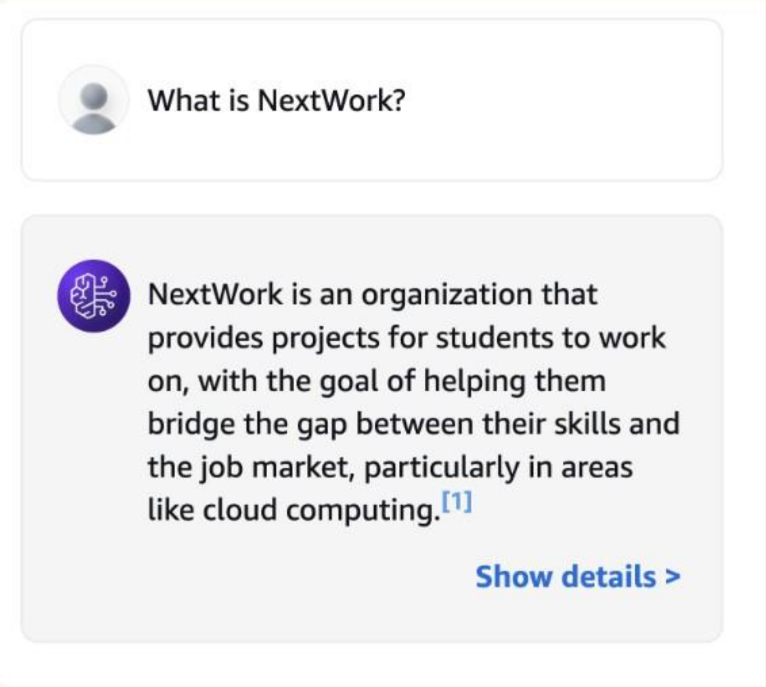
NextWork.org

RAG-based Chatbot with Bedrock



What is NextWork?

NextWork is an organization that provides projects for students to work on, with the goal of helping them bridge the gap between their skills and the job market, particularly in areas like cloud computing.[1]

Introducing Today's Project!

RAG (Retrieval Augmented Generation) is an AI technique that lets user train an AI model on their own personal documents. In this project, I will demonstrate RAG by setting up a RAG chatbot in Amazon Bedrock.

Tools and concepts

Services I used were Amazon Bedrock, S3 and Opensearch Serverless. Key concepts I learnt include knowledege bases, requesting access to AI models, how chatbot generate responses, Vector stores.

Project reflection

This project took me approximately 2 hours. The most challenging part was running into error with AI models and understanding on-demand vs preprovisioned inference. It was most rewarding to level up chatbots response!

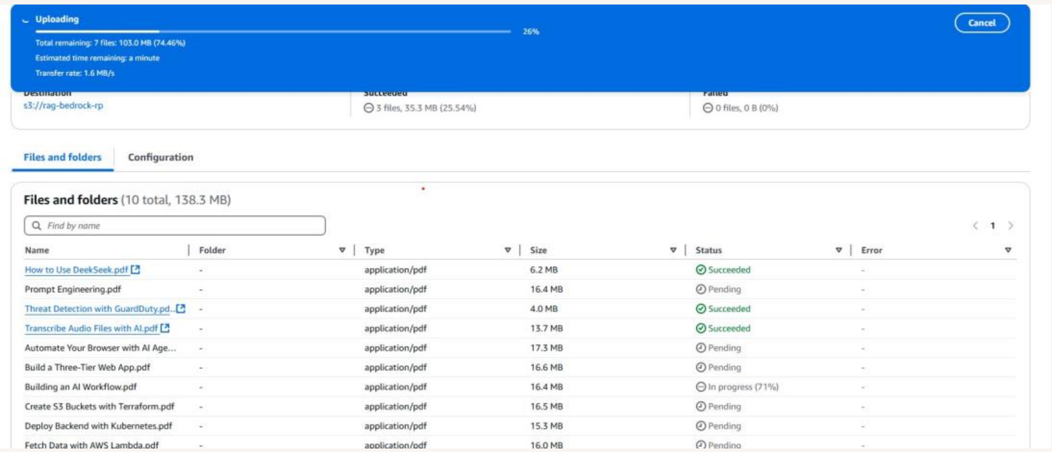
i did this project to Learn more about AWS and How i turn the data into an AI chatbot. YES this project met my goals of learning on of a new key things to learn for today!

Understanding Amazon Bedrock

Amazon Bedrock is an AWS service that makes it easy to build AI applications as bedrock is like an AI model marketplace which we can find use test models from different users/providers, using bedrock to create knowledge base.

My Knowledge Base is connected to S3 because S3 is going to be a storage/source for our knowledge bases raw documents. S3 is AWS's storage service, wehere we can store all kind of objects such as videos, docus, audio in the same bucket :)

In an S3 bucket i uploaded the documents that will make up thw chatbots knowledge. S3 bucket is in the same region as our knowledge base as bedrock is regional service data must live in the same region as bedrock resources!



Uploading

26%

Total remaining: 7 files: 103.0 MB (74.46%)

Estimated time remaining: a minute

Transfer rate: 1.6 MB/s

Destination	Succeeded	Failed
s3://rag-bedrock-rp	3 files, 35.3 MB (25.54%)	0 files, 0 B (0%)

Files and folders (10 total, 138.3 MB)

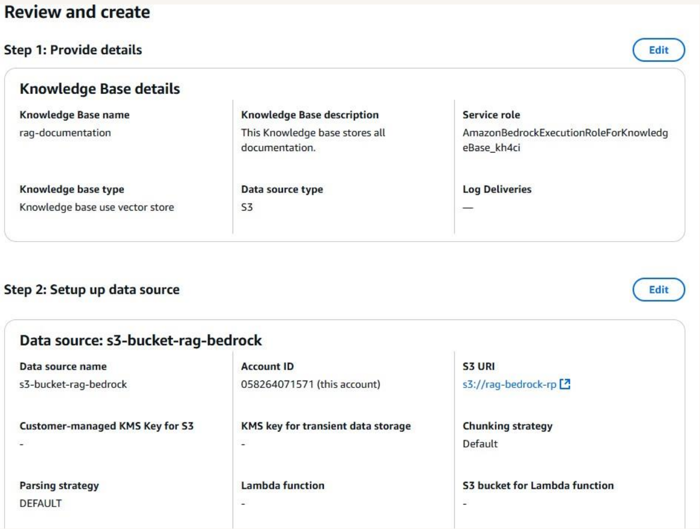
Name	Folder	Type	Size	Status	Error
How to Use DeekSeek.pdf	-	application/pdf	6.2 MB	✓ Succeeded	-
Prompt Engineering.pdf	-	application/pdf	16.4 MB	Pending	-
Threat Detection with GuardDuty.pd...	-	application/pdf	4.0 MB	✓ Succeeded	-
Transcribe Audio Files with AI.pdf	-	application/pdf	13.7 MB	✓ Succeeded	-
Automate Your Browser with AI Age...	-	application/pdf	17.3 MB	Pending	-
Build a Three-Tier Web App.pdf	-	application/pdf	16.6 MB	Pending	-
Building an AI Workflow.pdf	-	application/pdf	16.4 MB	In progress (71%)	-
Create S3 Buckets with Terraform.pdf	-	application/pdf	16.5 MB	Pending	-
Deploy Backend with Kubernetes.pdf	-	application/pdf	15.3 MB	Pending	-
Fetch Data with AWS Lambda.pdf	-	application/pdf	16.0 MB	Pending	-

My Knowledge Base Setup

My Knowledge Base uses a vector store, which means a searxh engine that stores data based on their sematic meaning when query our knowledge base, openSearch will find relevant chunks of data to Query and passed it to bedrock.

Embeddings are vector representations od the semantic meaning of a chunk of text. The embedding model that we will be using is titan Text embeddings v2 as its fast, accurate, and a lot more affordable

Chunking is a process of splitting up text into smaller pieces in chucks, chunks are set to be about 300 tokens in size each.



Review and create

Step 1: Provide details

Knowledge Base details

Knowledge Base name rag-documentation	**Knowledge Base description** This Knowledge base stores all documentation.	**Service role** AmazonBedrockExecutionRoleForKnowledgeBase_kh4ci
Knowledge base type Knowledge base use vector store	**Data source type** S3	**Log Deliveries** —

Step 2: Setup up data source

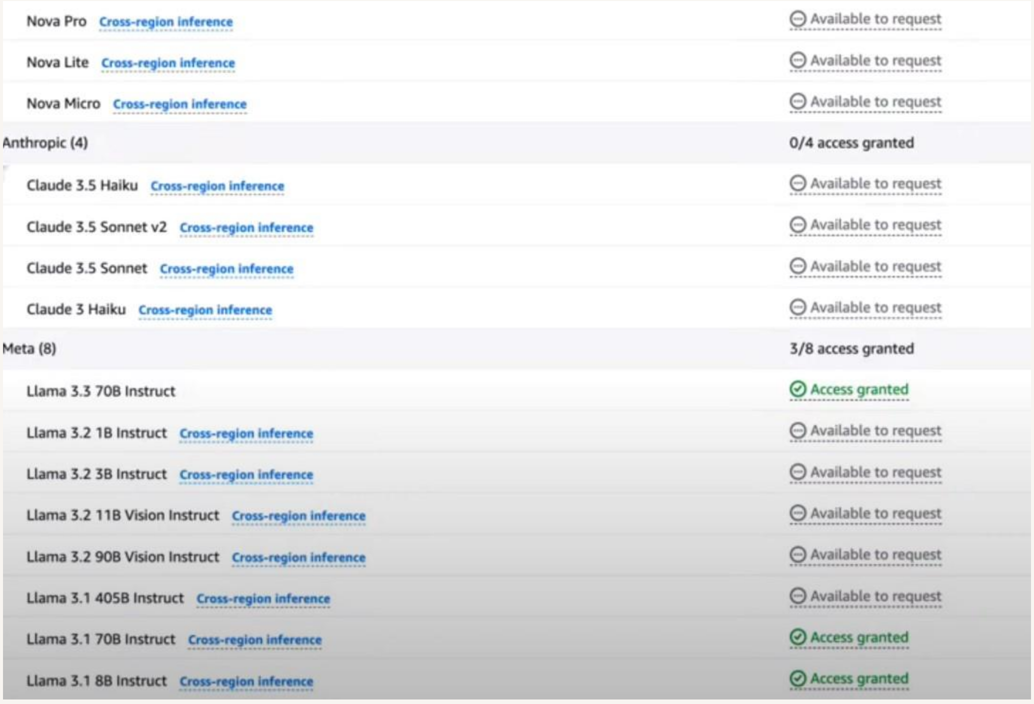
Data source: s3-bucket-rag-bedrock

Data source name s3-bucket-rag-bedrock	**Account ID** 058264071571 (this account)	**S3 URI** s3://rag-bedrock-rp
Customer-managed KMS Key for S3 -	**KMS key for transient data storage** -	**Chunking strategy** Default
Parsing strategy DEFAULT	**Lambda function** -	**S3 bucket for Lambda function** -

AI Models

i nthis step i will set up an AI model that will become the brains of the chatbot it is going to help the chatbot respond like chat like human like messages rather than chunks of text like a bot.

To get access to AI model in Bedrock we visit the "Model Access" page and request Access explicity. AWS needds explicit access because some AI model providers have extra forms/rules if we wanted to use them and AWS needs to check for availability.

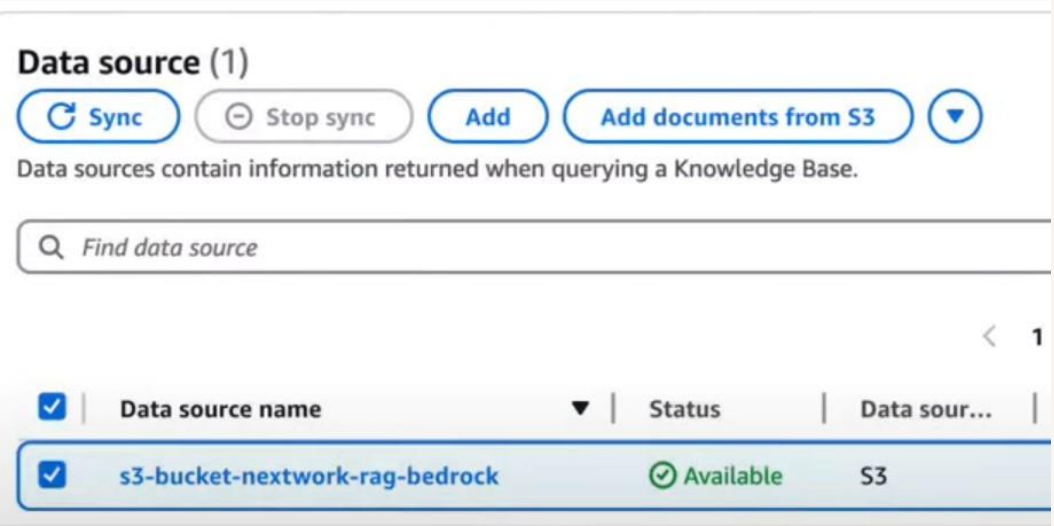


Nova Pro Cross-region inference	⊖ Available to request
Nova Lite Cross-region inference	⊖ Available to request
Nova Micro Cross-region inference	⊖ Available to request
Anthropic (4)	0/4 access granted
Claude 3.5 Haiku Cross-region inference	⊖ Available to request
Claude 3.5 Sonnet v2 Cross-region inference	⊖ Available to request
Claude 3.5 Sonnet Cross-region inference	⊖ Available to request
Claude 3 Haiku Cross-region inference	⊖ Available to request
Meta (8)	3/8 access granted
Llama 3.3 70B Instruct	✓ Access granted
Llama 3.2 1B Instruct Cross-region inference	⊖ Available to request
Llama 3.2 3B Instruct Cross-region inference	⊖ Available to request
Llama 3.2 11B Vision Instruct Cross-region inference	⊖ Available to request
Llama 3.2 90B Vision Instruct Cross-region inference	⊖ Available to request
Llama 3.1 405B Instruct Cross-region inference	⊖ Available to request
Llama 3.1 70B Instruct Cross-region inference	✓ Access granted
Llama 3.1 8B Instruct Cross-region inference	✓ Access granted

Syncing the Knowledge Base

The sync process involves three steps: Ingesting which Bedrocks takes the data S3,
Processing which Bedrock chunks and embeds the datasync and
Storing which Bedrock stores the processed data int the vector store, OpenSearch Serverless

The sync process involves three steps: Ingesting which Bedrocks takes the data S3,
Processing which Bedrock chunks and embeds the datasync and
Storing which Bedrock stores the processed data int the vector store, OpenSearch Serverless



Data source (1)

Data sources contain information returned when querying a Knowledge Base.

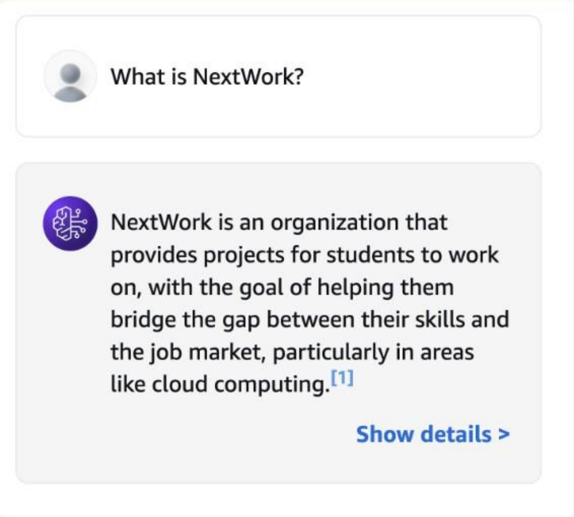
☑	Data source name	Status	Data sour...
☑	s3-bucket-nextwork-rag-bedrock	✓ Available	S3

Testing My Chatbot

I initially tried to test my chatbot using Llama 3.1 8B as the AI model but it occured an error it was not availabe on demand! I had to switch to Llama 3.3 70B because it was offered by AWS as its newer and efficient model.

When I asked about topics unrelated to my data, my chatbot responds that it cannot helps us with this request. Hence, it proves that the chatbot only knows the information that we provided.

You can also turn off the Generate Responses setting to see the raw chunks of data directly from our knowledge base.



What is NextWork?

NextWork is an organization that provides projects for students to work on, with the goal of helping them bridge the gap between their skills and the job market, particularly in areas like cloud computing.[1]