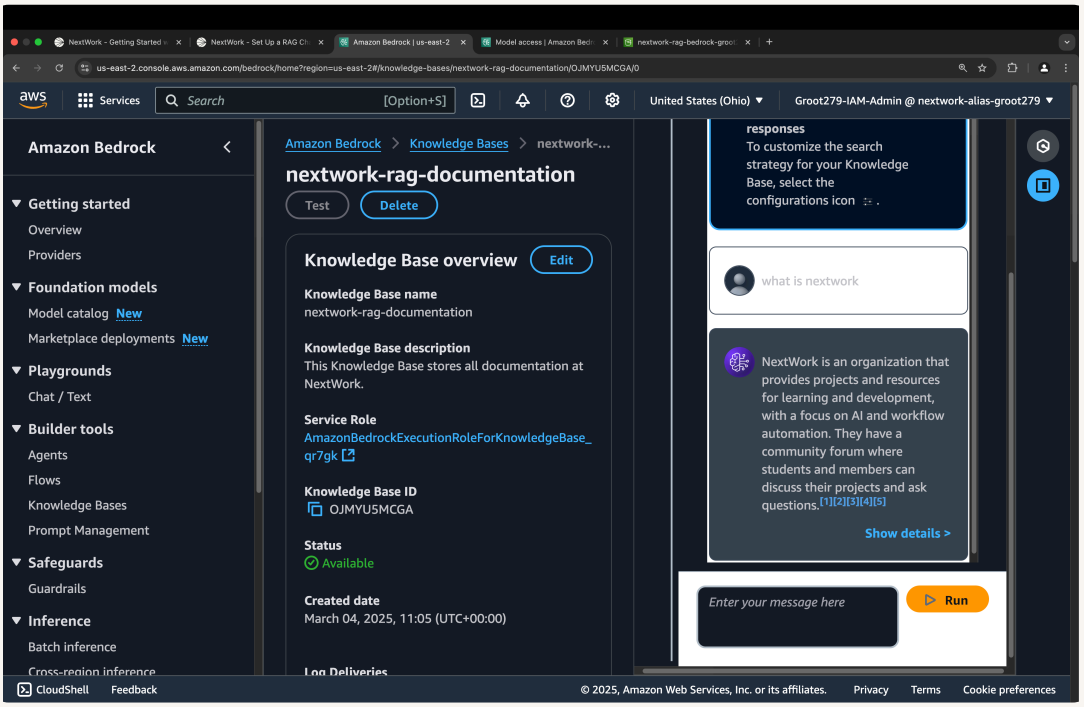
nextwork.org

Set Up a RAG Chatbot in Bedrock

TA tahirgroot@gmail.com



TA
tahirgroot@gmail.com
NextWork Student
NextWork.org

Introducing Today's Project!

RAG (Retrieval Augmented Generation) is an AI technique that lets you train an AI model on your 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 Severles. Key concepts I learnt include Knowledge bases, requesting access to AI modles, how chatbots geberate responses (i.e. AI nodles + Knowledge base), vector stores.

Project reflection

This project took me approximately 1 hour The most challenging part was the error with AI models(and understanding on-demand vs pre-provisioned inference). It was most rewarding to update level up my chatbot's responses.

I did this project today to learn more about RAG and Bedrock. This project definitely met my goals - awesome to learn both over a hands on project.

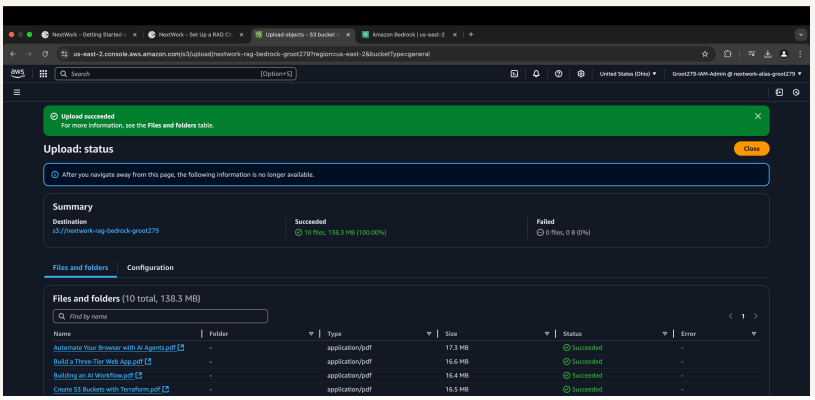
TA
tahirgroot@gmail.com
NextWork Student

Understanding Amazon Bedrock

Amazon Bedrock is an AWS service that make it easy to build generative AI applications - it's like an AI model marketplace that lets us find, use and test models form different providers I'm using Bedrock in this project to create a knowledge base.

My Knowledge Base is connected to S3 because S3 is going to be the strage/source for our knowledge Base's raw documents. S3 is AWS's storage service, wherei can store all kinds of objects (e.g. videso, documents, audio) in the same bucket.

In an S3 bucket, I uploaded the documents that will make up my chatbot's knowledege. My S3 bucket is in the same region as my Knowledge Base because Bedrock is a regional service - data must live in the same region as the Bedrock resource (kbase).



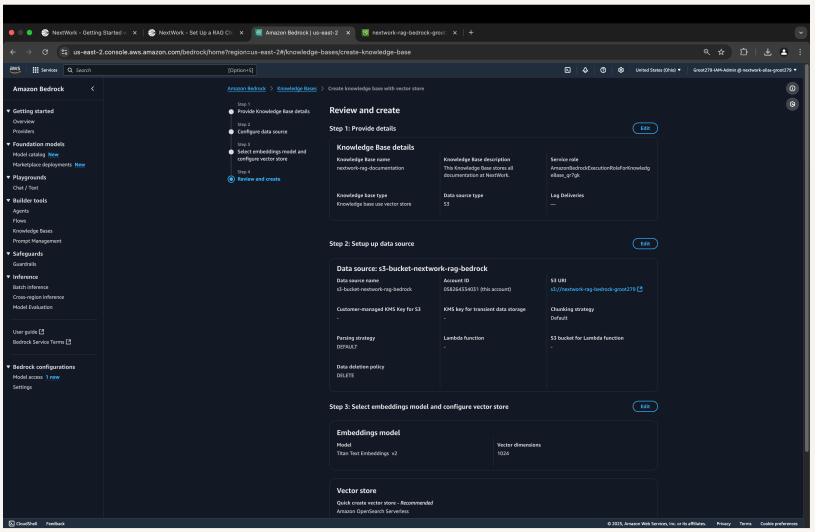
TA
tahirgroot@gmail.com
NextWork Student

My Knowledge Base Setup

My Knowledge Base uses a vector store, which means a search engine/database that stores data based on their semantic meaning! When I query my Knowledge Base, OpenSearch will find the relevant chunks of data to the query and passed it to Bedrock.

Embeddings are vector represntations of the semantic meaning of a chunk of text. The embedding model I'm using is Titan Text Embeddings v2 because its fast, accurate and a lot more affordable!

Chunking is the process of splitting up text into smaller pieces i.e. chunks. This helps with searching for data more efficiently in the vectir store. In our knowledge Base, chunks are set to be about 300 tokens in size each!

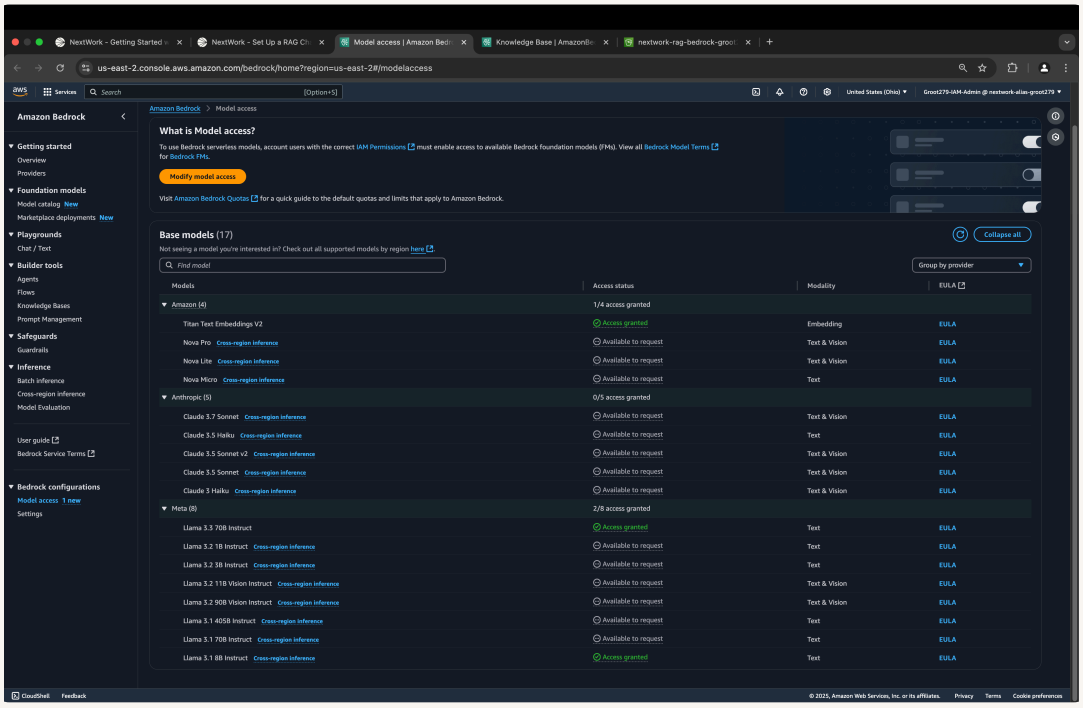


TA
tahirgroot@gmail.com
NextWork Student

AI Models

In this step, I will setup an AI model that will become the brains of the chatbot! This is going to help the chatbot repsond with chat-like, human-like-messages (rather than raw chuncks of text).

To get access to AI models in Bedrock, I had to visit the "Model Access" page and request access explicitly! AWS needs explicit access because some AI model providers have extra forms/rules if I wanted to use them, AWS needs to check availability

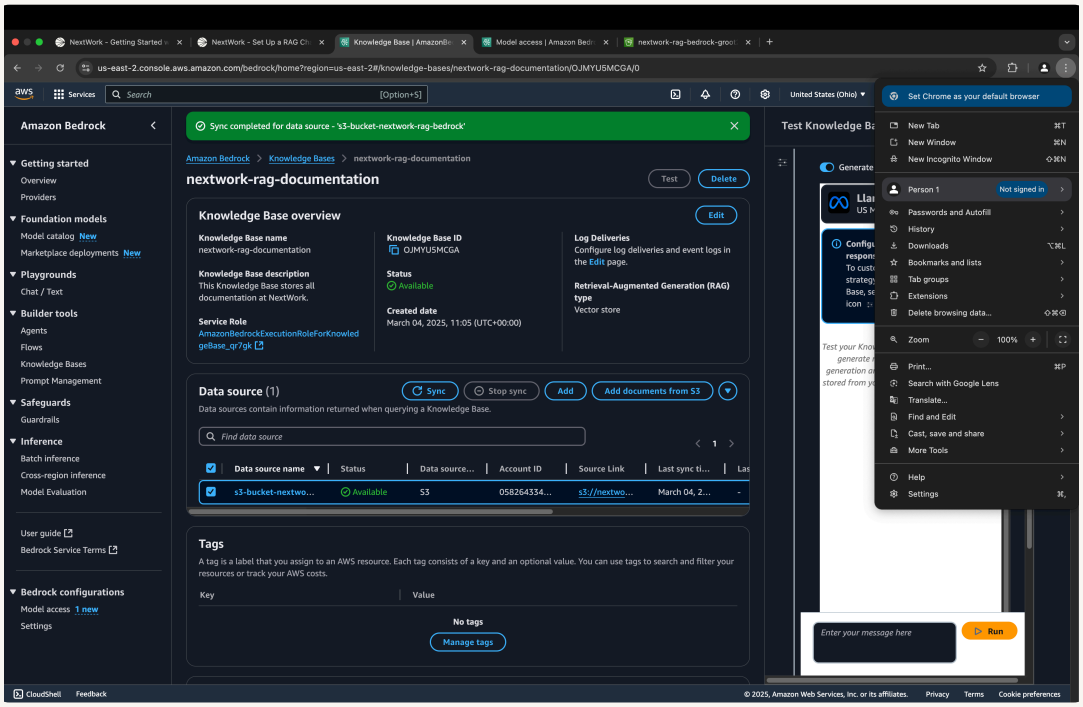


TA
tahirgroot@gmail.com
NextWork Student

Syncing the Knowledge Base

Even though I already connected my S3 bucket when creating the Knowledge Base, I still need to sync my data. This is because syncing is what actually moves the data from S3 into my Knowledge Base + OpenSearch Severeless

The sync process involves three steps: Ingesting(i,e Bedrock tajes the data from S3), Processing (i.e. Bedrock chunks and embeds the data) and storing(i.e. Bedrock stores the oricessed data in the vector store, Open search Severless).



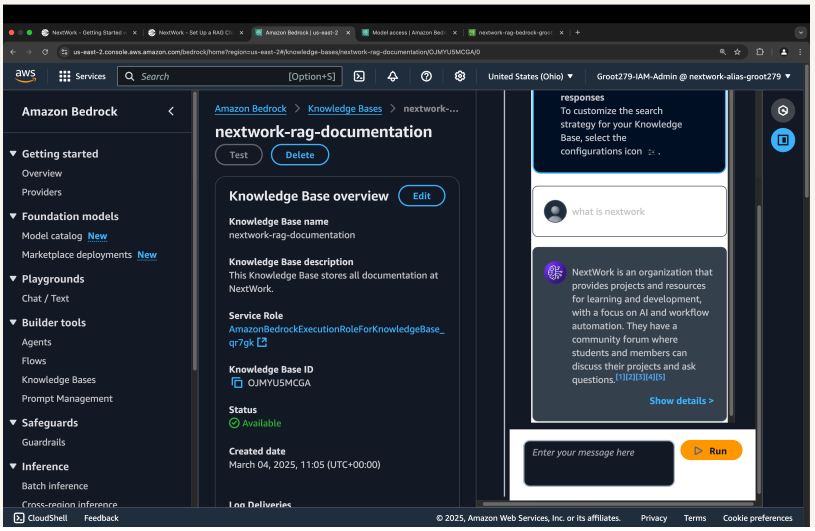
TA
tahirgroot@gmail.com
NextWork Student

Testing My Chatbot

I initially tried to test my chatbot using Llama 3.1 8B as the AI model, but it triggered an error - it was not available on demand. we had to switch to Llama 3.3 70b becuase it was offered on-demand by AWS (since it's a newer, efficient model).

When I asked about topics unrelated to my Knowledge Base's data, our chatbot responds that it cannot help us with this request. This proves that the chatbit only know the information that we gave it, - it wont know even coomon knowledge outside!

I can also turn off the generate responses setting to see the raw chunks of data directly from the Knowledge Baase. When i tested this, the chatbot gave a list of 5 paragraphs to answer a question, whereas the AI model will convert it to a sentence.



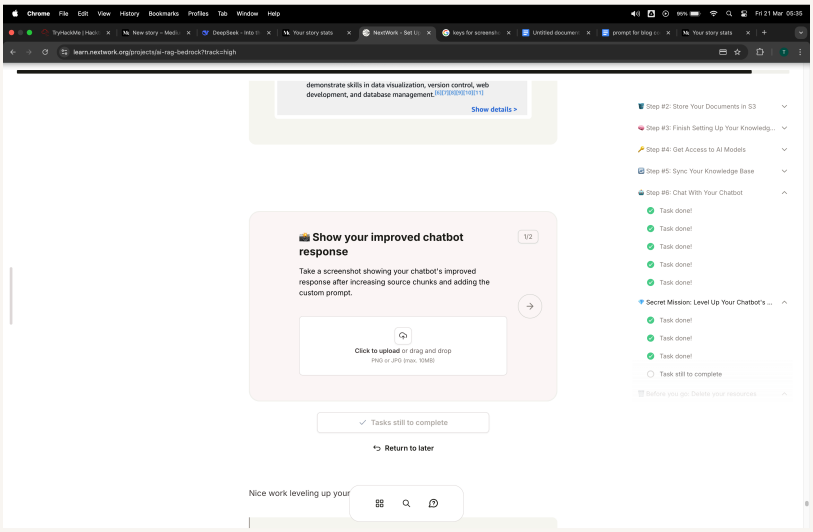
TA
tahirgroot@gmail.com
NextWork Student

Upgrading My Chatbot

In a project extension, I looked into increasing the number of source chunks, which are the number of chunks of text that the knowledge base will return for a query. This will improve my chatbot's responses by giving our AI model more data to use.

I also added a custom prompt that tells my chatbot to always respond by mentioning skills that the students learnt, even if the students didnt ask for it directly, I also gave context that the the database contains documnetation for projects.

After adjusting the source chunks and the generation prompt, my chatbot's response became so much more detailed! The chatbot would talk about skills we've learnt, and used 10 or more examples within the response



NextWork.org

Everyone should be in a job they love.

Check out nextwork.org for more projects