Answering your 4 biggest questions about generative AI security

Rapidly adopt generative AI—while helping to ensure security, privacy, and compliance

This eBook is written for business leaders, particularly IT decision makers and security team leads who are planning or thinking about how to safely integrate generative AI into their organizations.

Table of contents

Ready, set, generate: Adopt generative AI quickly and safely

The race for generative AI is on. Businesses are rushing to reinvent customer experiences and applications, driven by potentially massive improvements to productivity and experience.

While the generative artificial intelligence (AI) era has only just begun, organizations are already realizing tangible benefits across virtually all business units. However, security professionals advise caution. They cite data privacy, model bias, harmful content creation (such as deepfakes), and the risks of malicious input on models as reasons to approach generative AI adoption with care.

It is imperative that organizations approach generative AI with a clear strategy for how to protect their data, users, and reputation—while still enabling rapid adoption and improving customer experience.

While this represents a multifaceted challenge, organizations should remember that standard best practices for AI, machine learning (ML), data protection, and cloud workload security still apply. In fact, your organization may be better prepared to secure generative AI than you think.

Establishing proper protections for generative AI workloads now will help drive innovation across your organization—giving your teams the confidence to pursue big ideas and the freedom to focus on growing your business.

In this eBook, you will explore 4 key questions to ask as you begin your journey toward more secure generative AI workloads.

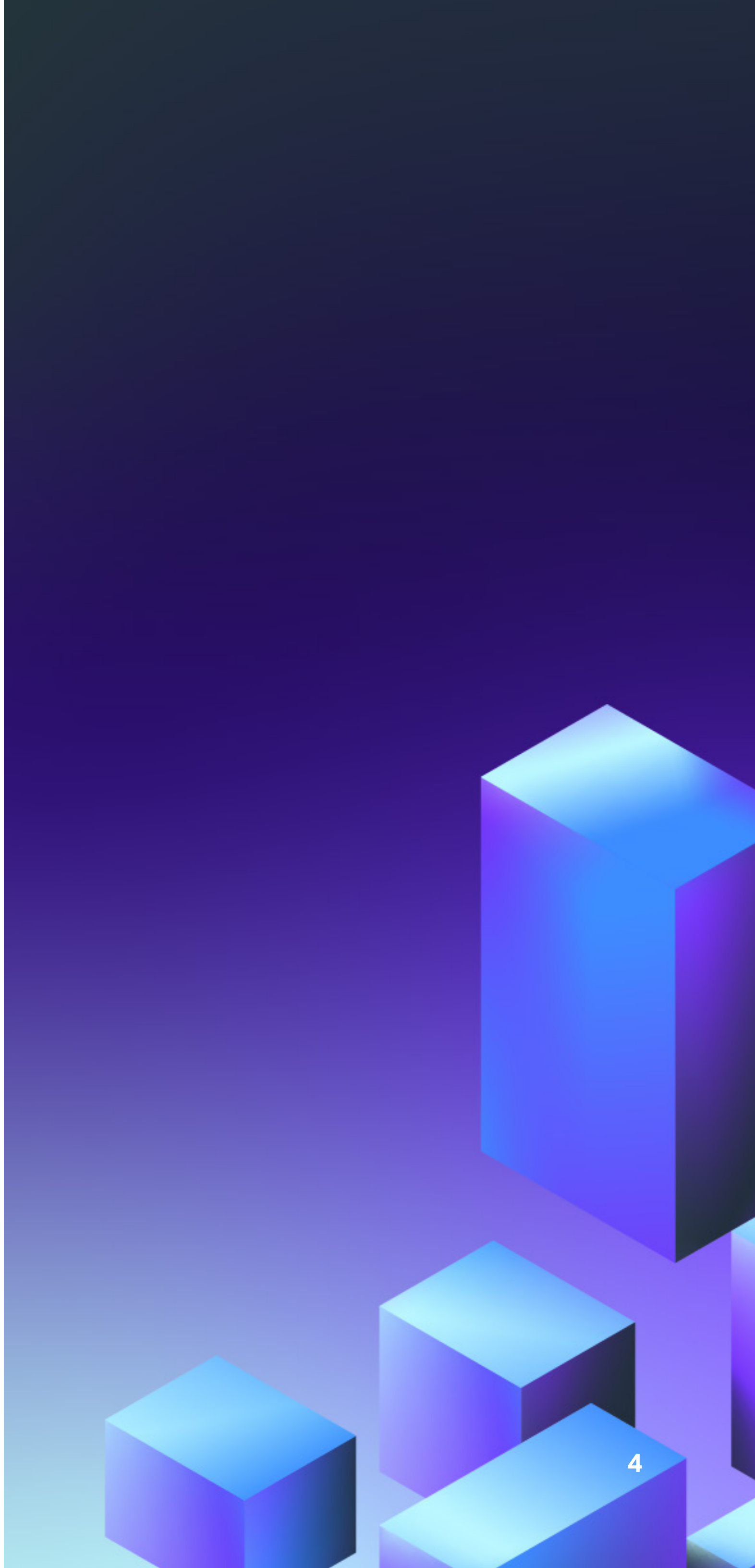
1. **What do you need to protect?**
2. **How can you address compliance concerns?**
3. **How can you ensure the models perform as intended?**
4. **Where should you start?**

Question 1:

What do you need to protect?

Before you can safely develop and deploy generative AI applications, it's important to understand what exactly needs protecting. It may be useful to group these efforts into three categories:

- **Protecting your cloud workloads**
- **Protecting your data**
- **Protecting your generative AI applications**



Protecting your cloud workloads

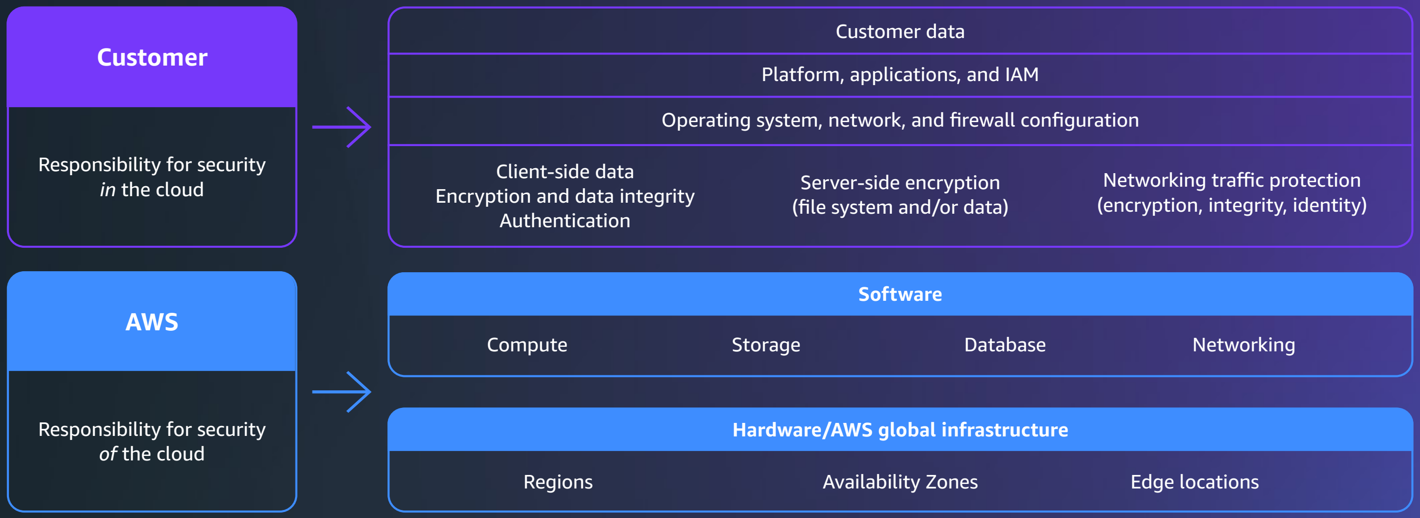
Using generative AI while meeting your security and privacy goals starts with protecting your overall cloud infrastructure, services, and configurations. To do that, you will first need to distinguish your security responsibilities from those handled by your cloud provider.

Amazon Web Services (AWS) customers can refer to the **shared responsibility model** for guidance in this area. It explains that broadly speaking, AWS is responsible for operating, managing, and controlling the infrastructure that runs all the services offered in the AWS Cloud—referred to as "security *of* the cloud."

AWS customers, on the other hand, are responsible for managing the guest operating system (including updates and security patches) and other associated application software, as well as the configuration of the AWS-provided security group firewall. The scope and specific duties customers are required to perform depend on the AWS services they opt to use. This is called "security *in* the cloud."

While the popularity of generative AI may be new, traditional security best practices remain a helpful starting point. This includes basic security hygiene practices for:

- Identity and access management (IAM)
- Detection and response
- Infrastructure protection
- Data protection
- Application security



Customer

Responsibility for security *in* the cloud

Customer data

Platform, applications, and IAM

Operating system, network, and firewall configuration

Client-side data Encryption and data integrity Authentication

Server-side encryption (file system and/or data)

Networking traffic protection (encryption, integrity, identity)

AWS

Responsibility for security *of* the cloud

Software

Compute | Storage | Database | Networking

Hardware/AWS global infrastructure

Regions | Availability Zones | Edge locations

Protecting your data

Next, you will need to help ensure the security and privacy of the data used by your generative AI applications. This may include proprietary information, valuable intellectual property (IP), and personal identifiable information (PII).

Generative AI applications are powered by foundation models (FMs), which are trained on vast quantities of data. FMs analyze this data to identify patterns and learn how to generate new, similar content. To build generative AI applications that meet your specific business requirements, you will typically need to customize an existing FM by training it on your organization's data.

To help protect this data, you will need to consider data privacy controls and IAM policy best practices.

When customizing an FM, ensure your teams are working with a version of the model that is securely stored and not being used to improve the FM itself. By setting up a single-tenant dedicated capacity in **Amazon Bedrock**, the service can attach its inference instances to your **Amazon Virtual Private Cloud** (Amazon VPC) in order to read from and write to **Amazon Simple Storage Service** (Amazon S3).

Effective IAM helps validate that the right people and machines have access to the right resources under the right conditions. The **AWS Well-Architected Framework** describes design principles and architectural best practices to help manage identities. This resource is a useful tool for developing IAM policies—and addressing other security concerns, such as threat detection and network security.

Protecting your generative AI applications

To secure generative AI at the application level, you must continuously identify, classify, remediate, and mitigate risks. A first step is implementing existing best practices for keeping environments and data safe.

From there, you should consider how you can shift security to earlier in the development process. This can streamline your efforts and allow development teams to innovate faster and with greater freedom while avoiding making security a bottleneck.

Next, you should consider how to protect the three critical components of any AI application: inputs, outputs, and the model itself.

Protecting inputs

Start by reviewing the data that enters into your AI system. Users should not have direct access to the FM without input filtering to reduce the risk of integrity attacks like tampering, spoofing, or prompt injection. These attack techniques circumvent controls or abuse the model. Other strategies to consider for protecting inputs are data quality automation, continuous monitoring, and threat modeling.

Protecting outputs

Risks to generative AI application outputs include information disclosure, IP incidents, and misuse or abuse of the model that can damage your organizational reputation. When developing your threat model, consider the information footprint and usage context and include complex behavior detection and monitoring.

Protecting the model itself

Finally, consider how adversaries may attempt to remove data from the model itself or its associated components. Risks include misrepresentations of the real world or data in the model and damage to the model's integrity or availability. Model threats to your business objectives and implement monitoring for these threat scenarios.

Question 2:

How can you address compliance concerns?

By mitigating the risks of designing and developing generative AI applications, your organization can build trust with your partners and customers, maintain your brand reputation, and continue to address your compliance requirements.

Legislative regulation of generative AI applications is still in its early stages, and there is no consensus yet on best practices. As such, navigating the maze of conflicting standards and oversight across different jurisdictions presents a complex and ongoing challenge.

Engage with your legal advisors and privacy experts to assess the requirements and implications of building your generative AI application. This may involve vetting your legal rights to use specific data and models and determining the applicability of laws around privacy, biometrics, antidiscrimination, and other use case-specific regulations.

Be mindful of differing legal requirements across states, provinces, and countries and new AI regulations that are being proposed around the world. Revisit these considerations at future deployment and operational phases.

Collaborating with peers, AI experts, and government organizations can also help you maintain compliance while showing customers that you take legal and ethical AI standards seriously. Recently, Amazon joined the White House and six leading AI companies in making **voluntary commitments to responsible and secure AI development**—demonstrating the value of such engagements while laying the groundwork for future collaboration.

Risks inherent to artificial intelligence

As with any solution that uses ML, generative AI applications present risks beyond those of traditional software. To safely build and deploy applications with generative AI, you will need to address and develop strategies for mitigating these risks, which include:

- Outputs that are biased, untrue, misleading, harmful, or offensive
- Complexities and costs at scale
- Datasets that grow too large, stale, or detached from their intended context
- Concerns over increased opacity and reproducibility
- Underdeveloped testing standards and procedures

In the next section, we will cover broad strategies for reducing some of these risks—and best practices for defining the professional, organizational, and societal impacts of your generative AI applications.

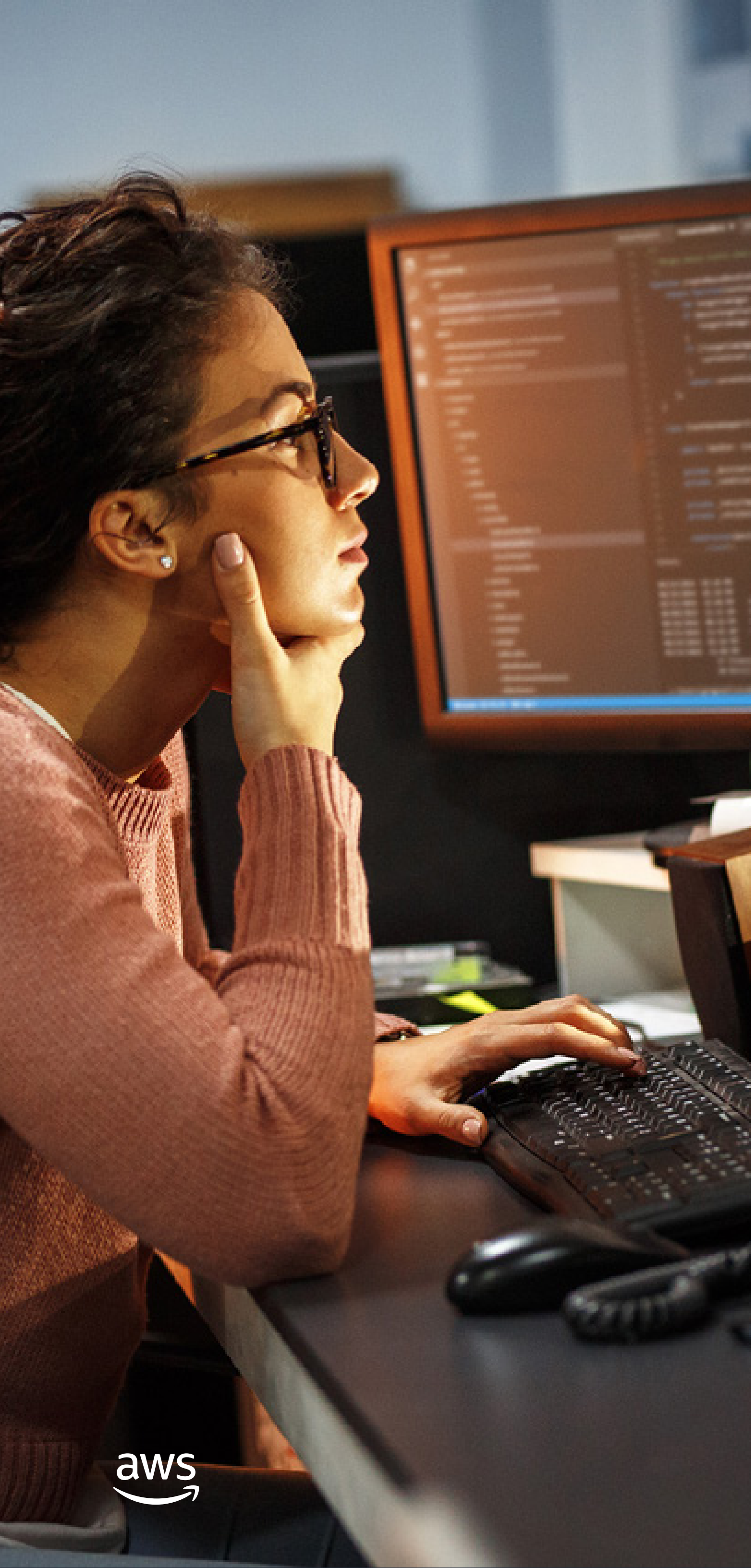
Question 3:

How can you ensure the models perform as intended?

Ensuring the responsible use of generative AI has emerged as an essential business task—and a critical enabler of continuous innovation.

FMs train on massive datasets, performing complex analyses that help them understand how to generate similar content. While many FMs deliver remarkable results, the age-old dictum of “garbage in, garbage out,” or GIGO, still applies. If an FM is fed inaccurate, incomplete, or biased data, its outputs may display similar flaws.

Flawed data opens opportunities for misuse, malicious actions, and other risks. As your generative AI application expands in users, scope, and function, the potential impact of these issues grows.



Fostering responsible AI

Committing to a responsible AI strategy will help you address these risks. Dimensions of responsible AI include explainability, fairness, governance, privacy, security, robustness, and transparency. It also includes understanding the ways in which different cultures and demographics are viewed, treated, and impacted by the application.

It’s best to begin considering responsible AI early in your generative AI journey—then continue to address it throughout the application lifecycle as a key part of your vision. Start with relatively small and simple actions. Next, scale how responsible AI affects your design, development, and operations over time.

When drafting responsible AI and governance policies, consider how your generative AI application will affect your users, customers, employees, and society. Be sure to address algorithmic fairness, diverse and inclusive representation, and bias detection.

Tackling toxicity

Toxicity in large language models (LLMs) refers to the generation of rude, disrespectful, or unreasonable text. There are many strategies to help prevent toxicity and ensure fairness in your generative AI applications. For example, you might identify and remove offensive language or biased phrases from your training data. You can also conduct more narrow fairness tests focused on your application's specific use case, target audiences, or the prompts and queries it is most likely to receive.

You can also train guardrail models on annotated datasets that identify varying types and degrees of toxicity. This can help the FM learn to detect and filter unwanted content across training data, input prompts, and generated outputs in an automated way.

Protecting privacy

There are several steps you can take to help prevent the unwanted exposure of sensitive information, trade secrets, and IP when working with generative AI applications.

Model deletion is one method to help address privacy concerns. This involves eliminating improperly used data as soon as it has been identified, helping to remove the effects of that data on any component of the FM.

Another approach is sharding, where training data is divided into smaller portions on which separate sub-models are trained—with the sub-models eventually combining to form the overall FM. This practice can make it much simpler to remediate FMs that have or are at risk of exposing private information. Rather than retraining the entire model, you only need to remove the unwanted or improperly used data from its shard and then retrain that sub-model.

Filtering and blocking can also be effective approaches. These methods explicitly compare protected information to generated content before the user sees it. If the two are too similar, the content is suppressed or replaced to avoid exposure. Limiting the number of times any specific piece of content appears in the training data can also prove helpful.

Enhancing explainability and auditability

To further support responsible AI, consider the need to explain the methodology and key factors that influence your application's output. Auditability is another important component of responsible AI. Implement mechanisms that allow you to track and review the development and operation of your generative AI application. This will help you trace the root cases of any problems and help meet governance requirements.

Consider documenting relevant design decisions and inputs throughout the development lifecycle. Establishing a traceable record can help internal or external teams evaluate the development and functioning of your generative AI application.

Staying responsible

Finally, think about how you will help ensure continued adherence to your responsible AI policies. Be sure to apply the lessons you learn and the experience you gain to evolve your security and privacy practices. Regularly educate all employees in your organization on their obligations to safe and secure generative AI practices. Foster a culture of responsible AI, use the right tools to help you monitor model performance and inform risk, and allow your teams to inspect the model and its components when needed. Test, test, and—when in doubt—test again.

Question 4:

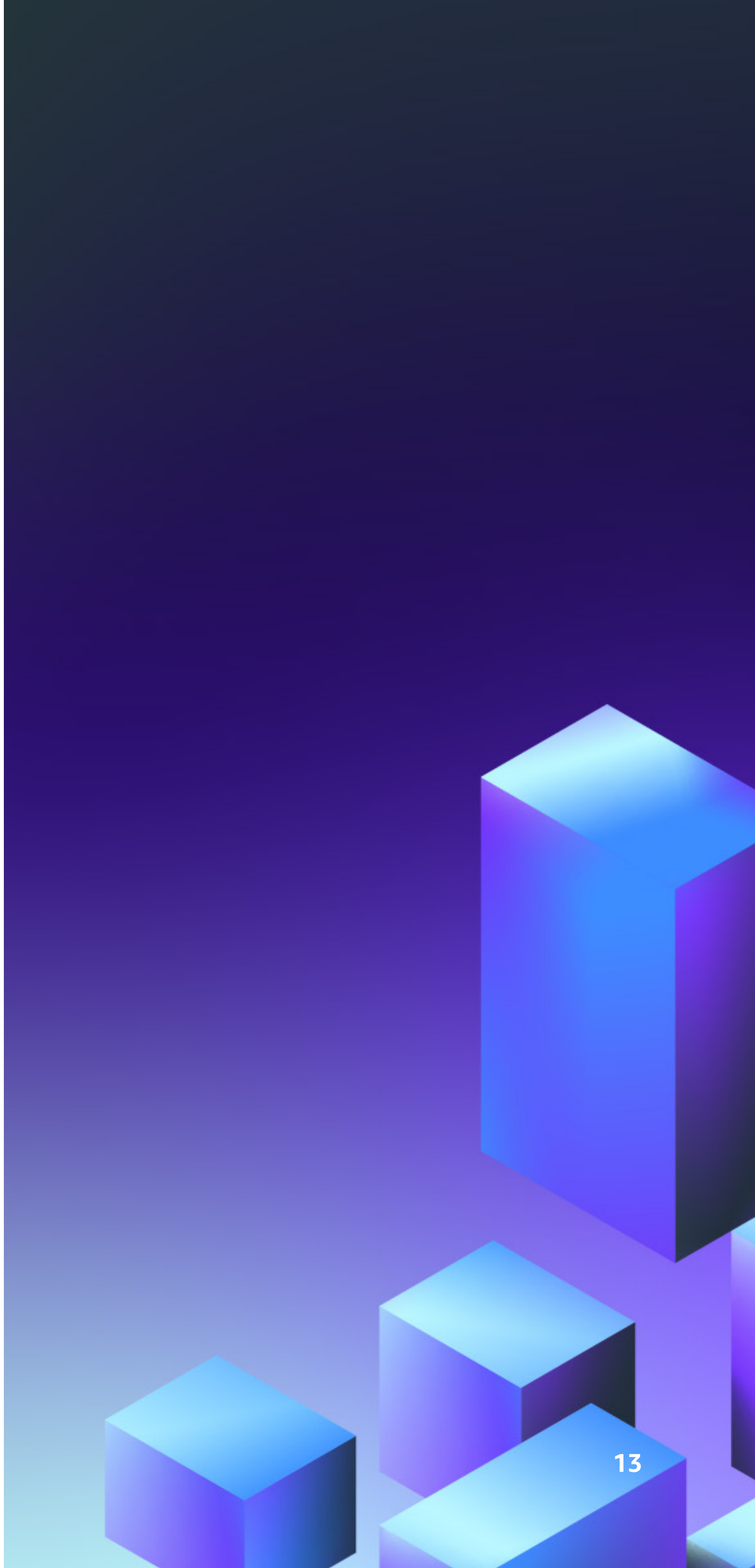
Where should you start?

Securing generative AI applications is no simple feat, and there's no universal set of actions you can take to accomplish it. When you work with the right vendor and deploy the right tools, however, the path to success becomes much clearer.

Using **Amazon Bedrock**, for example, can drastically simplify and accelerate your journey to developing secure generative AI applications. Amazon Bedrock is a fully managed service that makes FMs from Amazon and leading AI startups available through an API.

When you customize a model with Amazon Bedrock, the service can fine-tune the model for a particular task without your team having to annotate large volumes of data. Then, Amazon Bedrock makes a separate copy of the base FM that is accessible only to you and trains this private copy of the model. None of your data is used to train the original base models, which helps keep your proprietary data private and secure.

You can also configure your **Amazon VPC** settings to access Amazon Bedrock APIs and provide your model with fine-tuning data in a secure manner. Your data is always encrypted in transit and at rest through service-managed keys. Plus, with **AWS PrivateLink**, you can pass your AWS Cloud data to Amazon Bedrock exclusively through the AWS network—never via the public internet.



Improving privacy with AWS

Whether you build generative AI applications using Amazon Bedrock, another service (such as **Amazon SageMaker**), or your own tooling, when you run and manage your applications on AWS, you gain industry-leading privacy protections and controls.

AWS supports 143 security standards and compliance certifications, helping to satisfy the requirements of our customers worldwide. All your data can be encrypted at rest using your own **AWS Key Management Service** (Amazon KMS) keys, providing full control and visibility into how your data and FMs are stored and accessed.

CONCLUSION

Next steps

AWS is committed to helping you build generative AI applications that grow your business while helping you meet your security, privacy, and compliance goals.

We stand firm in our belief that generative AI applications can be securely designed, developed, and operated. We also acknowledge the validity of security and privacy concerns about these technologies. **Generative AI raises new challenges** in defining, measuring, and mitigating issues around data privacy, IP, legislative oversight, equality, and transparency.

With the introduction of new products, the growing complexity and scale of solutions, new training parameters, and ever-growing datasets, generative AI security will become even more essential in the days ahead. By developing an effective and comprehensive security strategy for generative AI workloads now, you can maximize your competitive advantage—and stand prepared for the rapidly approaching future.

The good news: The basic controls needed to securely design, develop, and run generative AI applications have been in place for years—and are aligned with trusted, proven principles of cloud security, such as those found in the **AWS Well-Architected Framework**.

By exploring the practices outlined in this eBook, you've already taken your first step toward securing your generative AI workloads.

Now, take the next step with AWS. We can provide you with deep insights and specific guidance needed to stay up to speed on emerging topics, think through your unique challenges, and unlock the full benefits of generative AI—all while protecting your data, your customers, and your business.

Learn more about generative AI on AWS ›

Get started quickly with Amazon Bedrock ›

Build and customize FMs on Amazon SageMaker ›

Elevate your security in the cloud with AWS ›

Transform responsible AI from theory into practice ›