SPOTLIGHT
Sponsored by: Broadcom

A private AI approach in highly regulated industries can enhance efficiency and compliance, while addressing challenges related to data privacy, security, and regulatory adherence.

Realizing the Value of GenAI in Regulated Industries While Controlling Costs and Risks

December 2024

Written by: Nimita Limaye, Research Vice President, Life Sciences R&D Strategy and Technology; Massimiliano Claps, Research Director, Worldwide National Government Platforms and Technologies; and Jerry Silva, Program Vice President, IDC Financial Insights

Introduction

Industries, from government to life sciences and healthcare, from banking to insurance, from retail to manufacturing, have used AI capabilities for years. However, most implementations were focused on selected use cases, such as anomaly detection supporting cybersecurity and anti-fraud analyses. The advent of GenAI prompted a surge of experimentation in both public and private sectors to drive automation of tasks, provide virtual assistants for contact center agents, develop and test software code, summarize meeting minutes, screen media and news, and countless more.

AT A GLANCE

WHAT'S IMPORTANT

- Organizations in financial services, government, healthcare, and life sciences are seeking to leverage the full potential of GenAI, while complying with all regulatory and ethical requirements.
- Many organizations in key regulated industries are expecting to use private AI built on dedicated on-premises/colocation computing and storage infrastructure for GenAI training, tuning, and inferencing.

Presently, public and private sector executives across industries, including regulated industries such as government, healthcare, life sciences, and financial services, have started to move beyond pilots and proofs of concept to realize the value of GenAI at scale. Due to this growing adoption, IDC forecasts that worldwide GenAI spending in financial services, healthcare, life sciences, and government will rise from $13.2 billion in 2024 to $72.7 billion in 2028.

Chief information officers, chief artificial intelligence officers (CAIO), and line-of-business leaders are collaborating to understand how to automate business processes end to end. They want to combine automation agents (including robotic process automation [RPA], AI, and GenAI) to reduce time to market for product and service innovations, improve financial performance, optimize operations, drive smarter decisions, and enhance customer/patient/constituent experiences.

Executives are now rethinking their strategy, governance, teams, and technology to adopt GenAI effectively. This transformation will require establishing senior leadership roles that can build AI organizational capacities, competencies, and literacy; designing and enforcing governance policies, structures, and processes for responsible AI innovation; and deploying data and AI infrastructure and platform capabilities aligned with the use cases they want to scale.

Private AI Framework and Adoption

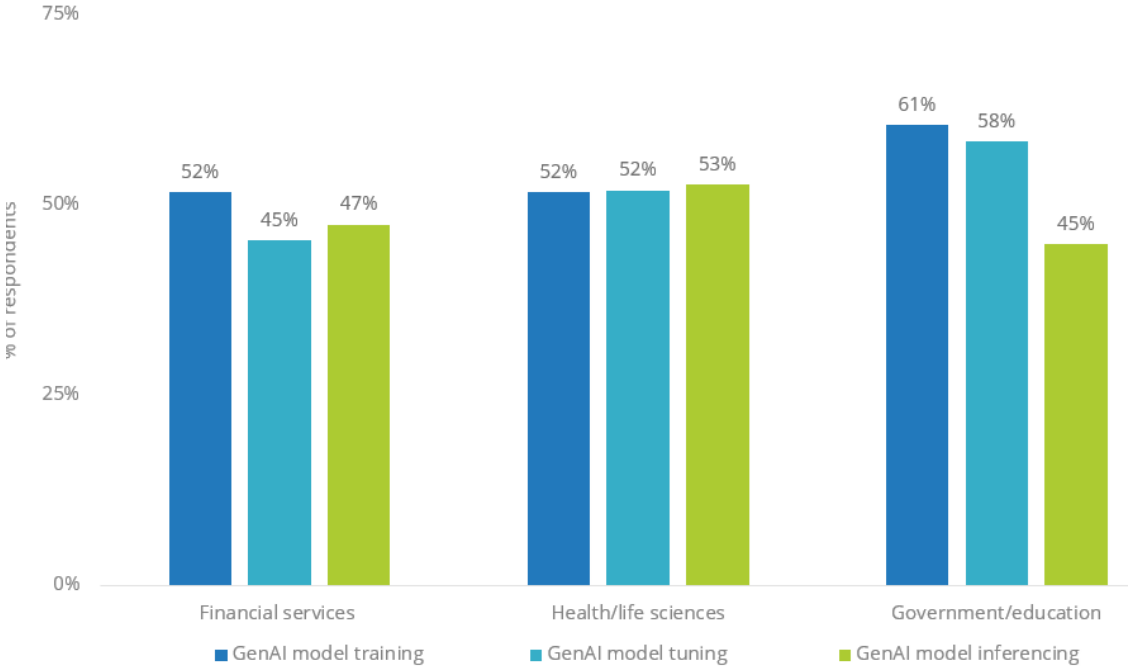
A private AI framework is an architectural approach that enables organizations to deploy and manage AI workloads in private and hybrid cloud environments. It focuses on delivering secure, scalable, and efficient AI capabilities, allowing organizations to build, train, and run AI models wherever critical and private business data is stored. This framework integrates with various AI software and hardware components, providing flexibility, control, and compliance with data protection regulations.

Private AI infrastructure refers to the dedicated hardware and software resources deployed within an organization's datacenter or in a third-party colocation facility. It includes high-performance computing servers, storage systems, networking equipment, and specialized AI accelerators. By maintaining control over the physical infrastructure, organizations can optimize performance, reduce latency, and meet specific regulatory requirements while leveraging advanced AI capabilities for their business operations.

A foundational aspect of a private AI infrastructure is dedicated on-premises/colocation computing and storage infrastructure for GenAI training, tuning, and inferencing. As shown Figure 1, roughly half of organizations in key regulated industries expect to primarily use dedicated on-premises/colocation computing and storage infrastructure for GenAI training, tuning, and inferencing.

FIGURE 1: ***Use of Dedicated On-Premises/Colocation Computing and Storage Infrastructure for GenAI***

Q *Over the next 18 months, what will be the primary type of computing and storage infrastructure used by your organization to support initial Gen AI model training/tuning/inferencing?*



n = 1,000

Source: IDC's Future Enterprise Resiliency and Spending (FERS) Survey, Wave 4, April 2024

Considerations and Benefits of Private AI

Regulated industries need to maintain highly confidential data with regulations on control, access, and security. Data is often disparate, from proprietary sources, and stored on-premises or in access-controlled or classified network enclaves. Private AI allows government, healthcare, life sciences, and financial services to bring AI models to the data sources they already have, maintaining privacy, governance, and controls that are already in place — using their existing toolset. Private AI can help CAIOs and other IT leaders accelerate the time to value of AI and GenAI while controlling its risks and costs. As leaders embrace private AI, they should consider the following:

- The need to invest in capacity and competencies to translate national policies into solutions that enable them to implement private AI with appropriate data governance controls and cybersecurity solutions
- The need to align private AI capabilities with use cases that have attributes, such as predictable scalability and performance requirements, and multiple interoperability dependencies that justify investment
- The need to invest in AI and data architecture and engineering competencies that enable best-in-class capabilities of private AI platforms and integrate them with workloads running in other environments

Leaders should also consider the benefits of a platform approach that allows increased flexibility to experiment with and utilize new AI models and services as market conditions change. These platforms should come with built-in automation and tools, significantly reducing the necessity for maintaining specialized internal skill sets to ensure success. By strategically investing in these areas and leveraging a platform approach, government CAIOs, and IT leaders can maximize the benefits of private AI while effectively managing its risks and costs

VMware Private AI Foundation with NVIDIA

Broadcom and NVIDIA have collaborated to develop the joint GenAI platform, VMware Private AI Foundation with NVIDIA. This joint GenAI platform enables enterprises to fine-tune LLM models, deploy RAG workflows, and run inference workloads in their datacenters, addressing privacy, choice, cost, performance, and compliance concerns. This joint platform simplifies Gen AI deployments for enterprises by offering self-service and automation to enable deployment of AI services in minutes, AI model governance to ensure that only approved models make it to production environments, a vector database for similarity searches, and GPU management and optimization.

Built and run on the private cloud platform, VMware Cloud Foundation (VCF), VMware Private AI Foundation with NVIDIA enables organizations to take advantage of the latest NVIDIA Inference Microservices (NIMs) that offer turnkey AI services for a variety of business use cases, and organizations can download the latest open source models from Hugging Face or use a variety of third party commercial applications. VCF offers a secure and scalable environment for building and operating GenAI workloads, providing organizations with agility, flexibility, and scalability to meet their evolving business needs. VCF empowers enterprises to easily integrate existing data pipelines, workloads, internal AI applications, and ISV applications onto a common platform and unify both infrastructure resources and data. VMware Private AI Foundation with NVIDIA is an Advanced Service for VCF.

By leveraging private AI, organizations and agencies can bring AI models directly to their data, ensuring that both data and AI insights remain private while maintaining full control, compliance, and resilience. VMware Private AI is an architectural approach built on these principles:

- **Secure:** Modern encryption protects training sets, model weights, and inference data, ensuring data confidentiality at rest, in transit, and in use. Agencies can benefit from GenAI while upholding existing privacy, security, and compliance standards.
- **Flexible:** An open ecosystem enables governments to quickly deploy and operate their AI models of choice.
- **Future-proof:** Investing in AI infrastructure through an open ecosystem allows organizations to operate at the speed of software, quickly onboarding new AI models or services as business needs require.
- **Accurate and reliable AI models:** Utilizing retrieval-augmented generation (RAG), AI models can fetch facts from internal sources, enhancing their accuracy and reliability.
- **Unified management and operations:** Management and operations are streamlined across private AI and other enterprise services, reducing total cost of ownership, complexity, and additional risks.
- **Time to value:** Built-in AI tools and automation allow AI environments to be quickly set up and dismantled, ensuring immediate resource availability.
- **Lower cost and increased efficiency:** Private AI infrastructure virtualizes and intelligently shares resources (GPU, CPU, memory, and networks) across multiple AI applications and services, reducing costs and enhancing efficiency.

Challenges

Private AI offers data security and privacy but does not substitute for strong governance, regulatory compliance, and risk management practices for enterprise data more widely. IT organizations can and must evolve to prioritize governance, security, and compliance over simply technical delivery. Investments in these areas are as essential, arguably more so, than ever before because regulators will be scrutinizing data throughout the life cycle.

Evolving AI Regulatory Frameworks

Private AI can help organizations align and comply with various regulatory frameworks that cover each industry or sector. Life sciences and healthcare industries already need to ensure compliance with multiple regulations, including:

- 21 Code of Federal Regulations (CFR) Part 11
- Health Insurance Portability and Accountability Act (HIPAA)
- 21st Century Cures Act
- Health Information Technology for Economic and Clinical Health Act (the HITECH Act)
- ISO 9001:2015, ISO/IEC 27001:2013, and ISO/IEC 26580:2021
- GxP compliance the General Data Protection Regulation (GDPR) as well as overall compliance

- State regulations such as the California Consumer Privacy Act (CCPA) and the Connecticut Data Privacy Act (CTDPA)

Senior leaders in regulated industries need to consider the rapidly evolving AI regulatory frameworks as well, such as:

- In January 2023, the U.S. National Institute of Standards and Technology (NIST) released its Artificial Intelligence Risk Management Framework (AI RMF 1.0) for all U.S. government organizations and agencies.
- In October 2023, U.S. President Joe Biden released an Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, directing new standards for AI safety and security.
- In November 2023, the U.S. Department of Defense (DoD) released its strategy to accelerate the adoption of AI within the DoD to "increase the efficiency of DoD business operations, make … warfighting capabilities and the people who command them more effective, and create opportunities to employ novel operational concepts."
- In March 2024, the U.S. Office of Management and Budget (OMB) released M-24-10: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence. This Memorandum provides guidance to all federal agencies consistent with the AI in Government Act of 2020, the Advancing AI Act, and Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.
- In March 2024, the U.S. Department of Homeland Security released its Artificial Intelligence Roadmap, which aims to enhance national security by ensuring responsible and ethical AI deployments
- In May 2024, the European Council approved the EU Artificial Intelligence Act. This regulation follows a risk-based approach, banning unacceptable risks such as social scoring and emotion recognition in workplaces. High-risk AI systems must comply with stringent standards, including risk management, data governance, and human oversight.
- In August 2024, the European Commission's AI Pact entered into force. It is a voluntary framework that runs parallel to the EU Act that encourages companies to adhere to ethical guidelines and best practices before mandatory regulations of the AI Act come into effect.
- In September 2024, OMB released M-24-18: Advancing the Responsible Acquisition of Artificial Intelligence in Government. The Memorandum provides guidance to federal agencies on procuring artificial intelligence.
- The European Commission has begun establishing an EU-level AI governance structure. In February 2024, it created the AI Office under the Directorate-General for Communications Networks, Content and Technology (DG CNECT) of the European Commission. The AI Office will be responsible for discharging the Commission's functions in relation to the implementation, monitoring, and supervision of AI systems and general-purpose AI models and responsibilities in accordance with the AI Act.
- South Korea's comprehensive AI act, which is under review by the National Assembly, includes a regulatory framework, standards on copyrights of AI-generated content, and guarantees developer access to AI technologies. It aims to promote growth within South Korea's AI industry while enforcing stricter requirements on high-risk systems.

- The G7 launched the Hiroshima AI Process under Japan's presidency in 2023, a comprehensive policy framework focused on analyzing AI risk, developing guiding principles for AI, creating a code of conduct, and promoting cooperation in support of responsible AI solution development.

As the strategic impact of GenAI grows and regulatory efforts advance, public and private sector senior leaders need to work with technology and industry partners to explore alternative pathways to adopt AI securely, on high-performing, scalable, and secure infrastructure and platforms.

Priority GenAI Use Cases in Regulated Industries

Regulated industries are beginning the move from early-stage experimentation to full production deployments. To achieve this level of automation, CAIOs need to work with line-of-business leaders to reengineer processes and systems so they can apply algorithms that recognize changes in customer/patient/constituent circumstances, identify the root causes, and trigger operational workflows or dynamically reconfigure services and programs to meet customer-evolving needs and preferences. This level of automation will require a combination of agents that will provide multimodal capabilities to process text, rules, and images and will be orchestrated to deliver intended outcomes for tangible use cases that can optimize and automate complex operational, customer-facing, and strategic decision-making workflows.

Financial Services

In the highly regulated financial services industry, the adoption of GenAI is carefully controlled to ensure customer data security, minimize institutional risks, and comply with numerous regulations across banking, capital markets, and insurance.

Financial institutions worldwide are experiencing benefits of GenAI, including faster decisioning, improved efficiencies, and increased accuracy. This last benefit helps to decrease the high costs of human intervention. GenAI is finding its most successful use cases in customer experience, risk management, marketing and sales, disputes and reconciliation, and business operations, with more use cases being piloted, like product pricing and fraud detection. Equally important, financial institutions cited IT and cloud infrastructure automation and support as the area of the biggest positive impact from GenAI (source: IDC's *Future Enterprise Resiliency and Spending Survey, Wave 4,* April 2024).

Government

In government, GenAI adoption means building on the benefits of traditional AI use cases — particularly in tax and revenue agencies, health and human service agencies, homeland security, defense, and intelligence — to hyper-personalize services and benefits, reduce the burden of tax compliance, streamline casework, simulate the impact of policy decisions, and protect critical national infrastructure, while balancing innovation with privacy and civil rights protections.

Life Sciences and Healthcare

In the life sciences industry, the maximum impact of GenAI is expected in product and software design, followed by marketing and customer experience. When it comes to what is being deployed today, quality assurance, risk management, and compliance lead the way, with 96% of R&D and half of commercial prioritizing these use cases. Half of the life sciences industry is focusing its R&D efforts on the use of GenAI for clinical trial optimization and drug discovery (source: IDC's *Life Sciences Generative AI Survey,* August 2024), with the ultimate goal of decreasing the time and cost of bringing new drugs to market.

For healthcare, the near-term adoption of GenAI will support hospitals with relieving the administrative burden on physicians and optimize the workflows of lean staffing, while longer-term developments may see GenAI playing a larger role in clinical diagnostics. While the life sciences and healthcare industries are embracing the use of AI with open arms, data security and privacy remain a top concern as regulatory compliance and patient trust are critical for these industries.

Conclusion

Realizing the immense promise of AI in regulated industries requires balancing innovation with the need for security and compliance. By prioritizing AI infrastructure platforms and deployment with strong security and compliance, companies can confidently explore innovative AI applications that drive efficiency and growth while adhering to ethical and legal obligations. This framework has the potential to not only mitigate risks associated with data breaches and privacy violations but also foster a more trustworthy environment for both businesses and consumers.

Private AI solutions can help organizations empower customers/patients/constituents and employees with improved usability and automation, navigate AI responsibly, and maximize returns on AI investments, all while maintaining security and privacy. Private AI solutions can help securely deploy GenAI solutions without increasing the risk of sensitive data loss to ensure that innovative solutions are secure and future proofed. The path forward will require a concerted effort from all stakeholders, including technology developers, regulatory bodies, and industry leaders, to create a balanced ecosystem where innovation thrives within a framework of robust security and compliance.

About the Analyst

Dr. Nimita Limaye, *Research Vice President, Life Sciences R&D Strategy and Technology*

Dr. Nimita Limaye provides research-based advisory and consulting services as well as market analysis on key topics related to R&D Strategy and Technology in the life sciences industry. Her research focuses extensively on AI/GenAI. She is the recipient of the 2024 DIA Global Inspire award and is the past chair of the board of SCDM.

Massimiliano Claps, *Research Director, Worldwide National Government Platforms and Technologies*

Massimiliano Claps is the research director for the Worldwide National Government Platforms and Technologies research in IDC's Government Insights practice. In this role, Max provides research and advisory services to technology suppliers and national civilian government senior leaders in the United States and globally.

Jerry Silva, *Program Vice President, IDC Financial Insights*

Jerry Silva is vice president for IDC Financial Insights responsible for the global retail banking practice. Jerry's research focuses on technology trends and customer expectations and behaviors in retail banking worldwide. Jerry draws upon over 35 years of experience in the financial services industry covering a variety of topics, from the back office to customer channels to governance in the technology shops at financial institutions. His work for both institutions and vendors gives Jerry a broad perspective on technology strategies.

MESSAGE FROM THE SPONSOR

VMware Private AI is an architectural approach from Broadcom that lets organizations unlock the business gains of AI while meeting their privacy and compliance requirements. It is a privacy-first approach to AI that provides choice of commercial and open source AI models and services.

Learn More at vmware.com/privateAI

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.

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