Service Name

Amazon Bedrock

Service Overview

Amazon Bedrock is a fully managed service that provides access to high-performing foundation models from leading AI companies and Amazon through a unified API. It enables developers to build generative AI applications with comprehensive capabilities including security, privacy, and responsible AI features. The service offers foundation models from providers like Amazon, Anthropic, AI21 Labs, Cohere, Meta, Mistral AI, OpenAI, and others. Users can experiment with various models, customize them privately with their data using techniques like fine-tuning and retrieval-augmented generation (RAG), and create managed agents that execute complex business tasks without managing infrastructure.

Key Use Cases

- Content generation and document processing for marketing materials, reports, and creative writing
- Conversational AI and chatbots for customer service and support applications
- Code generation, optimization, and vulnerability scanning for software development
- Data analysis and insights extraction from documents, images, videos, and audio files
- Automated business process workflows including travel booking, insurance claims processing, and inventory management
- Personalized recommendations and customer engagement across various industries

Features

- **Foundation Model Choice**: Access to industry-leading foundation models from multiple providers through a single API
- **Model Customization**: Fine-tuning, continued pretraining, and model distillation capabilities for domain-specific adaptation
- **Knowledge Bases and RAG**: Retrieval-augmented generation with support for multimodal data and structured data retrieval
- **Agents and Multi-Agent Collaboration**: Managed agents that execute complex tasks and coordinate with each other for sophisticated workflows
- **Guardrails**: Built-in safeguards against harmful content with automated reasoning and contextual grounding checks

- **Data Automation**: Automated processing and insight extraction from documents, images, videos, and audio files
- **Model Evaluation**: Automatic and human-based evaluation capabilities with LLM-as-a-Judge functionality
- **Service Tiers**: Priority, Standard, and Flex tiers to optimize performance and cost for different workload types

Value Proposition

Amazon Bedrock eliminates the complexity of managing AI infrastructure while providing enterprise-grade security, privacy, and compliance. The service offers unparalleled model choice from leading AI providers through a single API, reducing vendor lock-in and enabling easy model switching:

- Built-in capabilities like Guardrails ensure responsible AI deployment, while features like Knowledge Bases and Agents enable sophisticated applications without extensive development effort
- The serverless architecture means no infrastructure management, and flexible pricing options including service tiers allow cost optimization
- Integration with AWS services provides seamless deployment within existing cloud architectures

Service Integration

Amazon Bedrock integrates deeply with core AWS services including Amazon S3 for data storage and Knowledge Bases, AWS Lambda for serverless functions and agent actions, Amazon DynamoDB for state management, Amazon VPC for secure networking, AWS IAM for access control and permissions, Amazon CloudWatch for monitoring and logging, AWS CloudTrail for audit trails, Amazon EventBridge for event-driven architectures, Amazon OpenSearch for vector databases in Knowledge Bases, and AWS Step Functions for orchestrating complex workflows. The service also works with Amazon SageMaker for model training and evaluation, AWS Backup for data protection, and integrates into Amazon SageMaker Unified Studio for collaborative development environments.

Onboarding and Offboarding

Getting started requires an AWS account with proper IAM permissions for Amazon Bedrock. Beginning June 15, 2025, access to foundation models is enabled by default, though first-time Anthropic model users may need to submit use case details:

- Users can start through the Amazon Bedrock console playgrounds for text and image generation, access the service via APIs and SDKs, or follow step-by-step tutorials

- The setup involves creating IAM roles with necessary policies, requesting model access if needed, and choosing appropriate foundation models for specific use cases
- Multiple getting started resources including tutorials, workshops, and code examples are available to accelerate onboarding

Getting Started Guide: https://docs.aws.amazon.com/bedrock/latest/userguide/getting-started.html

Data Removal

Amazon Bedrock follows AWS data handling practices where customer data is not retained after processing. Users can delete custom models, Knowledge Bases, agents, and associated resources through the console or APIs. Data stored in integrated services like S3 must be deleted separately by the customer. The service provides secure data handling with no sharing of customer inputs or outputs with third-party model providers, ensuring complete data privacy during and after service usage.

Backup/Restore and Disaster Recovery

Amazon Bedrock is a managed service where AWS handles infrastructure resilience and availability across multiple Availability Zones. Customer-created resources like custom models, Knowledge Bases, and agent configurations are maintained by AWS with built-in redundancy. For customer data in integrated services like S3, standard AWS backup and disaster recovery capabilities apply. The service itself does not require customer-managed backup procedures as the foundation models and service infrastructure are managed by AWS.

Backup Capabilities

Amazon Bedrock does not require traditional backup capabilities as it provides access to foundation models and managed services. Custom models created through fine-tuning are stored and maintained by AWS:

- Customer data in Knowledge Bases relies on the underlying data sources in S3 or other supported services
- AWS Backup integration is available for the underlying storage services that customers use with Bedrock, but the service itself operates as a managed offering without customer backup requirements

Disaster Recovery

Amazon Bedrock provides built-in disaster recovery through AWS's multi-AZ architecture and global infrastructure. The service is designed for high availability with automatic failover capabilities:

- While the service does not directly integrate with AWS Elastic Disaster Recovery, the underlying infrastructure follows AWS disaster recovery best practices
- Customer applications using Bedrock should implement their own disaster recovery strategies for application-level components and data stored in integrated AWS services

Recovery Objectives

Amazon Bedrock supports customer RPO and RTO objectives through its highly available, managed architecture with automatic scaling and multi-AZ deployment. The service provides consistent API availability and can handle traffic spikes without manual intervention. For custom models and configurations, AWS maintains redundancy to ensure quick recovery. Customers can achieve low RTO by leveraging multiple AWS regions and implementing proper error handling and retry logic in their applications using Bedrock APIs.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/bedrock/latest/userguide/quotas.html

FAQ: https://aws.amazon.com/bedrock/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/bedrock/latest/userguide/what-is-bedrock.html

User Guide: https://docs.aws.amazon.com/bedrock/latest/userguide/getting-started.html

API Reference:
https://docs.aws.amazon.com/bedrock/latest/APIReference/welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/bedrock/pricing/

Cost Optimization

Amazon Bedrock offers unique cost optimization through service tiers (Priority, Standard, Flex) allowing workload-specific pricing, with Flex tier providing significant discounts for non-time-sensitive tasks. Prompt caching reduces costs up to 90% by caching repeated prompt prefixes:

- Model Distillation creates smaller, faster models that are up to 75% less expensive with minimal accuracy loss

- Intelligent Prompt Routing automatically selects the most cost-effective model, reducing costs up to 30%
- Batch processing provides cost advantages for large-scale workloads, while Provisioned Throughput offers predictable pricing for consistent workloads
- The pay-per-use model with token-based pricing ensures customers only pay for actual usage

Savings Considerations

Main cost savings come from the serverless architecture eliminating infrastructure costs and the ability to choose optimal models for each task rather than over-provisioning. Service tier selection allows matching cost to performance requirements, with Flex tier offering substantial savings for batch workloads:

- Prompt engineering and management reduce token usage through optimized prompts
- Model customization reduces the need for expensive fine-tuning by leveraging Knowledge Bases and RAG
- Multi-model access prevents vendor lock-in enabling cost-effective model switching
- The elimination of model hosting, scaling, and maintenance costs provides significant operational savings compared to self-managed AI infrastructure