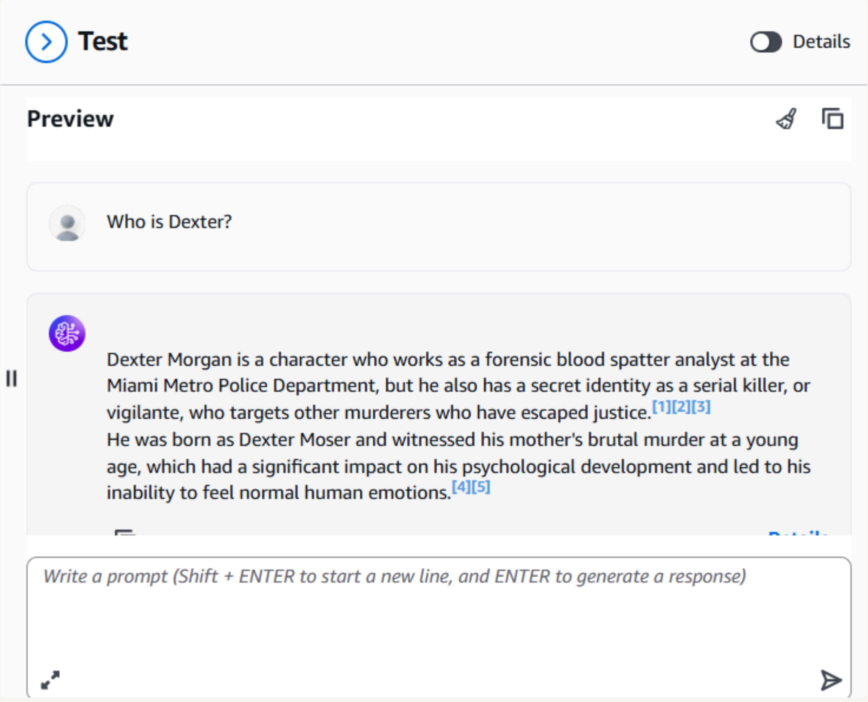
Set Up a RAG Chatbot in Bedrock

Hamza Sajid



Who is Dexter?

Dexter Morgan is a character who works as a forensic blood spatter analyst at the Miami Metro Police Department, but he also has a secret identity as a serial killer, or vigilante, who targets other murderers who have escaped justice.[1][2][3]
He was born as Dexter Moser and witnessed his mother's brutal murder at a young age, which had a significant impact on his psychological development and led to his inability to feel normal human emotions.[4][5]

Write a prompt (Shift + ENTER to start a new line, and ENTER to generate a response)

Introducing Today's Project!

RAG is an AI-technique where it recieves data from an external database and uses it to generate an accurate response. In this project, I will demonstrate RAG by using a fake-person info (which I should know) and use it to ask about that person.

Tools and concepts

Services I used were: 1. Claude - Generating everything about Dexter Morgan and store that Information in the S3 Bucket. 2. S3 Bucket - Storing the Info 3. K-Base - Chunking and get the necessary info

Project reflection

This project took me approximately took 30 mins to complete.

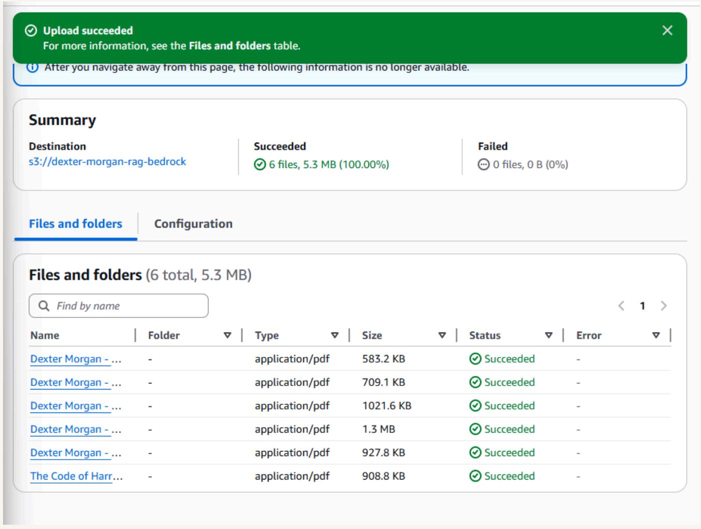
Enjoyed making a Info about a Psycho-Murder XD.

Understanding Amazon Bedrock

Amazon Bedrock is an AWS service that amkes it easy for us to build generative AI appliations - it's like a AI Model marketplace that lest us find, use and test moedls from different providers We're using Bedrock to create a Knowledge Base.

My Knowledge Base is connected to S3 becuase I've chosen it as a default data storage to know everything about Dexter Morgan. S3 is the go-to storage for your K-Base to use for retrieving volitial information about Dexter Morgan while being safe.

In an S3 bucket, I uploaded everything about Dexter Morgan. From his Psychological Analysis to his Key Relationships. My S3 bucket is in the same region as my Knowledge Base because it helps reduce latency and most importantly, cost.



Upload succeeded
For more information, see the **Files and folders** table.

After you navigate away from this page, the following information is no longer available.

Summary

Destination
s3://dexter-morgan-rag-bedrock

Succeeded
✓ 6 files, 5.3 MB (100.00%)

Failed
0 files, 0 B (0%)

Files and folders (6 total, 5.3 MB)

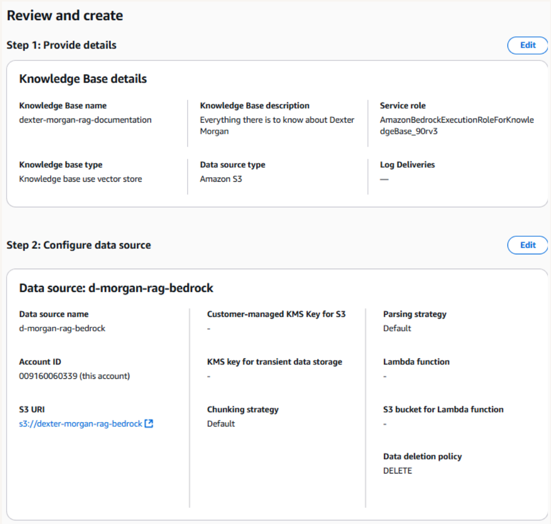
Name	Folder	Type	Size	Status	Error
Dexter Morgan - …	-	application/pdf	583.2 KB	✓ Succeeded	-
Dexter Morgan - …	-	application/pdf	709.1 KB	✓ Succeeded	-
Dexter Morgan - …	-	application/pdf	1021.6 KB	✓ Succeeded	-
Dexter Morgan - …	-	application/pdf	1.3 MB	✓ Succeeded	-
Dexter Morgan - …	-	application/pdf	927.8 KB	✓ Succeeded	-
The Code of Harr…	-	application/pdf	908.8 KB	✓ Succeeded	-

My Knowledge Base Setup

My Knowledge Base uses a vector store, which means it searchs for the most relevant information based on the User's Input Context. When I query my Knowledge Base, OpenSearch will search, analyze, and visualize large amounts of data quickly.

Embeddings are great to label and oraganise different datas that is stored in a database. The embeddings model I'm using is Titan Text Embeddings V2 because it works very well with AWS related-services.

Chunking is an effecient way to manage large amount text into small chunks for the AI Model to process efficiently. In my K-Base, Chunks are set to 300 tokens per scan (meaning 300 words).



Review and create

Step 1: Provide details

Knowledge Base details

Knowledge Base name
dexter-morgan-rag-documentation

Knowledge Base description
Everything there is to know about Dexter Morgan

Service role
AmazonBedrockExecutionRoleForKnowledgeBase_90rv3

Knowledge base type
Knowledge base use vector store

Data source type
Amazon S3

Log Deliveries
—

Step 2: Configure data source

Data source: d-morgan-rag-bedrock

Data source name
d-morgan-rag-bedrock

Account ID
009160060339 (this account)

S3 URI
s3://dexter-morgan-rag-bedrock

Customer-managed KMS Key for S3
-

KMS key for transient data storage
-

Chunking strategy
Default

Parsing strategy
Default

Lambda function
-

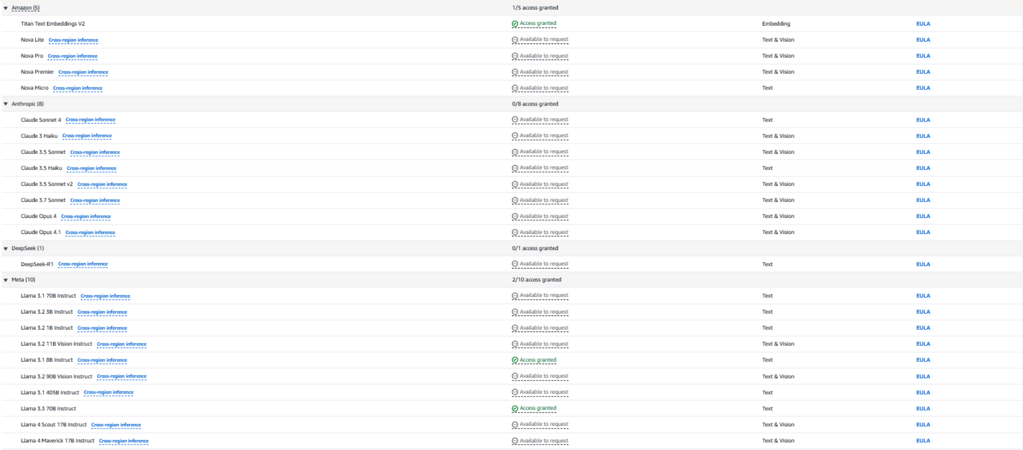
S3 bucket for Lambda function
-

Data deletion policy
DELETE

AI Models

AI models are important for my chatbot because they would convert the KBase output into human like response. Without AI models, my chatbot would only output raw results.

To get access to AI models in Bedrock, I had to go to Model Access and select the specific models required for my KBase. AWS needs explicit access so it can know how to use the AI Model

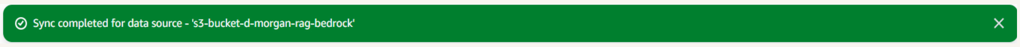


▼ Amazon (5)	1/5 access granted		
Titan Text Embeddings V2	✓ Access granted	Embedding	EULA
Nova Lite Cross-region inference	⊖ Available to request	Text & Vision	EULA
Nova Pro Cross-region inference	⊖ Available to request	Text & Vision	EULA
Nova Premier Cross-region inference	⊖ Available to request	Text & Vision	EULA
Nova Micro Cross-region inference	⊖ Available to request	Text	EULA
▼ Anthropic (8)	0/8 access granted		
Claude Sonnet 4 Cross-region inference	⊖ Available to request	Text	EULA
Claude 3 Haiku Cross-region inference	⊖ Available to request	Text & Vision	EULA
Claude 3.5 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 3.7 Sonnet Cross-region inference	⊖ Available to request	Text & Vision	EULA
Claude Opus 4 Cross-region inference	⊖ Available to request	Text & Vision	EULA
Claude Opus 4.1 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 (10)	2/10 access granted		
Llama 3.1 70B 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 1B 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 3.1 8B Instruct Cross-region inference	✓ Access granted	Text	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 3.3 70B Instruct	✓ Access granted	Text	EULA
Llama 4 Scout 17B Instruct Cross-region inference	⊖ Available to request	Text & Vision	EULA
Llama 4 Maverick 17B Instruct Cross-region inference	⊖ Available to request	Text & Vision	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 data hasn't actually moved from S3 into your Knowledge Base yet.

The sync process involves three steps: 1. Ingesting 2. Processing 3. Storing



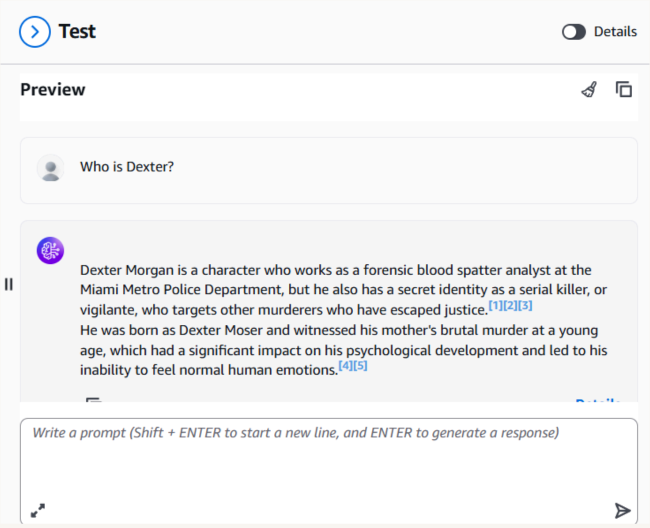
Sync completed for data source - 's3-bucket-d-morgan-rag-bedrock'

Testing My Chatbot

'I initially tried to test my chatbot using Llama 3.1 8B as the AI model, but I had to switch to Llama 3.3 70B because it supports on-demand inference!

When I asked about topics unrelated to my data (like Batman), my chatbot replied 'There is no information about Batman in the provided search results.' This proves that the model focus only on what it has been given and had been taught.

You can also turn off the Generate Responses setting to recieve RAW results from your Data.



Who is Dexter?

Dexter Morgan is a character who works as a forensic blood spatter analyst at the Miami Metro Police Department, but he also has a secret identity as a serial killer, or vigilante, who targets other murderers who have escaped justice.[1][2][3]
He was born as Dexter Moser and witnessed his mother's brutal murder at a young age, which had a significant impact on his psychological development and led to his inability to feel normal human emotions.[4][5]

Write a prompt (Shift + ENTER to start a new line, and ENTER to generate a response)

The place to learn & showcase your skills

Check out nextwork.org for more projects