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

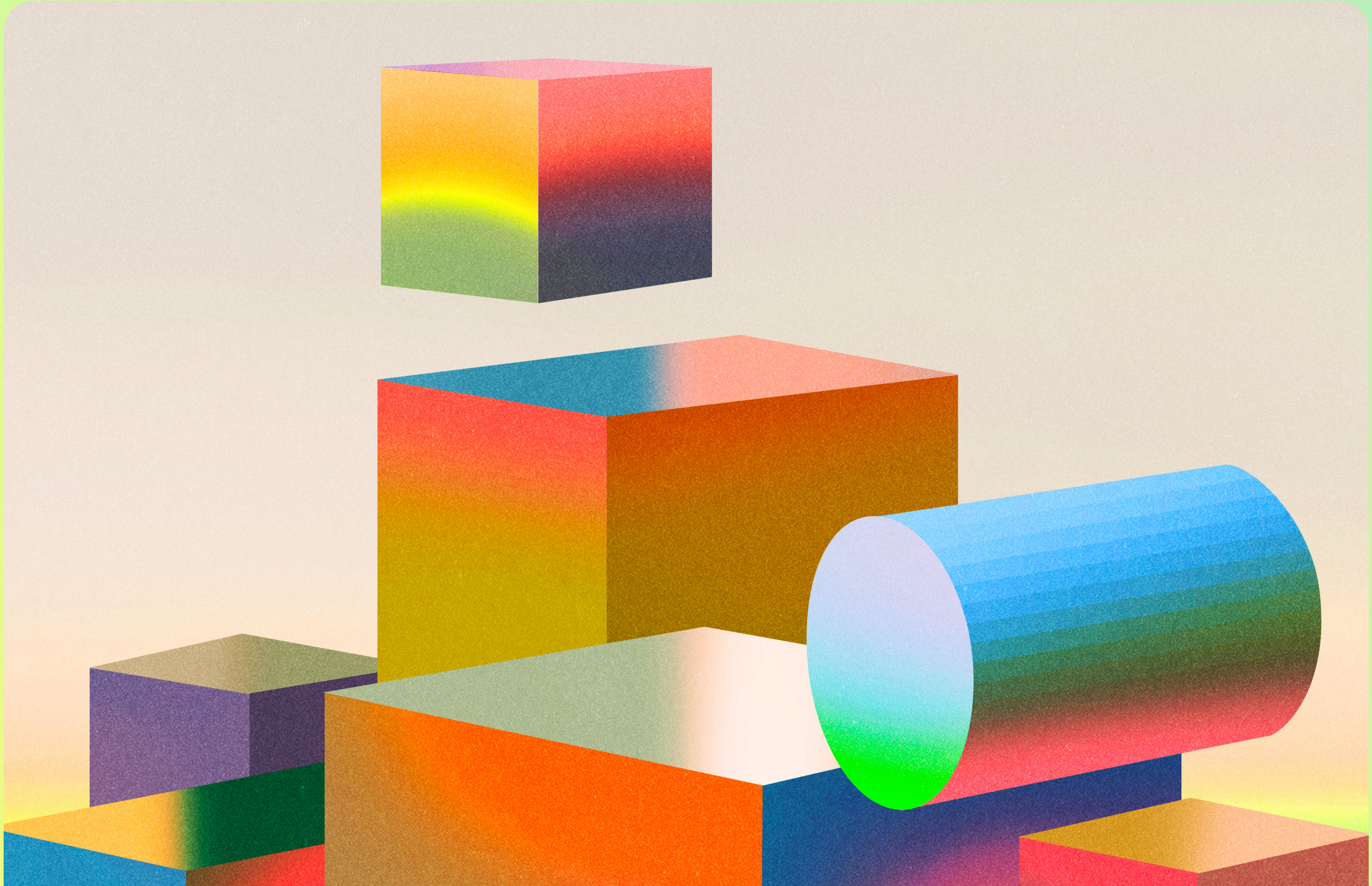


Table of contents

Overview

Foundation models are trained on extensive datasets to understand and generate humanlike 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 artificial intelligence (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 from Amazon Web Services (AWS) that simplifies building and scaling gen AI applications. It provides access to high-performing FMs from leading AI companies, such as AI21 Labs, Anthropic, Cohere, and Amazon, through a unified API. 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 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 organizations 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 provides access to a wide range of high-performing FMs from leading AI companies like AI21 Labs, Anthropic, Cohere, Meta, Mistral AI, Stability AI, and Amazon with its Amazon Titan models. The platform also provides a powerful model evaluation capability that allows customers to evaluate, compare, and select the best FM for their specific use case and requirements. Model evaluation streamlines the often time-consuming process of benchmarking and choosing the right model, reducing the time from weeks to only hours. This allows customers to quickly identify the best model fit and bring new gen AI applications to market faster.

Cultivating innovation with generative AI

The Arizona State University AI Cloud Innovation Center, powered by AWS, collaborated with CGIAR and CIP to develop an interactive, generative AI guide for potato and sweet potato seed requests. Using Amazon Bedrock and Claude Sonnet LLM, they created a chatbot that streamlines the process for researchers seeking germplasm accessions. The solution leverages managed AWS services like Amazon Elastic Container Service (Amazon ECS) for a scalable backend architecture. This innovative approach has transformed a 3-4 month back-and-forth process into a matter of days, significantly improving efficiency in agricultural research.

> “By working backwards, the ASU AI CIC helped us understand users’ most urgent problems and focus on solving them. Together with AWS, we co-created a chatbot to quickly gather information and provide germplasm recommendations, reducing processing time from months to minutes.”
>
> Dr. Bettina Heider, Genetic Resources Specialist, Genebank International Potato Center

10x+

The CGIAR chatbot reduced germplasm request processing time from months to days

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, organizations 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 organization 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 or other third-party tools, and access them on demand through the invoke model API found in Amazon Bedrock.

Scaling Generative AI Across Singapore’s Public Sector

GovTech, Singapore’s digital services agency, partnered with AWS Generative AI Innovation Center to integrate generative AI into MAESTRO, an end-to-end AI/ML development platform. This collaboration enabled cost-effective adoption of generative AI across government agencies, improving cost-performance by up to 75%. MAESTRO leverages Amazon Bedrock and Amazon SageMaker JumpStart for ready-made models, while SageMaker Studio and Canvas provide a no-code interface for building and deploying ML models. Within 9 months, 20 public sector organizations adopted MAESTRO, accelerating AI-powered solutions for enhanced public services.

> “We’re thrilled to partner with AWS to make generative AI more accessible and sustainable for our agencies. This allows us to harness generative AI’s potential while responsibly managing resources.”
>
> Jeffrey Chai, Product Manager, MAESTRO, Government Technology Agency

75%

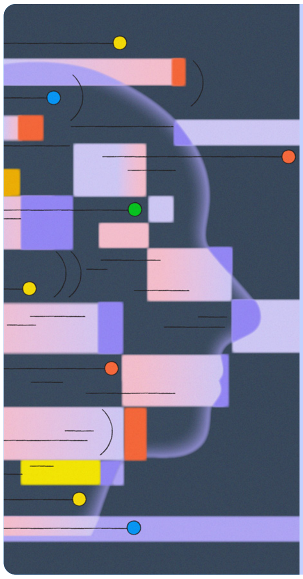
75% improved cost-performance for generative AI workloads

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 retrieval augmented generation (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 service that allows organizations to leverage RAG workflows. It enhances FM responses by incorporating relevant information from an organization's proprietary data sources. The service streamlines the RAG process, including ingesting data from Amazon Simple Storage Service (Amazon S3), converting it into embeddings using FMs, storing embeddings in a vector database, and retrieving pertinent information to augment prompts at query time.

By integrating enterprise data into the gen AI pipeline, Amazon Bedrock Knowledge Bases grounds FM responses in an organization's specific domain knowledge. This improves the accuracy, factual consistency, and relevance of the generated outputs.

Accelerating chatbot deployments with generative AI

LeadSquared leveraged Amazon Bedrock and Amazon Aurora PostgreSQL to enhance their chatbot capabilities. By integrating Retrieval Augmented Generation (RAG) with pgvector extension and large language models, LeadSquared improved chatbot responses and streamlined onboarding. The solution retrieves data from various sources to augment prompts, resulting in more personalized and context-aware interactions. This approach led to a 20% improvement in customer onboarding times and easier chatbot setup.

> “The integration of RAG capabilities using Amazon Aurora PostgreSQL with pgvector and LLMs from Amazon Bedrock has empowered our chatbots to deliver natural language responses, enhanced dialogue management, and reduced manual efforts.”
>
> - Prashant Singh, COO and Cofounder, LeadSquared

20%

LeadSquared improved customer onboarding times by 20% with 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 are advanced AI systems that combine FMs with the ability to interact with external data sources, APIs, and tools. They enable organizations to build autonomous agents that can understand natural language instructions, orchestrate complex, multistep workflows; and take actions beyond generating text responses.

Amazon Bedrock Agents work by first parsing the user's natural language input using an FM. Based on the instructions provided during agent creation, the agent then determines the appropriate course of action, such as retrieving relevant information from a knowledge base, invoking external APIs or tools, or breaking down the request into smaller subtasks. The agent can iteratively refine its understanding, gather additional context from various sources, and ultimately provide a final response synthesized from multiple inputs.

Sírio-Libanês Hospital accelerates reporting with an AI agent

The Sírio-Libanês Hospital implemented Cloudinho, a generative AI agent powered by Amazon Bedrock, to revolutionize its reporting process. By leveraging Claude 3 Sonnet through Amazon Bedrock, Cloudinho reduced report generation time by 99%, from 3 hours to just 1.5 minutes. The solution integrates with Microsoft Teams, providing users with autonomous access to cloud application data. This AWS-based innovation has enhanced financial transparency, enabled near real-time decision-making, and freed up the FinOps team from manual reporting tasks.

> “AWS gives us the best framework to work with digital assets. Working together is key here.”
>
> Vitor Bellot, Coordinator of Operational Excellence, Sírio-Libanês Hospital

99%

Amazon Bedrock helped Sírio-Libanês Hospital reduce reporting time by 99%

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 organizations to implement 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 filter out harmful content, block denied topics, redact sensitive information like personal identifiable information (PII), and enforce content moderation based on their requirements.

Accelerating technical documentation creation

Skyflow, a data privacy vault provider, faced challenges in keeping documentation up-to-date with rapid product releases. To address this, they developed VerbaGPT, a generative AI tool powered by Amazon Bedrock. VerbaGPT uses Contextual Composition and Retrieval Augmented Generation to create accurate first drafts of technical content in minutes.

The solution leverages Amazon Bedrock's foundation models, particularly Anthropic's Claude 3 Sonnet, for its substantial context length. VerbaGPT's architecture includes a RAG pipeline, reusable prompt templates, and an LLM gateway for flexible model selection. A Streamlit-based UI allows easy document uploading and content generation.

> "Using Amazon Bedrock and VerbaGPT, we've dramatically reduced our content creation time from weeks to days. This efficiency boost allows us to keep pace with our rapid product development while maintaining the highest standards of data privacy and security."
>
> Manny Silva, Head of Documentation, Skyflow

3 Days

Reduced content creation time from 3 weeks to as little as 3 days

10 Min

Enabled creation of multiple content types from a single source in 10 minutes

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