

Enterprises that recognize how decisions related to data, model selection, and type of AI impact infrastructure cost and performance will be best positioned to drive maximum ROI from AI investments.

# Strategic Infrastructure Decisions Are Key to Delivering GenAI ROI

[April 2025]

**Written by:** Nancy Gohring, Senior Research Director, AI

## Introduction

The introduction of ChatGPT in late 2022 set off a period of intense interest in the potential to harness the power of generative AI (GenAI) in the enterprise. Organizations around the world across sectors have been experimenting with and preparing to scale up the use of GenAI technologies. The broad adoption will drive notable benefits, with IDC estimating that AI will generate a cumulative global economic impact of \$19.4 trillion by 2030.

However, over the past 18–24 months, enterprises have taken a fragmented approach to developing and deploying GenAI, resulting in a relatively low rate of conversion from trial to production. A number of factors have held back deployments, with IDC's *Global GenAI Technology Trends Survey* revealed that the most common reason GenAI initiatives did not make it to production was that excessive costs prevented them from meeting ROI objectives (see Figure 1).

In addition, nearly 37% of respondents cited inadequate infrastructure performance or availability as another deployment challenge. One key to improving ROI of GenAI deployments is careful consideration of infrastructure options available once an application is deployed to production. Volume and geography of end users, location of data, and workload variability are among the factors that can impact cost and performance in ways that organizations may not consider before deployment. Unexpected infrastructure costs can skew ROI estimates that were calculated during proof of concept (POC) or trial phases.

## AT A GLANCE

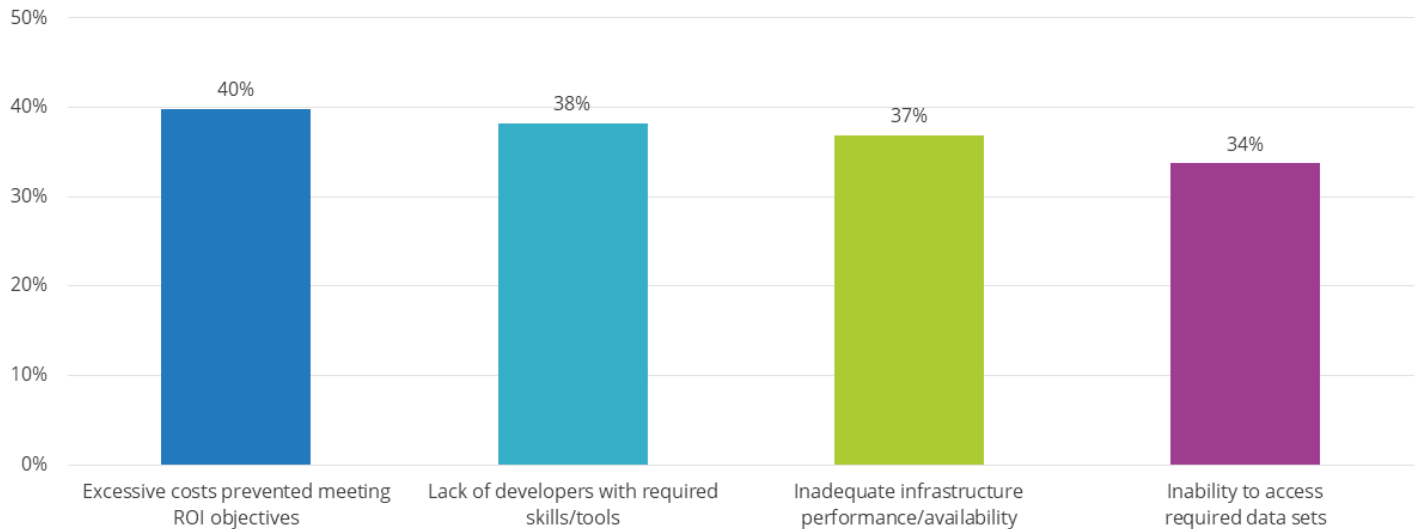
### KEY TAKEAWAYS

The most common reasons GenAI initiatives do not make it to production:

1. Excessive costs prevented meeting ROI objectives
2. Lack of developers with required skills/tools
3. Inadequate infrastructure performance/availability
4. Inability to access required data sets

FIGURE 1: *Challenges That Hold Back Production*

**Q What was the most common reason that your organization's GenAI initiatives did not make it to production?**



Source: IDC Global GenAI Technology Trends Survey, July 2024, n = 624

Organizations that make smart decisions related to infrastructure selection will be best positioned to reduce costs and to achieve ROI from their AI investments. To select the right infrastructure for AI workloads, organizations should examine the following components of the AI application and consider the implications related to infrastructure selection:

- » **Data:** A range of enterprise data may be useful in training, tuning, and adding knowledge to AI applications. Because valuable enterprise data may be structured, unstructured, or semistructured, many organizations require investments in preparing and managing data for use by AI applications.

Once those investments are made, privacy and security requirements have a notable bearing on infrastructure decisions since some data will be best retained in an on-premises environment to meet compliance and regulatory requirements while other data may already be stored in cloud environments. In addition, data volume, data timeliness, and the location in which the data is generated are all relevant factors that should shape infrastructure decisions and that impact cost.

For instance, if most corporate data resides on-premises, a retrieval-augmented generation (RAG) implementation that runs on premises may deliver the least latency, the highest levels of security, and lowest cost, compared with using cloud resources. A hybrid architecture, in which some data is tapped from the cloud and other data sets remain on premises might be the solution that best meets privacy and security demands, while delivering the desired performance and cost efficiency.

- » **Models:** Model selection also has a bearing on where an AI application can and should run. A small, purpose fit model that can be run on premises, for instance, could result in considerable savings when compared with large language model (LLM) that runs in the cloud. Organizations are increasingly turning to small language models and understanding the best infrastructure to run the model can have a notable impact on AI application cost and performance.

» **Type of AI:** GenAI and traditional predictive or prescriptive AI have different infrastructure demands; not every AI app requires graphics processing units (GPUs). For instance, even GenAI inferencing workloads, such as those with a relatively low number of users or that use relatively small models, may not need a GPU. Fitting the use case to the appropriate processor or accelerator can lead to notable savings.

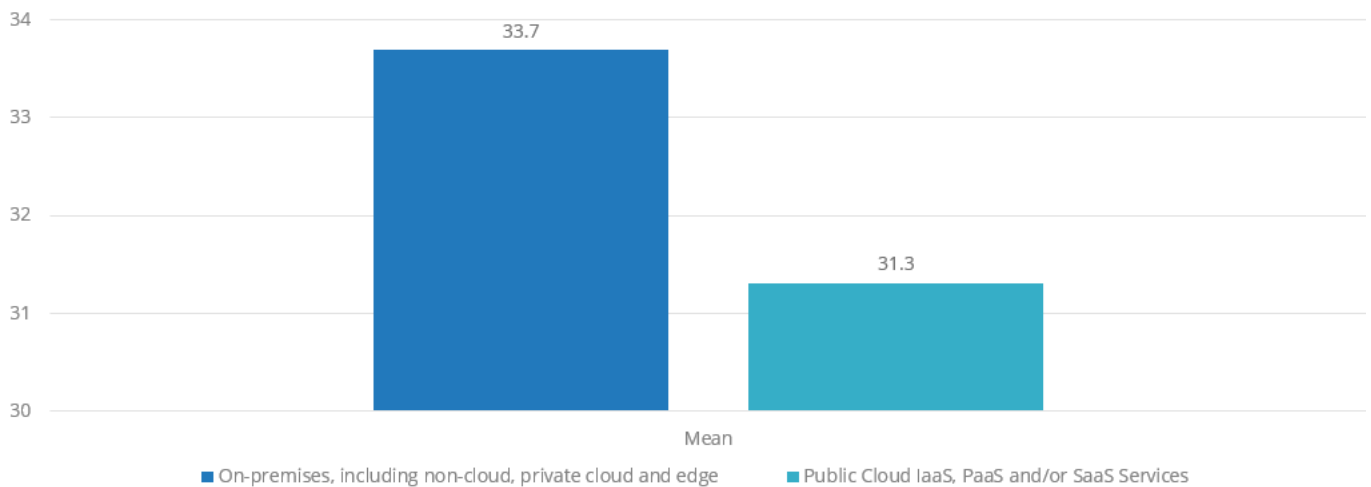
For applications that benefit from the use of GPUs, there are strategies that can help contain costs. Resource management tactics including GPU fractioning, resource policy management, and other AI workload orchestration tools can improve efficiency in ways that can reduce required spending.

Smart decision making in selecting the right infrastructure to support the type of AI can have a significant impact on cost. In some scenarios, an existing on-premises environment might best support an AI workload and in others, a cloud service with scalable GPU access, might meet the needs of an application. For some AI applications that combine traditional AI with GenAI, a hybrid approach may be the best solution.

Organizations are already taking advantage of an array of options when it comes to infrastructure that supports GenAI workloads. Although respondents to a recent IDC study said that they expected to deploy on average 31% of production GenAI workloads and data in the public cloud, they expect the bulk of their GenAI workloads and data to run in dedicated, noncloud; dedicated private cloud; or dedicated edge locations (see Figure 2). The finding indicates that organizations are already experimenting with fitting the right workload to the right infrastructure.

FIGURE 2: *AI Workloads Will Run in a Variety of Environments*

Q *What is your best estimate of where your organization will deploy production GenAI workloads and data across the digital infrastructure deployment options? (mean)*



Source: IDC FERS Wave 1, 2024, n = 881

Each infrastructure environment offers benefits and drawbacks and comes with cost implications. Infrastructure as a service (IaaS) may offer an easy on-ramp and pay-for-what-you-use pricing. An on-premises datacenter that is carefully managed could reduce costs and support requirements depending on where data is stored, where users are located and security is needed. For very low-latency demands, edge locations might satisfy both performance and cost requirements.

## Benefits

Leading organizations in various sectors have demonstrated that applying AI to the right use cases can result in notable productivity gains that cut costs. But realizing those benefits will require a disciplined and strategic approach in order to meet or exceed ROI expectations. Organizations that carefully weigh the ways that data, model selection, and types of AI are related to infrastructure decisions will be best positioned to contain costs and drive ROI. A positive ROI is the difference between a successful or a failed AI implementation, and organizations with successful implementations are positioned to win against the competition and even to disrupt industries.

In many cases, hybrid will be the right infrastructure approach to support AI workloads. There are a number of benefits to hybrid deployments including:

- » **Security:** Using tightly controlled on-premises environments to manage the security, privacy, and regulatory requirements of sensitive enterprise data, while deploying other application workloads in the cloud, may be the most cost effective approach to protecting valuable enterprise data.
- » **Scalability:** With sensitive data resources securely deployed on-premises, enterprises can employ cloud resources to meet the scale requirements of other application functions including AI inferencing, for example.
- » **Cost-efficiency:** A hybrid approach has a number of opportunities to reduce costs. For instance, some applications may benefit from batch workloads that can be strategically executed in the cloud to reduce costs. Or AI applications can burst to the cloud when additional scale is required.
- » **Access to new innovation:** Cloud providers offer some capabilities that are otherwise difficult or even impossible to access in an on-premises environment. A hybrid architecture allows enterprises to access these capabilities, while still benefiting from the use of on-premises environments for other reasons, including data security and privacy.

## Technology or Vendor Profile

HCLTech AI Foundry is a full-stack suite of services and software that delivers a cohesive blueprint for data, infra, and AI to scale AI-led outcomes across the enterprise value chain. This bundled, modular offering brings together process templates to accelerate AI adoption responsibly at scale, data assets to make estates ready for AI, and build the pipelines to turn data to actionable insight, prebuilt AI applications that deploy faster and accelerate time-to-value, all backed by flexible, build-and-run capabilities for the foundational infrastructure needed to operate effectively and efficiently.

Cohesive, outcome-led and ecosystem-configurable with a network of leading hyperscalers, OEMs, and platform providers: AI Foundry offers industry-specific frameworks, accelerators, hardware, and software tools that help businesses simplify AI journeys, moving from experimentation to production faster. Supported by robust governance and integrated data, infra, and AI operations, AI Foundry scales with your business needs, delivering outcomes at scale, across the value stream.

HCLTech's Cognitive Infrastructure Services, offered as part of AI Foundry, represent the infrastructure foundation that supports the AI Foundry services suite of offerings. With a focus on enabling scalability, security, and efficiency, Cognitive Infrastructure Services is designed to support the overall adoption and success of GenAI at scale and drive innovation across industries. HCLTech offers integration with existing customer infrastructure in order to drive rapid AI adoption without disruption.

HCLTech Cognitive Infrastructure Services includes:

- » **Advisory and assessment:** This offering includes assisting enterprises in identifying AI use cases and feasibility, assessment of current infrastructure environment and path to an ideal state, help with buy versus build decision making, cost analysis, security consultation, and more.
- » **Platform and infrastructure build services:** HCLTech can configure and build an AI platform and hybrid infrastructure. It can also prepare enterprise data for use by AI applications, including data engineering services, and advise on whether RAG or fine tuning is the best approach, and then build the required infrastructure.
- » **Operate and manage services:** Once AI applications are in place, they require ongoing operations and management. HCLTech can perform availability and performance monitoring, infrastructure monitoring, machine learning operations (MLOps) and large language model operations (LLMOps) services, and security and compliance services.
- » **Flexible delivery models:** Enterprises can purchase compute, storage, backup, and GPU as a service from HCLTech as part of Cognitive Infrastructure Services.

This can go into flexibility whether on-prem, hybrid cloud, edge, and so forth — all optimized for cost, and ability to integrate with leading cloud hyperscalers (Microsoft Azure, AWS, IBM, and Google Cloud) and physical infrastructure providers (such as Dell, NVIDIA, HPE, IBM, Cisco), ensuring that Cognitive Infrastructure part of HCLTech AI Foundry can be tailored to meet the unique needs of each organization. Whether piloting select AI use cases or moving into full production, HCLTech also offers T-shirt-sized solutions tailored to the needs and budgets of different organizations. This ensures controlled and streamlined experimentation and inferencing, fitting deployments to match the scale of your AI maturity and aspirations.

### Challenges

Most enterprises lack the skills and expertise to quickly deploy AI technology that delivers business outcomes and ROI. Partners are key, especially in guiding enterprises through the complex decision-making process related to hybrid AI deployments. However, the market is very crowded and moving rapidly, creating difficulties for enterprises in selecting the right partners that have the expertise, capacity, technology, and capacity to respond. Technology and services providers also face a number of challenges in meeting the needs of enterprises, including their own lack of internal expertise and the requirement to build differentiated, AI-driven products and services, in a hypercompetitive environment.

### Conclusion

AI and GenAI innovations are moving at an incredibly fast pace and will continue to do so. Enterprises that are able to scale the adoption of AI have the potential to establish competitive advantage. The challenges are significant though, and the complexity of the AI landscape demands a strategic approach to laying the groundwork for deployments that balance cost and returns. Careful consideration of the impact that moving AI workloads into production has on infrastructure costs will determine whether investments in AI deliver the returns that enterprises require. The stakes are high: organizations that make the right infrastructure decisions that support the growing adoption of AI into the future will be best positioned for success.

## About the Analyst



### ***Nancy Gohring, Senior Research Director, AI***

Nancy Gohring is a senior research director, co-leading IDC's Generative AI Strategies program alongside Ritu Jyoti. Nancy covers big picture trends related to enterprise AI adoption (including GenAI). Key research themes include business, organizational, and technology architecture transformation, in the context of AI and GenAI. As part of the worldwide AI, automation, data, and analytics research practice, Nancy supports a range of clients across the technology stack.

### MESSAGE FROM THE SPONSOR

HCLTech is a global technology company, home to more than 220,000 people across 60 countries, delivering industry-leading capabilities centered around digital, engineering, cloud, and AI. We work with clients across all major verticals, providing industry solutions for Financial Services, Manufacturing, Life Sciences and Healthcare, Technology and Services, Telecom and Media, Retail and CPG, and Public Services.

We help businesses navigate the complexities and opportunities of GenAI and Agentic AI through our flagship offerings—AI Force, AI Foundry, AI Labs, and AI Engineering—enabling service transformation and value stream innovation.

As part of AI Foundry, HCLTech Cognitive Infrastructure Services provides a full-stack solution for scalable GenAI adoption. From ideation and LLM fine-tuning to enterprise-grade deployment, we accelerate innovation and hyper-automation using a validated ecosystem and flexible delivery models. Discover how HCLTech can accelerate GenAI adoption for your organization at [Cognitive Infrastructure Services & Solutions | HCLTech](#)



The content in this paper was adapted from existing IDC research published on [www.idc.com](http://www.idc.com).

**IDC Research, Inc.**  
140 Kendrick Street  
Building B  
Needham, MA 02494, USA  
T 508.872.8200  
F 508.935.4015  
blogs.idc.com  
www.idc.com

IDC Custom Solutions produced this publication. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis that IDC independently conducted and published, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies. This IDC material is licensed for external use, and in no way does the use or publication of IDC research indicate IDC's endorsement of the sponsor's or licensee's products or strategies.

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives.

©2025 IDC. Reproduction is forbidden unless authorized. All rights reserved. [CCPA](#)

