

Streamline critical business tasks across your enterprise with generative AI

Generative models are at the forefront of revolutionizing the capabilities of artificial intelligence (AI). However, these models can have limitations. Trust is a primary concern, as AI-generated content can sometimes be unreliable or biased. Potential issues like hallucinations, where the AI generates plausible-sounding but incorrect information, pose challenges for business use. Additionally, access to specific and current data can be limited, impacting the model's ability to provide up-to-date information.

Retrieval-augmented generation (RAG) can help generative AI models build trust with users. RAG systems improve the relevance and accuracy of generative AI responses by incorporating information from data sources that were not a part of pre-trained models.

Cohere's Generative AI and Advanced Retrieval models work seamlessly together for applications requiring RAG. With Cohere's models on Amazon Web Services (AWS), organizations can seamlessly connect and retrieve information from documents and enterprise data sources to build powerful AI applications.

How it works

Cohere allows developers and enterprises to build large language model (LLM)-powered applications by offering simple-to-use models on AWS to deploy them securely and privately.

In collaboration with:



Better together: AWS and Cohere

Cohere, a pioneer in AI, is redefining intelligent enterprise solutions. Cohere builds world-class foundational models that enable businesses to augment their competitive advantage with AI. The company's models are uniquely suited for enterprises, providing ease of use, security, and privacy controls. Cohere has three primary AI models: Command, Embed, and Rerank as well as RAG solutions. Cohere is a third-party (3P) model provider partner of AWS, and the company's models can be accessed through all AWS platforms, including Amazon Bedrock, Amazon SageMaker, Amazon SageMaker JumpStart, and AWS Marketplace.

Cohere's foundational models

Command R family

Command R is a scalable LLM family that includes Command R and R+ purpose-built to excel at real-world enterprise use cases. Command R family is designed to offer:

- Advanced RAG with citation to help reduce hallucinations.
- Multilingual coverage in 10 languages to support global business operations.
- Tool Use to automate complex business processes.

Embed

Embed models can be used to generate embeddings from text or classify text based on various parameters. Embeddings can be used to estimate semantic similarity between two texts. For example, embeddings can choose a sentence that is most likely to follow another sentence, or categorize user feedback. When used with Cohere's Classify endpoint, embeddings can be used for any classification or analysis task.

Rerank

Rerank models are designed to enhance enterprise search and RAG systems. The models sort text inputs by semantic relevance to a specified query. They are often used to sort search results returned from an existing search solution.

RAG capabilities

Command R and R+ are optimized for long context tasks, such as RAG and using external application programming interfaces (APIs) and tools. The models are built to work in concert with Cohere's Embed and Rerank models to provide seamless integration for RAG applications and excel at enterprise use cases.

Customer story: Building a knowledge assistant with RAG

A large professional services firm wanted to build a RAG-based knowledge assistant using a knowledge base of tens of millions of documents, initially for internal users and ultimately as a product for its clients. Its proof-of-concept had a combination of commercial generative models, open-source embeddings, and reranking solutions for search. However, the firm was looking for a long-term solution that could work in secure customer environments.

The firm replaced its open-source retrieval models with Cohere Embed and Cohere Rerank after testing them side-by-side and seeing a 40 percent increase in retrieval accuracy. The team is currently replacing their generative model with a fine-tuned Cohere Command. The company also gained productivity improvements of over 20 percent by automating repetitive tasks. With Cohere, the firm can deploy the solution anywhere within their own network and in any client environment while staying fully compliant with demanding security requirements.

Simplify your generative AI journey

Cohere on AWS provides businesses with the tools to help you augment a competitive advantage with generative AI including:

Streamlined access to LLMs

Enterprises can seamlessly access Cohere's models as an API on Amazon SageMaker and Amazon Bedrock and incorporate AI features into their applications and workflows without the need to train models or build the infrastructure required for deploying and managing LLMs.

Security and data privacy

With Cohere on AWS, your data remains securely managed and protected. AWS provides enterprise-grade security and privacy tools integrated into the environment, enabling encryption, access control, and privacy.

Scalable, cost-efficient deployments

Amazon Bedrock and Amazon SageMaker are optimized for AI workloads, with high-performance compute, storage, and networking resources. AWS provides tools that make it seamless to deploy models at scale. Cohere's models provide high accuracy, low latency, and high throughput at an industry-leading cost.

Cohere in the AI/ML stack

Cohere products

Application layer

Framework layer

- Inference
- Model serving
- Fine-tuning
- Connectors

Foundation model layer

- Command
- Rerank
- Embed

Infrastructure

Generative AI on AWS

Start generating innovation with these performant, cost-effective, and easy-to-use tools purpose-built for generative AI.



Amazon Bedrock

The easiest way to build and scale generative AI applications with foundation models.



Amazon SageMaker

Build, train, and deploy machine learning models at scale.



Amazon SageMaker JumpStart

Deploy prebuilt machine learning solutions in just a few clicks.

[Learn more >](#)

Benefits and outcomes of generative AI

1. Enhanced efficiency and productivity

- **Automated tasks:** Cohere's models can be used to automate repetitive tasks such as text summarization, content generation, and data analysis, freeing up time and resources.
- **Improved accuracy and speed:** Cohere's LLMs can process large datasets and perform tasks with greater accuracy and speed compared to traditional AI methods.

2. Deeper insights and understanding

- **Identifying patterns and trends:** The models can analyze vast amounts of text and data to identify hidden patterns, trends, and relationships, enabling better decision-making.
- **Generating creative content:** The models can generate diverse and creative text formats, fostering innovation and exploration in various domains.

3. Improved user experience and personalization

- **Conversational interfaces:** Cohere's models power chatbots and virtual assistants, allowing for natural language interactions and personalized user experiences.
- **Tailored content and recommendations:** The models can personalize content recommendations, product descriptions, and marketing messages based on individual user preferences and needs.

4. Increased accessibility and language agnostic

- **Breaking down language barriers:** Cohere's models can translate languages accurately and efficiently, promoting communication and collaboration across cultures.
- **Democratizing access to information:** The models can summarize complex information, making it easier for people with varying levels of expertise to understand and utilize.

5. Continuous learning and improvement

- **Evolving with new data:** Cohere's LLMs are continuously trained on new data, allowing them to improve their performance and capabilities over time.
- **Adapting to specific needs:** Cohere models can be fine-tuned for specific tasks and domains, catering to diverse industry needs and applications.

Ready to gain a competitive advantage with generative AI?

[Learn more](#) about Cohere's deployment options on AWS.