WHITEPAPER

6 key guidelines for building secure and reliable generative AI applications on Amazon Bedrock

Follow these guidelines to build high-performing, secure, and responsible generative AI applications

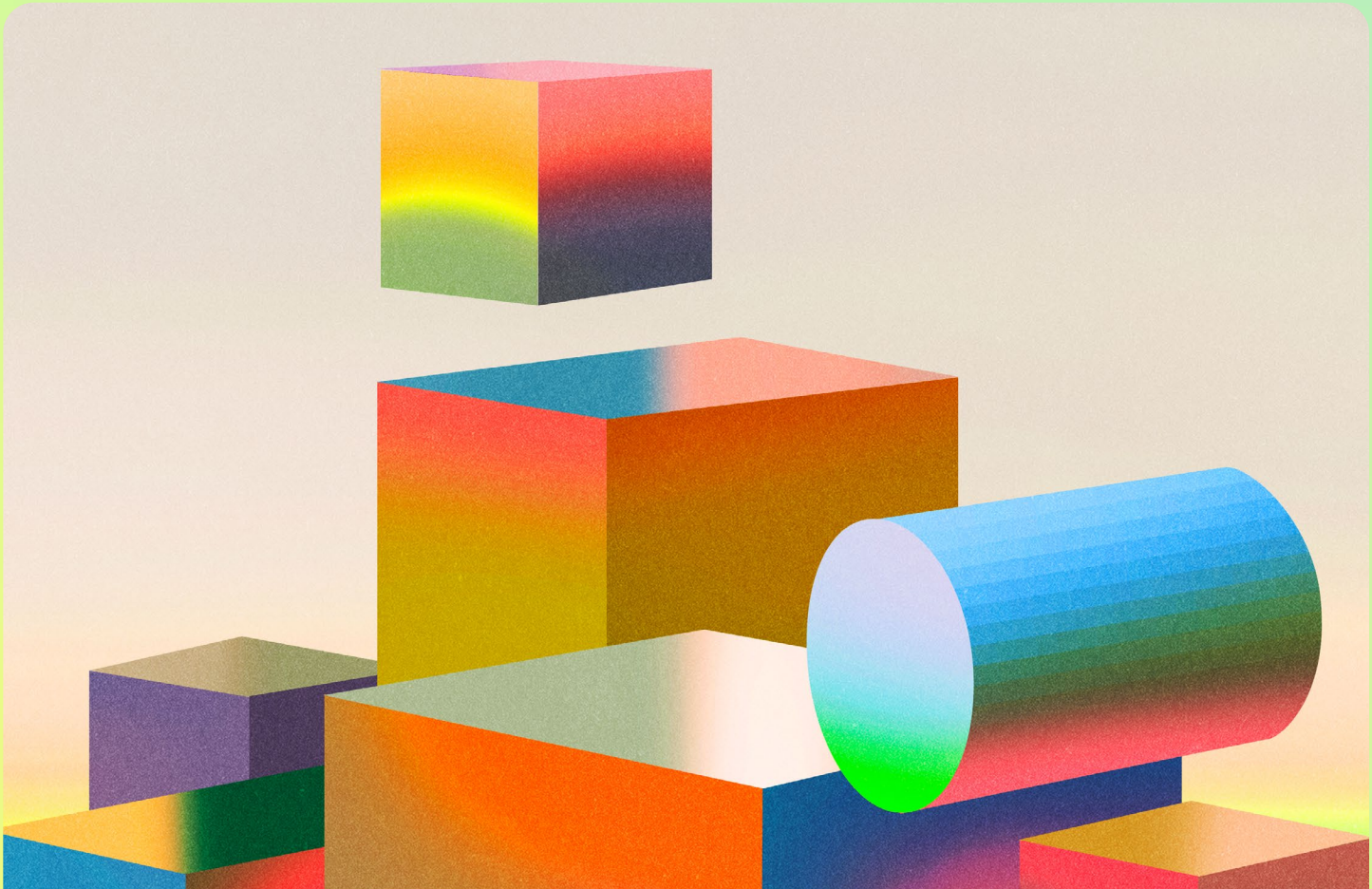


Table of contents

Overview

Foundation models are trained on extensive datasets to understand and generate human-like responses

Foundation models (FMs) have advanced significantly in recent years, resulting in widespread adoption in a variety of industries, including customer service, content generation, and healthcare with notable improvements in natural language understanding (NLU) and natural language generation (NLG).

Developing and deploying generative AI (gen AI) applications into production is a complex process that demands careful planning and implementation. As these models get more advanced, their integration into real-world applications brings both opportunities and challenges. Key considerations include selecting the best-suited FM for your use case, ensuring reliable performance through rigorous evaluation, getting access to powerful tools and capabilities that allow you to build your apps with this model, mitigating risks, such as hallucinations, and managing model responses effectively.

Amazon Bedrock is a fully managed service that offers a choice of high-performing FMs from leading AI companies through a single API, along with a broad set of capabilities needed to build gen AI applications with security, privacy, and responsible AI. By removing the need to manage infrastructure, Amazon Bedrock allows teams to focus on developing powerful applications without worrying about scalability or system complexity. With seamless scalability and flexibility, Amazon Bedrock can easily handle varying workloads, enabling organizations to build secure, reliable gen AI solutions that meet their business needs.

This guide outlines the key challenges in developing gen AI applications and shows how Amazon Bedrock addresses these challenges to boost productivity, efficiency, and innovation. Additionally, you will gain insights from real-world examples into how leading businesses use Amazon Bedrock to build gen AI applications securely.

1 Choose the right model for your use case

When developing generative AI applications for production, it's crucial to recognize that no single model fits all needs

The choice of FM significantly impacts the application's performance, scalability, and suitability for specific tasks. Different FMs excel in various areas, with capabilities varying widely based on factors such as model size, training data, cost, and underlying architecture. For instance, some FMs may be better suited for tasks requiring a deep understanding of context and nuance, while others may be better suited for image processing and generation.

Experimenting with multiple FMs and performing thorough model evaluations using representative data and test cases helps ensure an application remains effective and competitive. This approach allows for informed decisions based on empirical evidence rather than theoretical capabilities or marketing claims. As the field evolves rapidly, staying up to date with the latest developments and periodically reevaluating your choice is essential.

Amazon Bedrock helps you rapidly adapt and take advantage of the latest gen AI innovations with easy access to a choice of high-performing FMs from leading AI companies like AI21 Labs, Anthropic, Cohere, Meta, Mistral AI, Stability AI, and Amazon. In addition. Amazon Bedrock Marketplace, lets you discover, test, and use over 100 popular, emerging and specialized FMs on fully managed endpoints. The single-API access of Amazon Bedrock, regardless of the models you choose, gives you the flexibility to use different FMs and upgrade to the latest model versions with minimal code changes.

Amazon Bedrock provides evaluation tools to help customers accelerate the adoption of gen AI applications. They can evaluate, compare, and select the FM for their use case with model evaluation. Customers can now use an LLM as a Judge to evaluate model outputs using custom prompt datasets with metrics such as correctness, completeness, and harmfulness. Additionally, customers can prepare their retrieval-augmented generation (RAG) applications built on Amazon Bedrock Knowledge Bases for production by evaluating the retrieve or retrieve and generate functions.

Hitting the road with generative AI

Ferrari, a luxury Italian auto manufacturer, is using the broad model selection of Amazon Bedrock to apply gen AI to several use cases, from accelerating the vehicle design process to providing personalized services to its customers. Using FMs in Amazon Bedrock, Ferrari developed a car configurator to make it easier and faster for customers to personalize their car, which increased sales leads and reduced vehicle configuration times by 20 percent. Ferrari also enhanced the after-sales experience with a gen AI chatbot. To assist its sales professionals and technicians, the company fine-tuned FMs in Amazon Bedrock—including Amazon Titan, Claude 3, and Llama—on its documentation.

> "Amazon Bedrock has simplified our approach. We can connect to a single layer of APIs to quickly test, benchmark, and deploy different models."
>
> Mauro Coletto, Head of Business Analytics & AI, Ferrari

Using FMs in Amazon Bedrock, Ferrari increased sales leads and reduced vehicle configuration times by 20%

2 Build models with your data and Custom Model Import

While pretrained FMs have achieved remarkable performance across a wide range of natural language tasks, they are often trained on broad, general-purpose datasets. As a result, these models may not perform optimally when applied to specific domains or use cases that deviate significantly from their training data. This is where the need for customizing models arises.

Customizing pretrained models involves fine-tuning them on domain-specific data, allowing the models to adapt and specialize for the unique characteristics, terminology, and nuances of a particular industry, organization, or application. By using customized models, businesses can unlock several key benefits.

Amazon Bedrock allows organizations to customize FMs with their own proprietary data to build applications tailored to specific domains, organizations, and use cases. This process, known as data gravity, enables customers to create unique user experiences that reflect their company's style, voice, and services.

There are two main methods for model customization in Amazon Bedrock:

1. Fine-tuning involves providing a labeled training dataset to specialize the model for specific tasks. By learning from annotated examples, the model's parameters are adjusted to associate the right outputs with corresponding inputs, improving its performance on the tasks represented in the training data.
2. Continued pretraining, on the other hand, utilizes unlabeled data to expose the model to certain input types and domains. By training on raw data from industry or business documents, the model accumulates robust knowledge and adaptability beyond its original training, becoming more domain-specific and attuned to that domain's terminology.

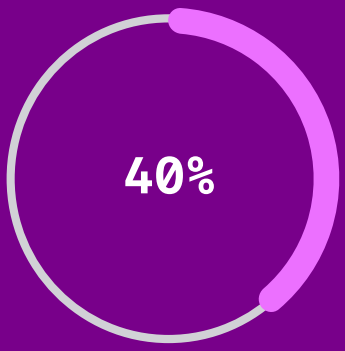
In addition to fine-tuning and continued pretraining, Amazon Bedrock now offers Custom Model Import, allowing customers to use prior model customization investments within the fully managed environment of Amazon Bedrock. With this new feature, organizations can import models customized outside of Amazon Bedrock, such as those fine-tuned or adapted using Amazon SageMaker AI or other third-party tools, and access them on demand through the invoke model API found in Amazon Bedrock.

Simplifying search with foundation models

Perplexity is revolutionizing online search with its AI-powered search companion that provides accurate and comprehensive answers to queries from more than 10 million monthly users. A key differentiator is giving users the ability to choose from multiple high-performing FMs, including cutting-edge offerings accessed through Amazon Bedrock. Amazon Bedrock has allowed Perplexity to simplify access to proprietary models while also fine-tuning open-source FMs on its AWS-powered infrastructure. By building on AWS, Perplexity has been able to continuously innovate and improve its language models and infrastructure. This includes fine-tuning models on Amazon SageMaker AI, which reduced training times by 40 percent when using Amazon SageMaker HyperPod. As Perplexity continues driving the future of intelligent search, AWS services like Amazon Bedrock allow them to stay on the cutting edge of gen AI while benefiting from the built-in security and responsible AI capabilities of Amazon Bedrock.

> "Using a high-performing service such as Amazon Bedrock means we are tapping into…powerful models in a way that allows our team to effectively maintain the reliability and latency of our product."
>
> William Zhang, Technical Team Member, Perplexity

perplexity



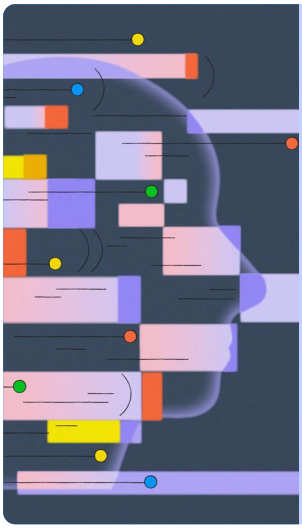
Amazon SageMaker HyperPod reduced training times by 40%

3 Ground foundation models with retrieval systems to improve accuracy

A key challenge with FMs is their tendency to generate hallucinations–outputs that may be incorrect, fabricated, or nonsensical–especially in response to open-ended queries. These hallucinations arise because FMs rely solely on their training data, which may be incomplete or biased, and do not inherently distinguish between plausible and factual information.

To mitigate this, grounding can be employed. Grounding involves integrating the FM with a retrieval system that searches external databases or document collections to find relevant, factual information during the model's inference process. The retrieved data is fed back into the model as additional input, ensuring that its responses are guided by real-world, verified information.

This technique, also known as RAG, allows FMs to produce outputs consistent with the external grounding data, improving factual accuracy and reducing hallucinations. By relying on current, trusted data sources, grounded models can condition their responses on facts rather than purely on the patterns learned during pretraining. This approach is particularly effective in scenarios where the model would otherwise lack sufficient context, significantly enhancing reliability for tasks requiring high accuracy.



Amazon Bedrock Knowledge Bases is a fully managed RAG workflow that enables customers to create highly accurate and secure custom generative AI apps using contextual information from their data sources. It supports various data sources, including S3, and Confluence, Salesforce, and SharePoint (in preview). Knowledge Bases converts unstructured data into embeddings, stores them in vector databases, and enables retrieval from diverse data stores. It integrates with Kendra for managed retrieval, supports structured data retrieval using NL2SQL, and supports graphRAG for more accurate and comprehensive responses.

Taking the next step in curated travel recommendations

Lonely Planet is using gen AI to reimagine the travel experience for customers and deliver curated travel itineraries. Using Amazon Bedrock, it took the company only seven months to translate manually published travel content from 150 million guidebooks, 270,000 mappable destinations, and 750 local experts into an evolving technology platform. Using the broad model choice available with Amazon Bedrock, the company was able to test and select the best model for its use case and reduced gen AI workload costs by 80 percent compared to other models and services.

> "We are evolving our technology platform on AWS to support the modern traveler while maintaining our 'ungoogleable' expert recommendations at a new scale and speed."
>
> Chris Whyde, SVP, Engineering & Data Science, Lonely Planet

lonely planet

Lonely Planet reduced gen AI workload costs by 80% on Amazon Bedrock

4 Integrate external systems and data sources to build artificial intelligence agents

Connecting FMs to external systems and tools enables them to access current information, execute complex, multistep actions, and overcome the inherent limitations of relying solely on training data

Integrating FMs with external data sources, tools, and systems is critical to realizing their full potential in production. This integration provides access to up-to-date, domain-specific information, enhancing accuracy, relevance, and functionality.

Agents are advanced AI systems that use the capabilities of FMs to exhibit autonomous behavior and perform complex tasks beyond only text generation. They play a crucial role in leveraging FMs' full potential. Agents are specialized components designed to handle specific tasks by interacting with both the FM and external systems. They can orchestrate complex workflows, automate repetitive tasks, and help ensure that the FM's outputs are actionable and relevant. By using agents, organizations can build applications that not only understand and generate language but also perform real-world actions, bridging the gap between language processing and practical application.

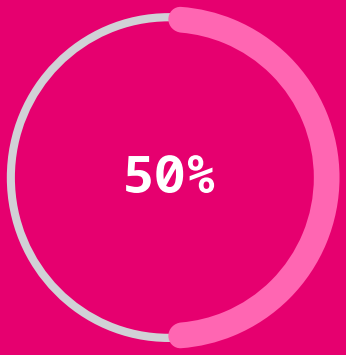
Amazon Bedrock Agents use the reasoning of FMs, APIs, and data to break down user requests, gather relevant information, and efficiently complete tasks—freeing teams to focus on high-value work. Building an agent is straightforward and fast, with setup in just a few steps. With Bedrock, customers can quickly create agents that handle sales orders, compile financial reports, analyze customer retention, and much more. However, as applications become more capable, the tasks customers want them to perform may exceed what a single agent can manage—either because the tasks are more specialized or they take longer. And coordinating potentially thousands of agents at scale is also challenging. For customers looking to advance their agents and tackle more intricate, multi-step workflows, Amazon Bedrock supports multi-agent collaboration, allowing developers to build, deploy, and manage multiple specialized agents working together seamlessly.

Amazon Bedrock's multi-agent collaboration enables developers to create networks of specialized agents that communicate and coordinate under the guidance of a supervisor agent. Each agent contributes its expertise to the larger workflow by focusing on a specific task. This approach breaks down complex processes into manageable sub-tasks processed in parallel. By facilitating seamless interaction among agents, Amazon Bedrock enhances operational efficiency and accuracy.

Delivering a superior contact center experience

Every day, DoorDash receives hundreds of thousands of requests for assistance through its contact center from consumers, merchants, and "Dashers" (independent contractors who deliver through the platform). To enhance its self-service offerings and elevate the user experience, DoorDash collaborated with AWS through the AWS Generative AI Innovation Center program to build a fully voice-operated self-service gen AI contact center solution. Using Amazon Bedrock, it got access to all the models it needed and selected Anthropic Claude. With the release of Claude 3 Haiku, DoorDash achieved the accuracy and speed it needed for its voice application, reducing response latency by 50 percent and achieving a response latency of 2.5 seconds or less. In addition, Amazon Bedrock helped the company reduce gen AI application development time by 50 percent and deliver a solution that was ready for live testing in only two months.

> "Using AWS…, we've built a solution that gives Dashers reliable and simple-to-understand access to the information they need, when they need it."
>
> Chaitanya Hari, Contact Center Product Lead, DoorDash
>
> DOORDASH



Amazon Bedrock helped DoorDash reduce gen AI application development time by 50%

5 Safeguard foundation model responses to build artificial intelligence responsibly

Prompt engineering is an effective approach to guiding the FM's generation process. Crafting specific prompts can set the tone, context, and boundaries for desired outputs, leading to the implementation of responsible AI. While prompt engineering defines the input and expected output of FMs, it might not have complete control over the responses delivered to end users. This is where guardrails come into play.

Implementing effective guardrails requires a multifaceted approach involving continuous monitoring, evaluation, and iterative improvements.

Guardrails must be tailored to each FM-based application's unique requirements and use cases, considering factors like target audience, domain, and potential risks. They contribute to ensuring that outputs are consistent with desired behaviors, adhere to ethical and legal standards, and mitigate risks or harmful content. Controlling and managing model responses through guardrails is crucial for building FM-based applications.

Within these guardrails, content filters and moderation systems are vital for detecting and filtering harmful, offensive, or biased language. These systems can be implemented at various stages of the generation process. Controlled generation techniques, such as top-k or top-p sampling, limit the model's output to the most probable or relevant tokens, improving coherence and relevance.

Amazon Bedrock Guardrails is a data governance feature that allows businesses to implement configurable safeguards and governance policies for their gen AI applications. It provides a way to customize the behavior of FMs and helps ensure they adhere to their organization's responsible AI policies. Amazon Bedrock Guardrails works by evaluating the inputs to and outputs from the FMs against the defined policies. With Amazon Bedrock Guardrails, organizations can define rules to prevent hallucinations using Automated Reasoning and contextual grounding checks, filter out harmful multimodal content, block denied words or topics, redact sensitive information like personally identifiable information (PII), and enforce content moderation based on the organization's requirements.

Driving secure mobility solutions

One of the most iconic names in the automotive industry, BMW Group has been at the forefront of using AI to enhance the driving experience for its customers. As of March 2024, its connected vehicle backend hosted on AWS processes 14.3 billion requests and 145 terabytes of traffic a day. To help over 450 DevOps teams continue to create business-critical applications that bring new experiences to over 22.3 million vehicles, BMW Group wanted to use gen AI to boost efficiency and productivity and optimize its cloud infrastructure.

The company used Amazon Bedrock to build its first-class gen AI–powered cloud assistant solution called In-Console Cloud Assistant (ICCA). The ICCA monitors infrastructure health at scale; answers questions related to cloud infrastructure; provides personalized recommendations for issue resolution, optimizing configurations and resizing resources; and even automatically generates and implements code. As a result, BMW Group was able to improve resource allocation and increase cost savings. With secure access to FMs in Amazon Bedrock—and with data encrypted in transit and at rest—BMW Group can securely deliver high-quality connected mobility solutions to motorists around the world.

> "Using Amazon Bedrock, we've been able to scale our cloud governance, reduce costs and time to market, and provide a better service for our customers. All of this is helping us deliver the secure, first-class digital experiences that people across the world expect from BMW."
>
> Dr. Jens Kohl, Head of Offboard Architecture, BMW Group

14.3B

Processed 14.3 billion requests and 145 terabytes of traffic a day

450+

Helped over 450 DevOps teams continue to create business-critical applications

6 Fortify security and safeguard privacy in foundation model-powered applications

Building FM-based applications involves unique security and privacy challenges.

These applications often handle vast amounts of data, some of which can be sensitive or proprietary. Key considerations include the risk of data breaches, which can lead to significant privacy infringements and intellectual property (IP) theft, making data protection through encryption and access controls paramount.

Another major concern is model manipulation, where adversaries might attempt to manipulate FM outputs, leading to biased or harmful results. Additionally, infrastructure vulnerabilities must be addressed to secure the hardware and networks supporting FMs, ensuring operational integrity. Ethical and legal risks are also significant, requiring FMs to comply with standards and regulations to avoid generating biased content or infringing on IP rights.

Amazon Bedrock security and compliance incorporates multiple strategies to address the security and privacy concerns inherent in gen AI-based applications. It employs industry-standard encryption protocols to protect data in transit and at rest and uses stringent access control mechanisms like role-based access control (RBAC) so that only authorized personnel can access sensitive data and functionalities.

By adhering to various compliance standards such as GDPR and HIPAA, the data handling practices of Amazon Bedrock meet regulatory requirements, while comprehensive logging and auditing capabilities allow continuous monitoring and tracking of all interactions, ensuring transparency and accountability.

Secure API integrations and privacy-preserving techniques are utilized within Amazon Bedrock to prevent data leakage during interactions with external systems and APIs. Finally, Amazon Bedrock has a robust incident response framework that includes regular security assessments and threat modeling. It also implements proactive measures like rate limiting, logging, and alerting mechanisms to prevent overreliance on FMs and enable accurate and secure model outputs.

Conclusion

Building generative AI applications demands rigorous planning and precise implementation to ensure high performance, robust security measures, and adherence to responsible AI practices

Amazon Bedrock, a fully managed AWS service, simplifies the development and scaling of gen AI applications by handling infrastructure, allowing teams to focus on innovation. Success depends on choosing the right FM, applying effective prompt engineering, and implementing guardrails to ensure safe, accurate outputs. Grounding models with real-time data through RAG further boosts accuracy and reduces hallucinations.

Amazon Bedrock integrates securely with external systems, protecting data through encryption and compliance with regulatory standards. Its comprehensive tools–from model customization to RAG–enable organizations to quickly build secure, high-performance gen AI applications, accelerating innovation and reducing time to market.

Learn more →

This content was adapted from a thought leadership article originally published in The New Stack, July 30, 2024.
Title: 6 Amazon Bedrock Guidelines for LLM-Based GenAI Apps
Author: Janakiram MSV