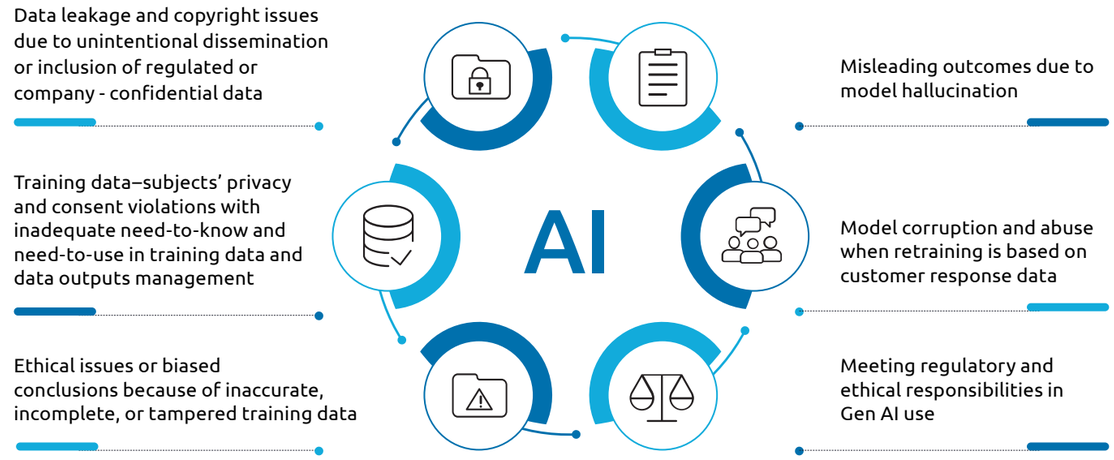
Generative AI and Cybersecurity:
A revisited classic

Capgemini

Gen AI follows a familiar pattern for adoption and cybersecurity, prompting questions reminiscent of those that accompanied the early days of cloud computing. The rapid rise of generative AI presents organizations with the usual innovation dilemma: is it better to adopt a cautious and restrictive approach, risking missing out on opportunities, or to grant more freedom, at the risk of exposing themselves to new risks?

The greatest risks when incorporating generative AI into a business structure are:



The biggest risks are to data

When designing for secure generative AI, data risks take priority. Broadly speaking, these risks originate from three activities:

The exposure of confidential and/or regulated information

Inaccurate information disrupts processes, whether decisional or operational

Potential reputational damage is caused when Gen AI tools are used as chatbots serving as interfaces between customers and an organization

These risks have common themes of identifying, scrubbing, and protecting the right data at the right time and putting the right guardrails in place around a Gen AI solution. Despite its potential and the excitement surrounding it, Gen AI is ultimately another enterprise tool: it requires the application and adaptation of policies, controls and measures implemented at enterprise level and within the AI ecosystem. It brings challenges of operating models internally and monitoring their input and output compliantly.

In a Gen AI system, foundational security must be done across four dimensions:

- Framework, governance, and risk management
- Data and identity security
- Trusted Gen AI models and their outcomes
- Infrastructure and application monitoring and delivery

Threat models are available from NIST, MITRE, Microsoft, Google, and others in the industry to build faster and be ready for new risks.

A Gen AI system can have different security scopes. Using cloud service providers (CSP) as examples, each CSP (also known as hyperscalers) offers generative AI systems with very different security scopes, and each provider defines this scope differently. Consider shared responsibility around the reference architecture found in Figure 1.

Various use cases that matters to the end users and are relevant business cases

Applications
Software applications that primarily use Gen-AI models to perform a task

Monitoring & Maintain
Monitor performance, user experience and outcome quality

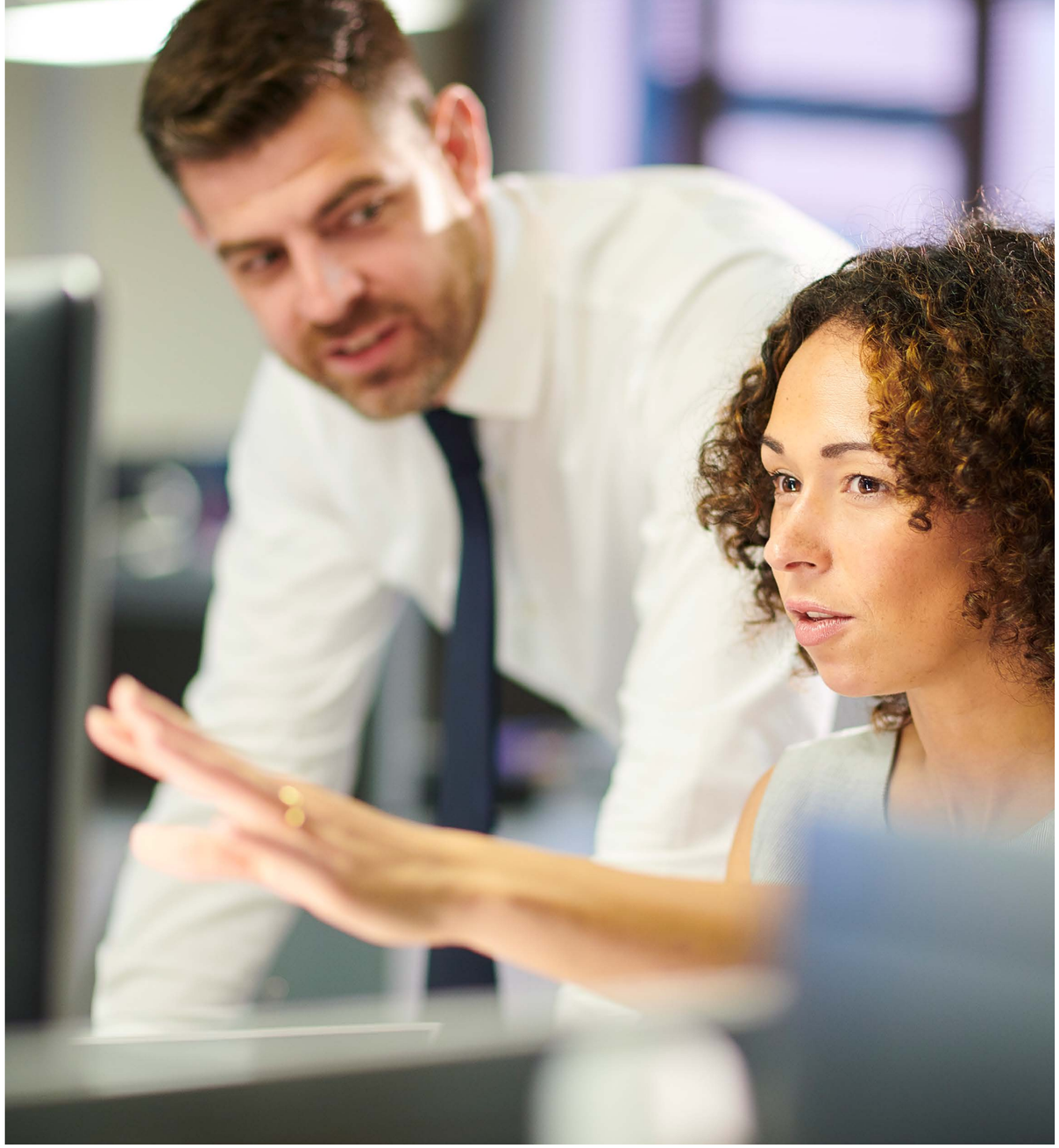
Models & Tools
Gen-AI foundation models & domain specific models that are tailored to a given industry or use case Tools to operationalize Gen-AI models

Data
Data collection, data preparation and transformation

Infrastructure
Infrastructure components used to build out Gen-AI applications such as compute, network and storage

Compute | Network Communication | Storage

Figure 1: Conceptual reference architecture for Gen AI shared responsibility.



Amazon Web Services focuses on providing the infrastructure for generative AI models, as with Amazon Bedrock. Various degrees of customization and ownership are possible. The client's system is defined as the provided infrastructure, and their part of shared responsibility includes the security of the models, data, and applications.

Google Cloud Platform's (GCP) approach focuses on the infrastructure and models, offering Vertex AI and the Model Garden to empower customers. Customers focus on the application layer, monitoring, and the Gen AI interface, while GCP has shared responsibility from the model down to data and infrastructure.

With Microsoft Azure's Co-pilot, the CSP takes ownership of infrastructure, model, application, and everything in between.. The customer focuses on data security and business purposes. Data interfaces define their system, while the models, infrastructure, and application interface are treated more as black boxes.

Establishing a security framework with governance

Positions on how to regulate Gen AI vary widely, from outright prohibition to complete laissez-faire. No single government or supranational political entity will be able to dictate how Gen AI proliferates. Nevertheless, enterprises must work within legal and regulatory structures based on their clients, geographies, and ethics.

To anticipate what's expected in generative AI governance, enterprises should consider the following:

- Existing and upcoming regulations that will influence AI use
- An enterprise's unique risk tolerances for technology and regulations
- Team member education on how Gen AI works, its inherent problems, and risks such as data leaks and the organization's own policies
- A secure Gen AI reference architecture describing how to manage risks

The reference architecture must address the risks of various models in diverse ways. A full proprietary solution, including Gen AI model development and pre-training, means an organization will have the ability and obligation to address its specific risks end-to-end.

In the case of Software-as-a-Service generative AI, many risks need to be addressed through contract and third- and fourth-party risk management. Organizations can also deploy more than one Gen AI solution with different architecture models, and hybrid models.

Governance bodies - such as a Generative AI Center of Excellence - are needed in enterprises to help shape the secure adoption of Gen AI. They help accelerate low-risk, high-impact business experiments while enforcing appropriate oversight of high-risk plans. By developing repeatable, enforceable, and disseminated guidelines, enterprises can leverage Gen AI solutions more quickly and securely.

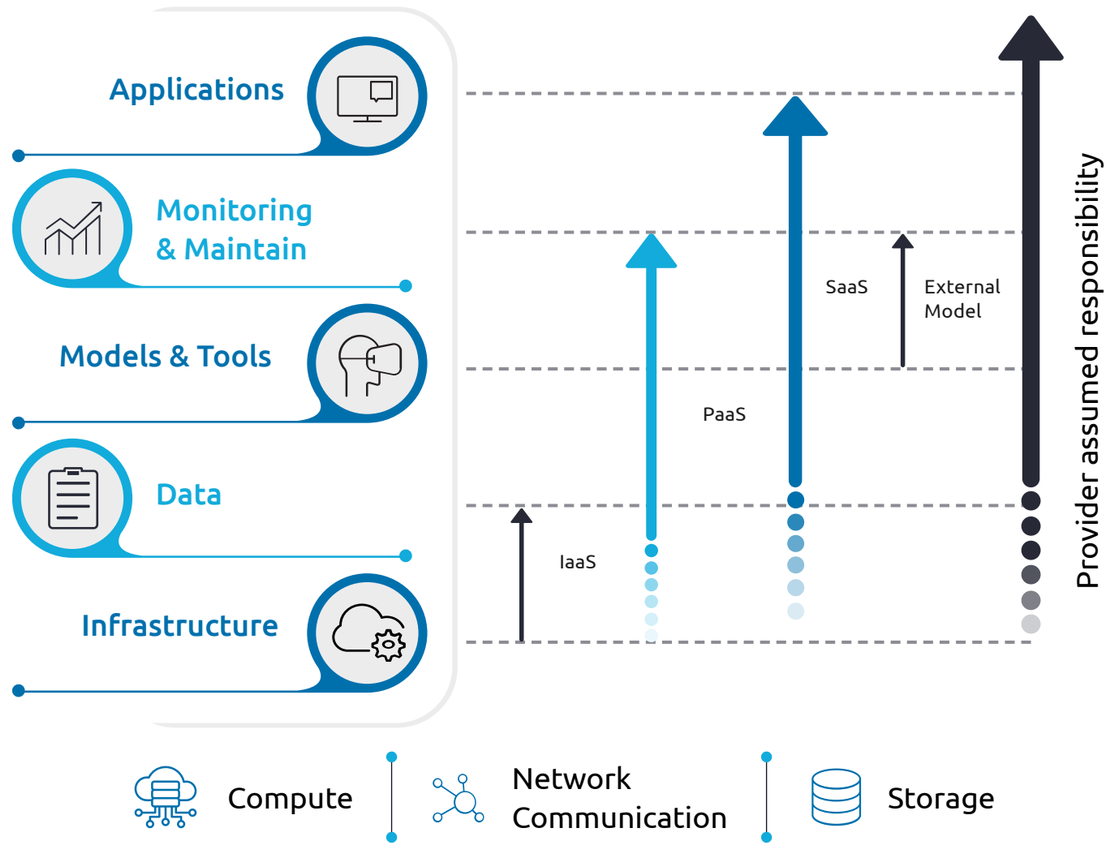


Figure 2: Shared responsibility models for various cloud provider Gen AI delivery models

Securing Data

Gen AI lacks human filters when it produces data: the machine searches through everything it can access and then reproduces this knowledge with complete candor regardless of sensitivity. It is, therefore, imperative to set limits. To do this, enterprises must inventory their data: classify it; implement controls for quality, representativeness, integrity, and access; and create repositories of authorized data for Gen AI applications.

Gen AI's consumption of data makes data classification even more essential to adequately protect an enterprise and customers. Classification allows tighter control of data used to train, specialize, and refine models. Access to its output can be restricted and data leak protection tools can be implemented; or a response can be limited using a subset of data based on a right-to-know rule.

With a third-party LLM, there is limited ability to build "native" guardrails around inputs and outputs. Likewise, the ability to implement guardrails inside the learning phases of a Generative Adversarial Network[1] is limited when using closed models in an application. It is critical to consider whether data can be inspected and validated, and whether its inputs and outputs can be observed when choosing components of a system.

A model's output must be subject to verification to detect hallucinations, malicious reinforcement, or drifts from expected behavior over time. When using real-time model output, such as with a chatbot, the observability of past performance to preempt unacceptable responses is important. A key point is to understand the data lifecycle and its sensitivity, as captured in Figure 3. Data security requirements can change over its lifecycle, depending on its proximity to, or comingling with other data.

Successfully securing Gen AI solutions is a multi-discipline approach that requires partnerships between cybersecurity, data governance, data science, and legal and compliance, since disciplined data management is at the heart of achieving Gen AI data security.

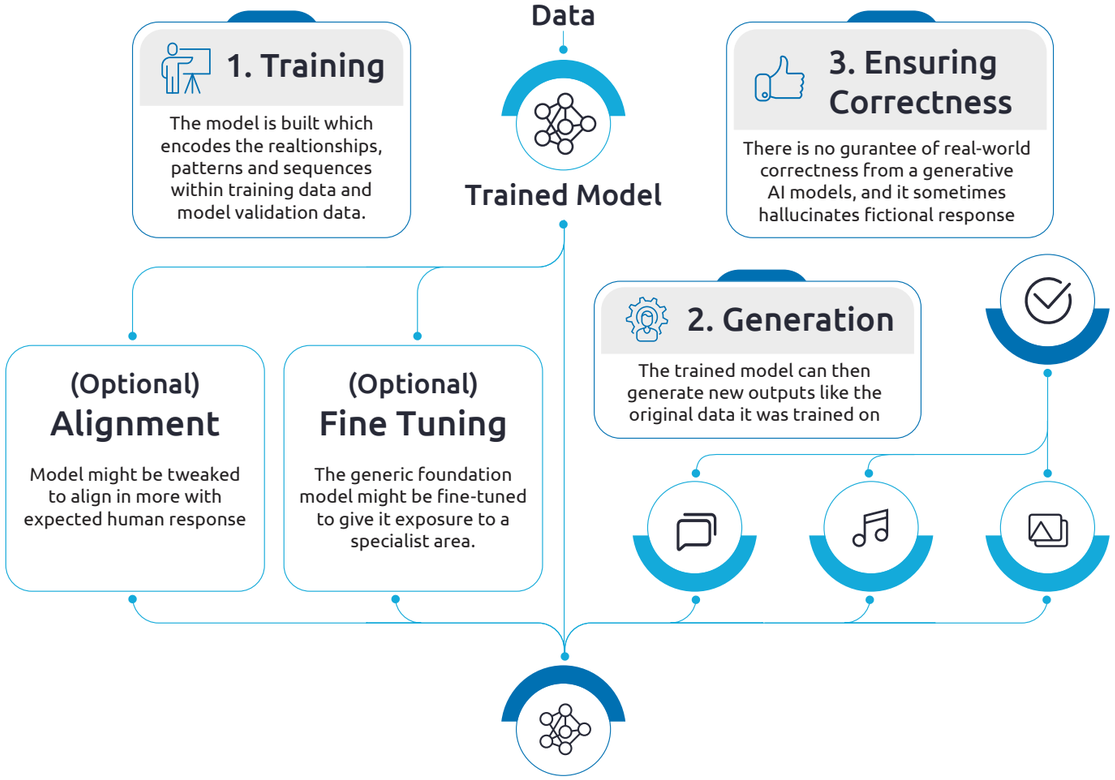


Figure 3: Data lifecycle inside a generative AI application

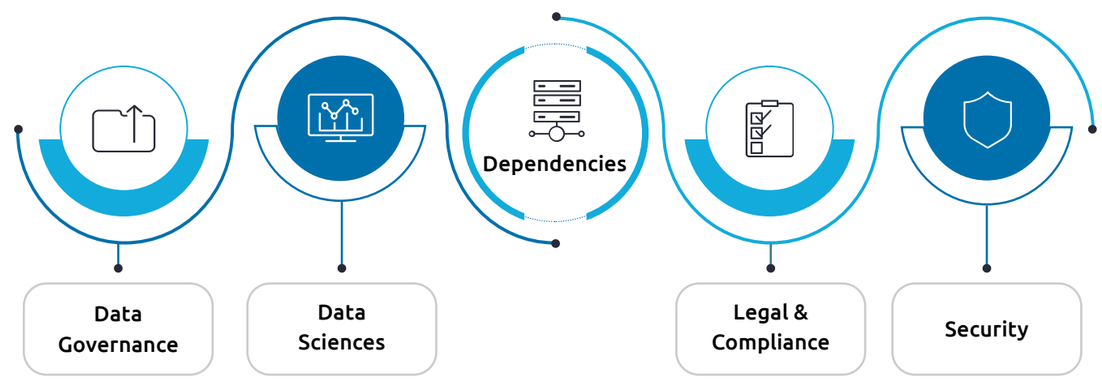


Figure 4: Multi-discipline interactions necessary for Gen AI success

Trusted Gen AI models and their outcomes

It may not be possible to gain access to and then validate all data sets used during the lifecycle of a Gen AI solution. A model such as the commonly used Large Language Model (LLM), multi-model models, and transformer-based models generating outcomes through user prompt or API requests can fall into one of the following model categories:

- Developed and initially trained by an external party (OpenAI's ChatGPT, for instance) and used "as is" by the enterprise
- Developed and initially trained by an external party, then specialized by the enterprise to a specific domain (i.e., specialism) with a new data set to address specific use cases
- Developed and trained by the enterprise entirely

Supply chain security and third/fourth-party risk management are crucial for the first two categories. It is even more important to integrate security controls such as model auditability, data leakage prevention, hallucination and bias detection (i.e. guardrails) into the application development pipeline.

Data quality

The recurrent use and provenance of training data is a focal point when using externally sourced models. Its composition, how often it changes, and how recursion between customer prompt/response pairings and reinforcement training of the model occurs should be clear.

When developing and training a proprietary model (third category above), some risks can be amplified while others are mitigated. The need to understand data's provenance and classification of training data while also testing for bias and derogatory responses falls on the enterprise, even though those can be different disciplines. At the same time, the risks of recursive training from prompt/response pairs are reduced as the information doesn't leave the local model.

For all models, organizations must apply their own additional, adaptable controls, such as:

- Specific security monitoring rules
- Completely original measures, such as controls to detect specific new attacks or user behaviors.
- For multi and hybrid architectures, API security and CI/CD secure-by-design domains

The key to assurance of data's integrity is due diligence on a provider's security, privacy controls, and compliance. Their commitments and responsibilities should be clearly defined in any contract.

Application and infrastructure monitoring and delivery

The final aspect of security for Gen AI is protecting applications from being rendered inoperative or unavailable. This requires deploying security controls within applications and infrastructure, covering compute, endpoint, network, and storage.

The same security and compliance hygiene applied to “classic security” must be applied here, especially those handling sensitive data. Corporate security policies and mandatory security controls over these layers are as important as ever.

Gen AI applications will require some new security controls, such as prompt analysis, and adaptation to existing security controls, such as edge protection, to be effective. Building adequate, automated governance around data classification and usage should be part of any security roadmap.

Software supply chain management is more important in generative AI application development, e.g., for pinning dependency versions in model development to ensure training runs do not become corrupted. This is important for monitoring and delivery since it is a part of the software delivery lifecycle. Continuous Integration (CI) and continuous delivery (CD) through a DevSecOps pipeline for application development can be used to secure model development. Red teaming[2], an application to test for vulnerabilities, should include testing of any prompts. This aims to stop malicious users from:

- Corrupting or recovering training data
- Manipulating results for other users
- Performing denial of service attacks
- Exfiltrating data

As generative AI evolves, security functions native to Gen AI will too, as will their capabilities to integrate with external security solutions.

Capgemini and generative AI security

Security does not have to be a barrier to tapping into generative AI’s power and, when done well, it can accelerate value creation. We can safely and sustainably deliver the innovation that generative AI offers based on our experience with Gen AI and adjacent expertise. We have the foundational, technical, and organizational means to help you secure your generative AI’s opportunities and deliver on transformations necessary for secure Gen AI use. Our experts and partners deliver advice, systems, tools, and operations to help you achieve the promise of Gen AI quickly and securely.

About Capgemini

Capgemini is a global leader in partnering with companies to transform and manage their business by harnessing the power of technology. The Group is guided every day by its purpose of unleashing human energy through technology for an inclusive and sustainable future. It is a responsible and diverse organization of nearly 350,000 team members in more than 50 countries. With its strong 55-year heritage and deep industry expertise, Capgemini is trusted by its clients to address the entire breadth of their business needs, from strategy and design to operations, fueled by the fast evolving and innovative world of cloud, data, AI, connectivity, software, digital engineering, and platforms. The Group reported in 2022 global revenues of €22 billion.

Get the future you want | www.capgemini.com

CIS_30.09.2025_Soumi Biswas

For more details contact:

cybersecurity.in@capgemini.com