optimization.

Explore the cost dynamics of agentic AI

Understanding the specific cost dynamics of agentic AI is critical for organizations aiming to maximize value and ensure sustainable returns on investment. These systems introduce new layers of operational expenses, from intensive computational needs to additional infrastructure and integration requirements. Carefully managing and planning for these costs is essential for successful deployment at scale.

Computational costs

Effectively managing computational costs is essential for organizations building or scaling multi-agent systems. Agentic AI technology introduces a new class of expenses, as advanced reasoning models require far greater computing resources than earlier AI systems. Factoring in these demands is critical for both budgeting and long-term resource planning.

Reasoning models that power agentic AI systems require substantially more computational resources than traditional AI implementations. A Johns Hopkins University study found reasoning models to require up to around **16x compute when compared to standard Gen AI models**.

The increased resource requirements stem from several factors inherent to agentic AI operations. Multi-step reasoning processes require multiple model invocations for a single task completion. Real-time adaptation capabilities mean systems must continuously process context and adjust their approaches. The ability to execute workflows toward defined business goals requires sophisticated planning algorithms that consume significant computational resources.

Barriers for enterprise adoption

These elevated computational costs create significant barriers, particularly for small to medium-sized enterprises or organizations with constrained budgets. The financial burden extends beyond direct model usage to include the high-performance computing infrastructure needed to support autonomous reasoning and decision-making at scale.

Organizations must weigh the higher operational costs against potential productivity gains and competitive advantages. For many businesses, the challenge lies not in justifying the value of agentic AI, but in managing the cash flow implications of the increased computational requirements.

Model training costs

Model training represents one of the most capital-intensive elements of the AI lifecycle, often influencing strategic decisions around technology adoption, vendor partnerships, and risk management. Training advanced models continues to rise.

Google's Gemini advanced training model costs estimated at $30–191 million and ChatGPT-4 at $41–78 million, excluding staff salaries that add another 29%–49%.

These cost dynamics directly impact accessibility, operational flexibility, and long-term budgeting for enterprises looking to integrate agentic systems.

1 Access through technology giants

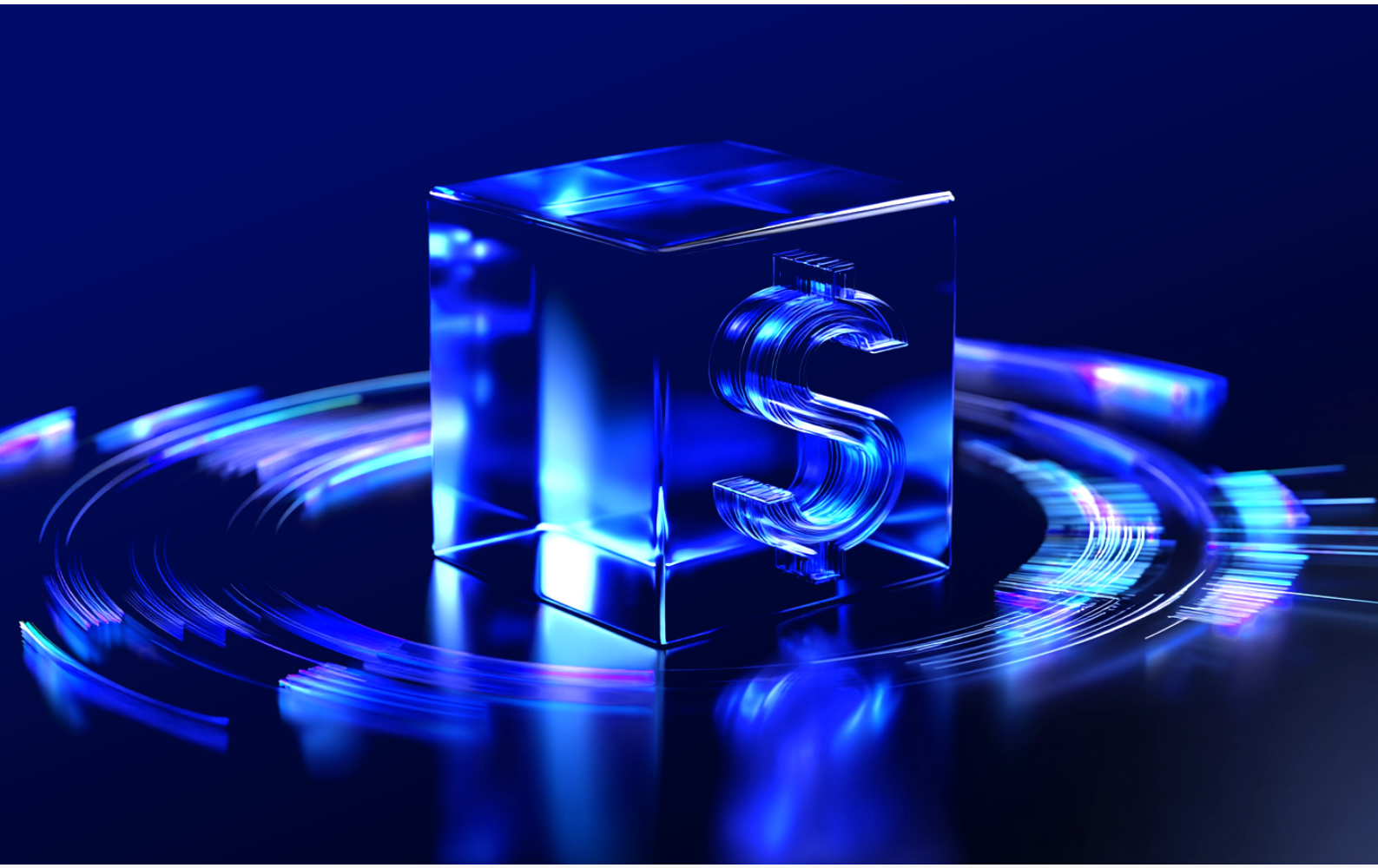
One of the most expensive components of agentic AI development, model training, remains largely absorbed by leading technology companies, including OpenAI, Anthropic, Meta, and NVIDIA. These organizations invest heavily in training large language models and advanced AI systems, shouldering the substantial infrastructure and research costs associated with developing reasoning capabilities.

This cost absorption by major technology companies creates opportunities for enterprises to access advanced agentic AI capabilities without bearing the full burden of model development. However, it also creates dependencies on external providers that organizations must factor into their long-term strategic planning.

To make advanced reasoning models accessible to businesses of all sizes, technology companies offer pay-as-you-go pricing models. This approach allows enterprises to leverage agentic AI capabilities while scaling their usage based on actual business needs rather than making large upfront investments.

Microsoft Copilot Studio exemplifies this approach with agent pricing that offers either pay-as-you-go metering or monthly capacity packs of **$200 for 25,000 agent messages**.

This pricing structure allows organizations to start small and scale their usage as they demonstrate value and build confidence in their agentic AI implementations.



2 Infrastructure alternatives

Selecting the right infrastructure approach is a central decision for organizations deploying agentic AI. Leadership must weigh the pros and cons of building in-house capabilities against relying on external providers. Each option presents its own financial, operational, and strategic considerations that can shape both the immediate and long-term outcomes of an AI investment.

Building in-house infrastructure

Organizations can choose to build dedicated infrastructure to support agentic workloads, gaining greater control over deployment, management, and data security. This approach enables customization of AI systems to meet specific business requirements and reduces dependency on external service providers.

However, in-house infrastructure requires substantial investment across multiple dimensions. Hardware costs include high-performance servers, specialized GPUs, storage systems, and networking equipment designed to handle the computational demands of reasoning models. Software licensing, system integration, and ongoing maintenance add additional layers of expense.

The human capital requirements for in-house infrastructure cannot be understated. Organizations need skilled personnel to design, implement, and maintain AI infrastructure, including specialists in machine learning operations, system architecture, and AI security.

Morgan Stanley and Wall Street Journal reporting project approximately

$2.9 trillion in AI infrastructure spending from 2025 to 2028

with roughly

$400 billion in capital expenditure expected in 2025

across major technology companies alone

3 Trade-offs between in-house and external providers

Organizations must carefully evaluate the trade-offs between the high initial costs of building in-house infrastructure and the long-term benefits of reduced dependency on external providers.

In-house solutions offer greater flexibility, control, and potential cost advantages for high-volume, predictable workloads, but they also require significant ongoing investment in maintenance, upgrades, and security.

External providers through pay-as-you-go models can reduce upfront costs and provide access to cutting-edge capabilities, but may result in higher cumulative expenses over time for organizations with extensive or continuous AI workloads. The decision often comes down to organizational capabilities, risk tolerance, and long-term strategic goals.

Beware of hidden operational costs

Understanding the hidden costs of agentic AI is essential for organizations aiming to deploy these systems responsibly and sustainably. While the upfront investment in infrastructure and model training is often the focus, long-term success depends on recognizing and managing less visible factors. These include the need for robust security, compliance, and governance frameworks that address the unique risks posed by autonomous decision-making. Without these safeguards, organizations may face regulatory challenges and operational vulnerabilities that undermine the value of their AI initiatives.

Agentic workflows consume more computational resources as they scale. Studies demonstrate that multi-step, tool-using agents consume significantly more input tokens and invoke models multiple times, trading higher cost and latency for improved performance and autonomous capability.

Empirical benchmarks confirm this overhead. In software engineering tasks, agentic workflows deliver better planning and verification capabilities but with higher computational cost and latency compared to single-shot prompting approaches.

Usage-based pricing combined with autonomous agent operations can lead to scaling cost challenges. Industry analyses note that token-metered agents can trigger unpredictable, escalating costs as workloads expand and agents autonomously chain multiple steps together. The emergence of "inference whales," heavy users executing long-running agent workflows, has forced some AI vendors to overhaul their pricing structures due to soaring inference costs that were not anticipated in original pricing models.

Operational efficiency also hinges on strategic decisions around model deployment.

Using high-cost models only where complex reasoning is required, and reserving lower-cost models for routine tasks, can significantly reduce expenses.

This requires a deep understanding of task complexity and careful alignment of model capabilities. Similarly, choosing between managed services and custom infrastructure should be based on workload predictability and long-term economic viability. For high-volume environments, investing in dedicated infrastructure may offer better cost control than usage-based pricing models.

Centralized approaches like the AI factory model help mitigate hidden costs by promoting reuse, consistency, and shared expertise across business units. However, organizations must also account for workforce adaptation, change management, and ongoing capability development. Evaluating the total cost of ownership, including these less visible elements, is critical to building a resilient and scalable agentic AI strategy. Readiness assessments that consider technical infrastructure, employee skills, and organizational culture can help ensure that deployments deliver lasting value (more on this in part three of the report).

Start planning

Agentic AI offers substantial economic opportunities for organizations willing to understand and master its unique cost dynamics. Multi-agent systems create new value streams, such as:

Automating Knowledge Discovery

Enhancing Customer Engagement

Optimizing Software Development Processes

For example, agentic AI has accelerated research, improved customer personalization, and streamlined coding tasks, delivering measurable productivity gains across industries like healthcare, financial services, and retail.

However, these advancements come with notable cost considerations, including the high computational demands of reasoning models, substantial infrastructure investments, and the ongoing expenses of model training and maintenance.

Organizations must carefully navigate these cost dynamics to maximize the benefits of agentic AI. The elevated computational requirements, driven by multi-step reasoning and real-time adaptation, can strain budgets, particularly for smaller enterprises. Additionally, businesses face decisions between building in-house infrastructure, which offers control but requires significant capital, and relying on external providers, which reduces upfront costs but may lead to long-term dependencies. Hidden operational costs, such as governance, compliance, and workforce adaptation, further complicate the economic landscape. Strategic planning and readiness assessments are essential to ensure sustainable deployment and long-term value realization.

Read part two and part three in this series to learn more about tokenization and strategies for implementing agentic AI that increase ROI.

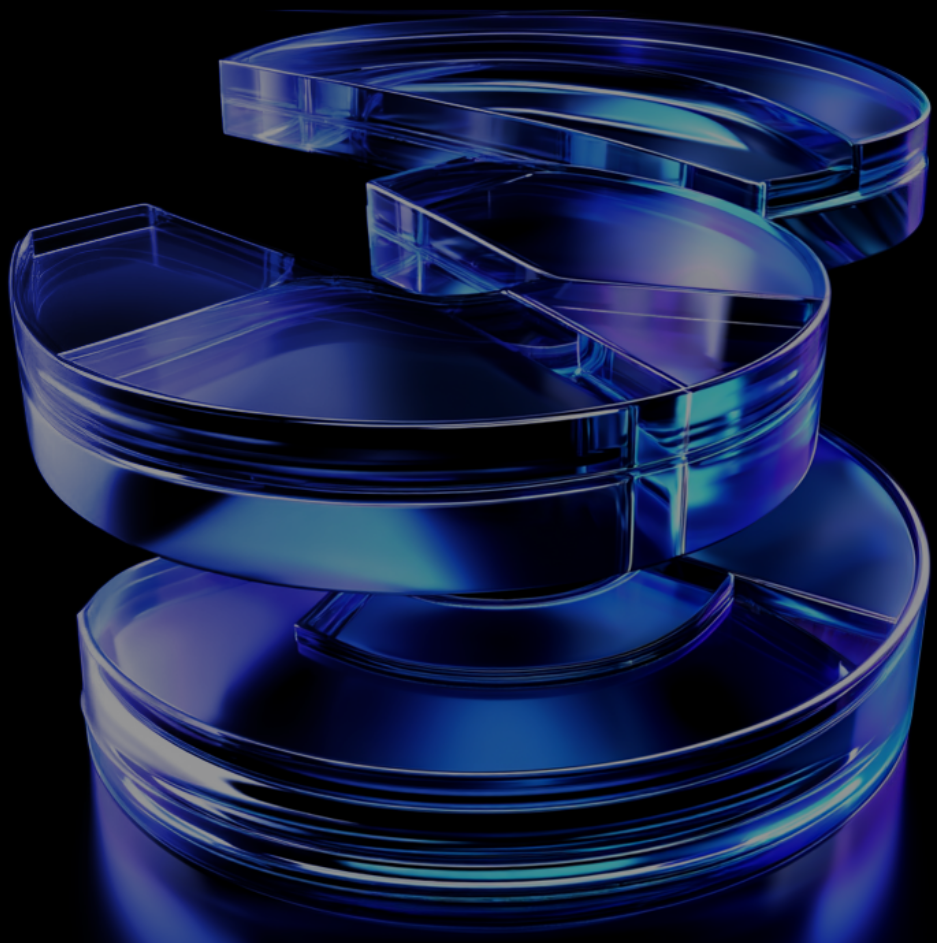
About Us

SoftServe is a premier IT consulting and digital services provider.

We expand the horizon of new technologies to solve today's complex business challenges and achieve meaningful outcomes for our clients. Our boundless curiosity drives us to explore and reimagine the art of the possible. Clients confidently rely on SoftServe to architect and execute mature and innovative capabilities, such as digital engineering, data and analytics, cloud, and AI/ML.

Our global reputation is gained from more than 30 years of experience delivering superior digital solutions at exceptional speed by top-tier engineering talent to enterprise industries, including high tech, financial services, healthcare, life sciences, retail, energy, and manufacturing.

Visit our website, blog, LinkedIn, Facebook, and X (Twitter) pages for more information.



AUSTIN HQ

201 W. 5th Street, Suite 1550
Austin, TX 78701
+1 866 687 3588 (USA)
Toll Free: +1 866 687 3588

LONDON

30 Cannon Street
London EC4 6XH
United Kingdom
+44 203 807 01 41

info@softserveinc.com
www.softserveinc.com

softserve