

nextwork.org

Set Up a RAG Chatbot in Bedrock



Teenamsh Kanamarlapudi

Test Knowledge Base >

Generate responses ⚙️ 🔍

∞ Llama 3.3 70B I... v1 ⏪ ⏴ On-demand

Hello

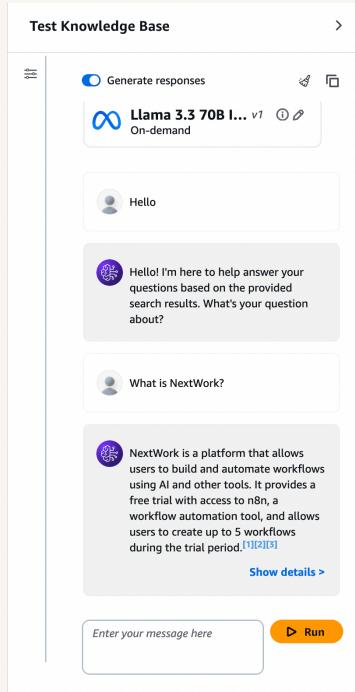
Hello! I'm here to help answer your questions based on the provided search results. What's your question about?

What is NextWork?

NextWork is a platform that allows users to build and automate workflows using AI and other tools. It provides a free trial with access to n8n, a workflow automation tool, and allows users to create up to 5 workflows during the trial period. [\[Read\]](#)

Show details >

Enter your message here ➔ Run





Teenamsh Kanamarlapudi

NextWork Student

NextWork.org

Introducing Today's Project!

RAG (Retrieval Augmented Generation) is an AI framework that works like LLMs by allowing them to access external data sources during the generation process. In this project, I will demonstrate RAG by making my bot answer my queries.

Tools and concepts

Services I used were Knowledge Base, S3 Bucket, and Vector Store. Key concepts I learned include embedding and key connections for all services and why they need to be connected.

Project reflection

This project took me approximately 3 hours. The most challenging part was getting my IAM admin user set up because I hadn't had one before to use, so I had to change permissions multiple times. It was most rewarding to get the result.

I did this project to learn about RAG and learn to make my own chatbot. This project definitely met my goal of learning RAG. This was definitely a fun project to work on and do.



Understanding Amazon Bedrock

Amazon Bedrock is an AI model marketplace that allows users to search and use different types of AIs. I'm using Bedrock in this project to use Knowledge Base to create my chatbot.

My Knowledge Base is connected to S3 because that is where my documents will be stored. S3 is a storage system for storing documents, videos, and pictures.

In an S3 bucket, I uploaded 10 different documents for my chatbot. My S3 bucket is in the same region as my Knowledge Base because Bedrock is a region-specific service. Adding on reduces latency.

The screenshot shows the AWS S3 console interface. At the top, a green banner indicates "Upload succeeded" with a note to see the "Files and folders table". Below this, the "Summary" section shows the destination as "s3://nextwork-rag-bedrock-tk". It displays two rows: "Succeeded" (10 files, 138.3 MB (100.0%)) and "Failed" (0 files, 0 B (0%)). The "Files and folders" tab is selected, showing a table of 10 uploaded files. The table includes columns for Name, Folder, Type, Size, Status, and Error. All files are listed as "Succeeded".

Name	Folder	Type	Size	Status	Error
Automate Your Browser with AI Age...	-	application/pdf	17.3 MB	Succeeded	-
Threat Detection with GuardDuty.pdf	-	application/pdf	4.0 MB	Succeeded	-
How to Use DeepSeek.pdf	-	application/pdf	6.2 MB	Succeeded	-
Transcribe Audio Files with AI.pdf	-	application/pdf	13.7 MB	Succeeded	-
Deploy Backend with Kubernetes.pdf	-	application/pdf	15.3 MB	Succeeded	-
Fetch Data with AWS Lambda.pdf	-	application/pdf	16.0 MB	Succeeded	-
Build a Three-Tier Web App.pdf	-	application/pdf	16.6 MB	Succeeded	-
Building an AI Workflow.pdf	-	application/pdf	16.4 MB	Succeeded	-
Create S3 Buckets with Terraform.p...	-	application/pdf	16.5 MB	Succeeded	-
Prompt Engineering.pdf	-	application/pdf	16.4 MB	Succeeded	-



My Knowledge Base Setup

My Knowledge Base uses a vector store, which means the AI will find the documents' embeddings to pair up the query and document data. When I query my Knowledge Base, OpenSearch will grab the most relevant chunks of text to answer my question.

Embeddings are special cards that give key phrases, so when a question is asked, it makes another card with the inquiry and connects them to extract information. The embeddings model I'm using is Titan Text Embeddings v2 because it's efficient by AWS

Chunking is a way AI processes larger texts by making them paraphrased so it is more efficient to process information from my documents. In my Knowledge Base, chunks are set to be paraphrased to 300 tokens or so.

The screenshot shows the configuration steps for a Knowledge Base:

Step 2: Configure data source

Data source name	Account ID	S3 URI
s3-bucket-nextwork-rag-bedrock	406095826214 (this account)	s3://nextwork-rag-bedrock-tk

Customer-managed KMS Key for S3: -

Parsing strategy: DEFAULT

Data deletion policy: DELETE

Step 3: Select embeddings model and configure vector store

Embeddings model	Embedding type	Vector dimensions
Model: Titan Text Embeddings v2	Embedding type: Float vector embeddings	Vector dimensions: 1024

Vector store

Quick create vector store - Recommended
Amazon OpenSearch Serverless



AI Models

AI models are important for my chatbot because they will give responses human-like responses. Without AI models, my chatbot would only respond with chunks from the documents.

To get access to AI models in Bedrock, I had to select the models to request access to the models. AWS needs explicit access because some are expensive to use, to check if there is enough capacity, and to check all the rules and conditions.

Models	Access status	Modality	EULA
▼ Amazon (4)	1/4 access granted		
Titan Text Embeddings V2	Access granted	Embedding	EULA
Nova Pro Cross-region inference	Available to request	Text & Vision	EULA
Nova Lite Cross-region inference	Available to request	Text & Vision	EULA
Nova Micro Cross-region inference	Available to request	Text	EULA
▼ Anthropic (5)	0/5 access granted		
Claude 3.7 Sonnet Cross-region inference	Available to request	Text & Vision	EULA
Claude 3.5 Haiku Cross-region inference	Available to request	Text	EULA
Claude 3.5 Sonnet v2 Cross-region inference	Available to request	Text & Vision	EULA
Claude 3.5 Sonnet Cross-region inference	Available to request	Text & Vision	EULA
Claude 3 Haiku Cross-region inference	Available to request	Text & Vision	EULA
▼ DeepSeek (1)	0/1 access granted		
DeepSeek-R1 Cross-region inference	Available to request	Text	EULA
▼ Meta (8)	2/8 access granted		
Llama 3.3 70B Instruct	Access granted	Text	EULA
Llama 3.2 1B Instruct Cross-region inference	Available to request	Text	EULA
Llama 3.2 3B Instruct Cross-region inference	Available to request	Text	EULA
Llama 3.2 11B Vision Instruct Cross-region inference	Available to request	Text & Vision	EULA
Llama 3.2 90B Vision Instruct Cross-region inference	Available to request	Text & Vision	EULA
Llama 3.1 405B Instruct Cross-region inference	Available to request	Text	EULA
Llama 3.1 70B Instruct Cross-region inference	Available to request	Text	EULA
Llama 3.1 8B Instruct Cross-region inference	Access granted	Text	EULA



Syncing the Knowledge Base

Even though I already connected my S3 bucket when creating the Knowledge Base, I still need to sync because the Knowledge Base doesn't have any information inside yet.

The sync process involves three steps: ingesting - receiving data from the data source, processing - chunk and embed the data, and storing - storing in vector store.

Sync completed for data source - 's3-bucket-nextwork-rag-bedrock'

Amazon Bedrock > Knowledge Bases > nextwork-rag-documentation

nextwork-rag-documentation

Knowledge Base overview

Knowledge Base name nextwork-rag-documentation	Knowledge Base ID CRXK7M1WDO	Log Deliveries Configure log deliveries and event logs in the Edit page.
Knowledge Base description This Knowledge Base stores all documentation at NextWork.	Status Available	Retrieval-Augmented Generation (RAG) type Vector store
Service Role AmazonBedrockExecutionRoleForKnowledgeBase_o1h5u	Created date March 29, 2025, 20:52 (UTC-05:00)	

Data source (1)

Data sources contain information returned when querying a Knowledge Base.

	Sync	Stop sync	Add	Add documents from S3	▼
<input checked="" type="checkbox"/> Data sou...					< 1 >
<input checked="" type="checkbox"/> s3-bucket...	Available	S3	40609582...	s3://next...	March 29,...
					-
					Default
					DEFAL

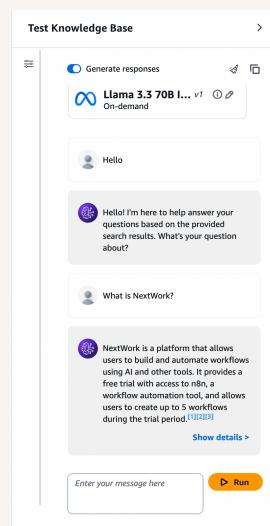


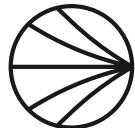
Testing My Chatbot

I initially tried to test my chatbot using Llama 3.1 8B as the AI model, but I got an error for my query. I had to switch to Llama 3.3 70B because it is newer and more efficient, they are less resource-intensive.

When I asked about topics unrelated to my data, my chatbot gave me a response saying, "Sorry, I am unable to assist you with this request." This proves that anything outside of the data provided will not be answered.

You can also turn off the Generate Responses setting to get the chunks from the documents to dive deeper into them rather than getting human-like responses.





NextWork.org

Everyone should be in a job they love.

Check out nextwork.org for
more projects

