



Retrieval-Augmented Generation (RAG)

Overview and Interactive Demos

Jeff Maruschek

Amazon Web Services

Solutions Architect – DoD – Air Force

Introduction



Introduction



djmoose on Mixcloud:



Agenda

What is RAG?

What are Embeddings?

Amazon Bedrock ChatBot Demo

Amazon Bedrock Knowledge Bases Demo

What book genre is The Three-Body Problem?

What instrument does Joshua Bell play?

What instrument does the musician Joshua Bell play?

Who is Joshua Bell?

WIKIPEDIA

The Free Encyclopedia

Search Wikipedia

Search

Contents

hide

(Top)

Early life and education

Career

Washington Post experiment

Personal life

Selected discography

Soundtrack albums

References

External links

Joshua Bell

28 languages

Article

Talk

Read

Edit

View history

Tools

From Wikipedia, the free encyclopedia

For other people named Joshua Bell, see [Joshua Bell \(disambiguation\)](#).

Joshua David Bell (born December 9, 1967) is an American [violinist](#) and conductor. He plays the [Gibson Stradivarius](#).

Early life and education [[edit](#)]

Bell was born in [Bloomington, Indiana](#), to Shirley Bell, a therapist, and [Alan P. Bell](#), a psychologist, professor emeritus at [Indiana University](#) (IU), and former [Kinsey researcher](#).^{[1][2]} His father is of Scottish descent and his mother is [Jewish](#) (her father was born in [Mandatory Palestine](#) and her mother was from [Minsk](#)).^[3]

Bell began playing the violin at age four after his mother discovered that he had taken rubber bands from around the house and stretched them across the handles of his nine dresser drawers to pluck

Joshua Bell



Bell in 2010



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

9

Lack of Context

Human

in one sentence where in new york city does joshua bell live?

Assistant (bedrock-claude-v2)

Unfortunately, I do not have enough information to determine in one sentence where in New York City the violinist Joshua Bell lives. Public records of celebrities' private home addresses are generally not available.

Human

On October 5, 2019, Bell married opera singer Larisa Martinez at their home in Mount Kisco, New York.[38][39] They also live in the Gramercy Park neighborhood of Manhattan [40]

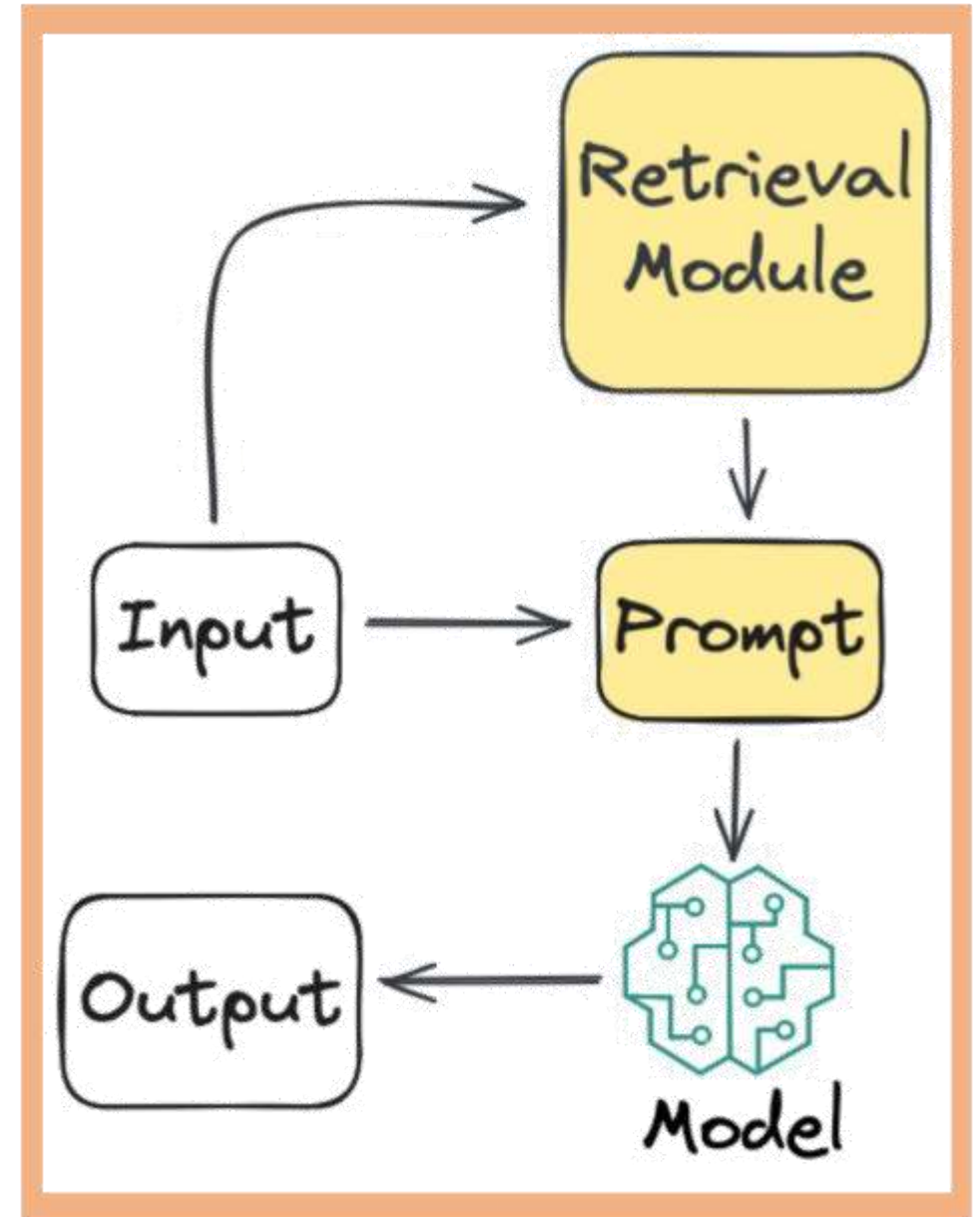
in one sentence, where in new york city does joshua bell live?

Assistant (bedrock-claude-v2)

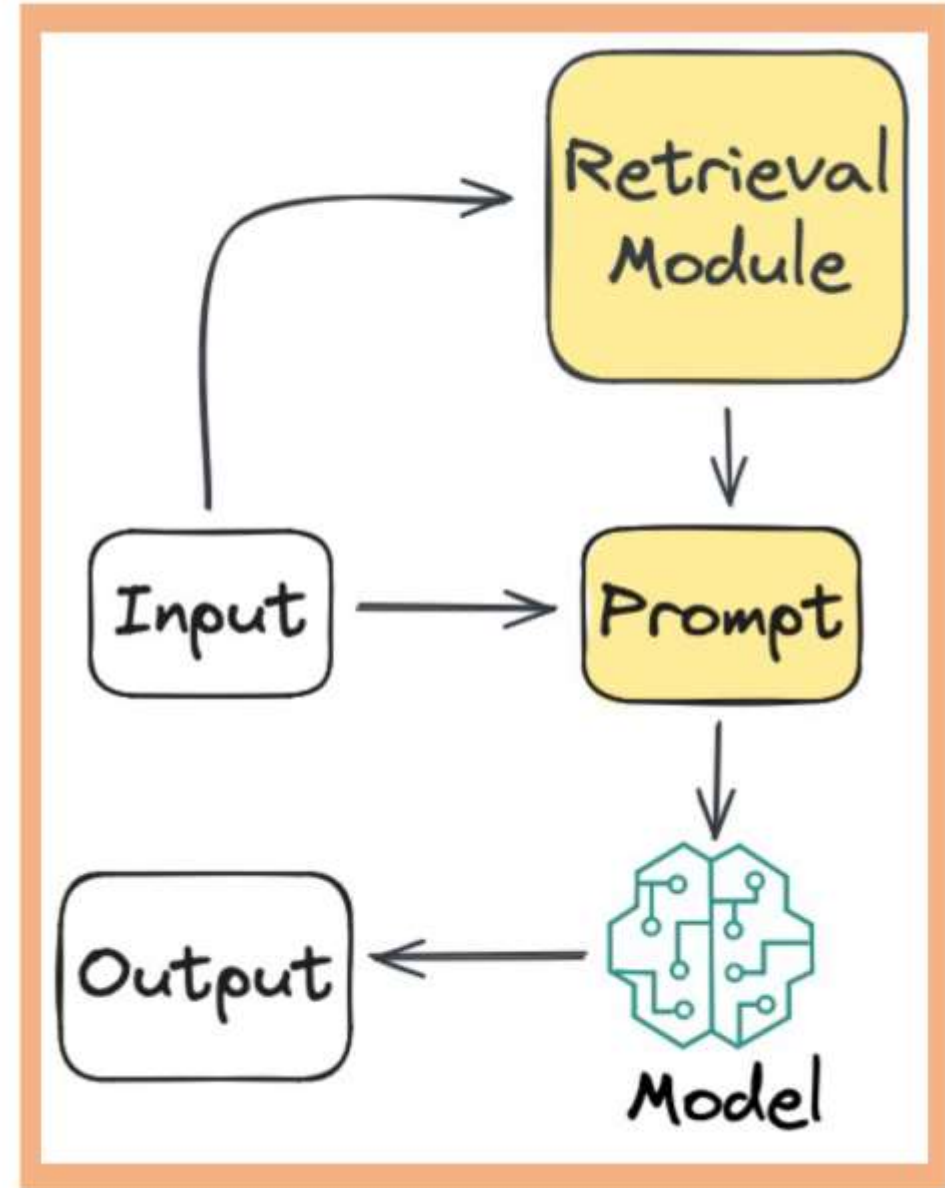
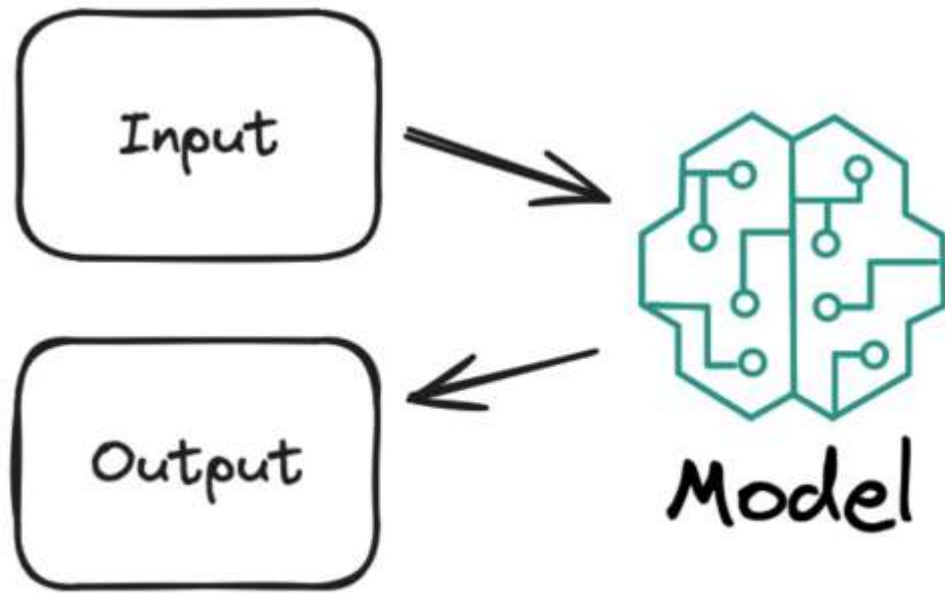
Joshua Bell lives in the Gramercy Park neighborhood of Manhattan, New York City.

What is RAG?

- Retrieval
 - Fetches the relevant content from the external knowledge base or data sources based on a user query
- Augmentation
 - Adding the retrieved relevant context to the user prompt, which goes as an input to the foundation model
- Generation
 - Response from the foundation model based on the augmented prompt.

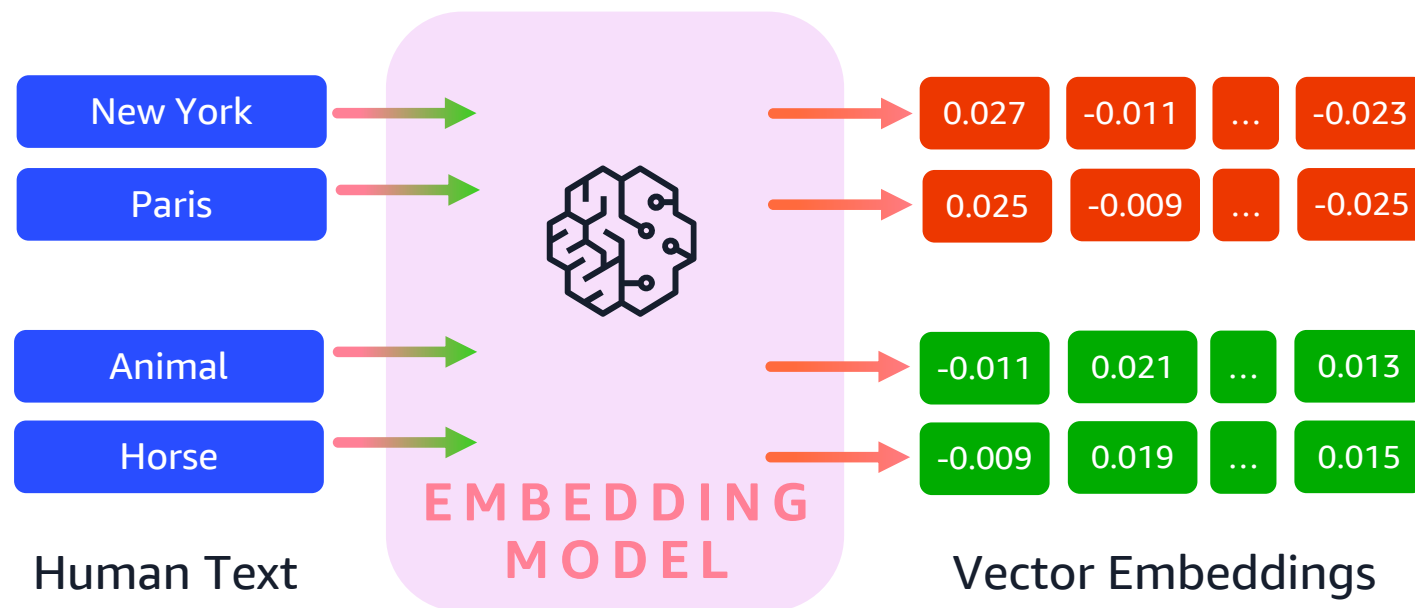


Models and Context

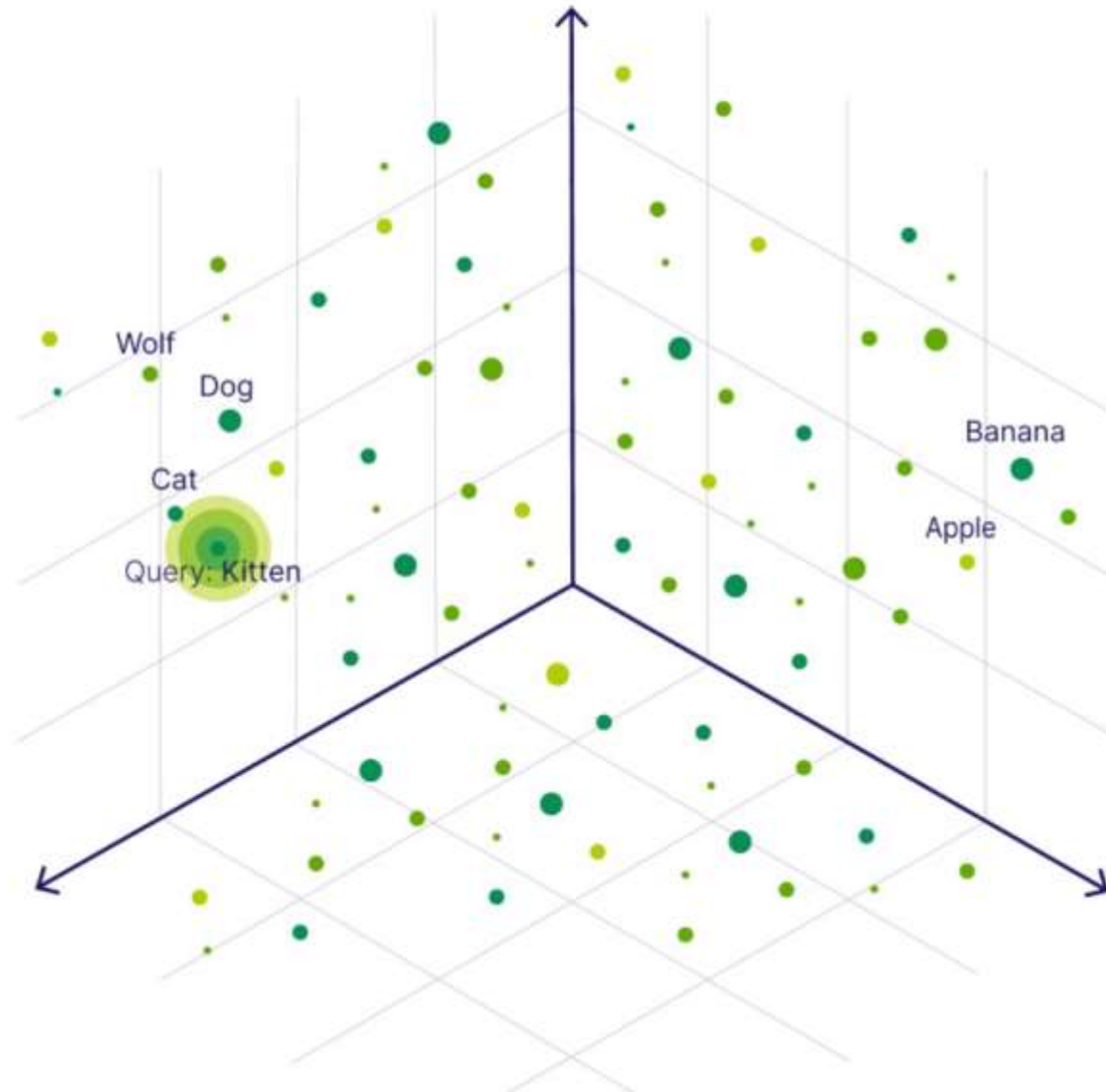


What are embeddings?

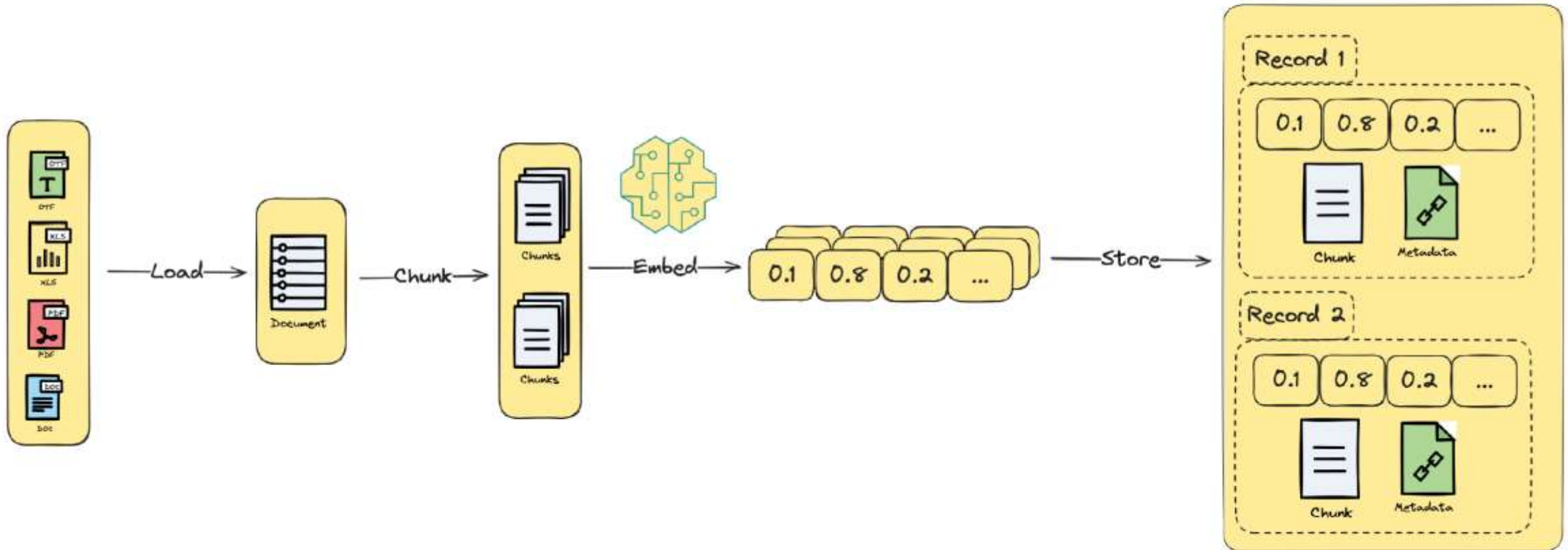
- Numerical representation of text (vectors) that captures semantics and relationships between words.
- Embedding models capture features and nuances of the text.
- Rich embeddings can be used to compare text similarity.
- Multilingual Text Embeddings can identify meaning in different languages.



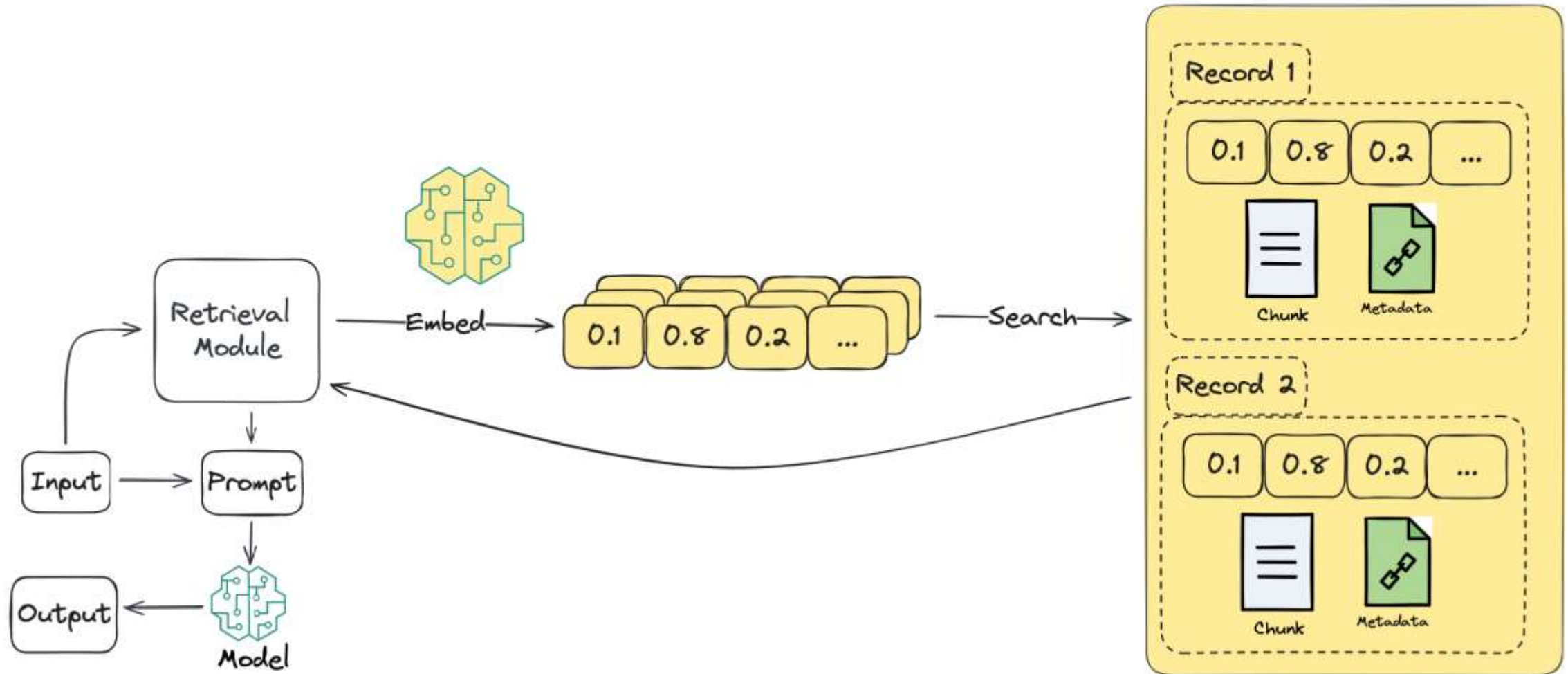
Embedding Similarity



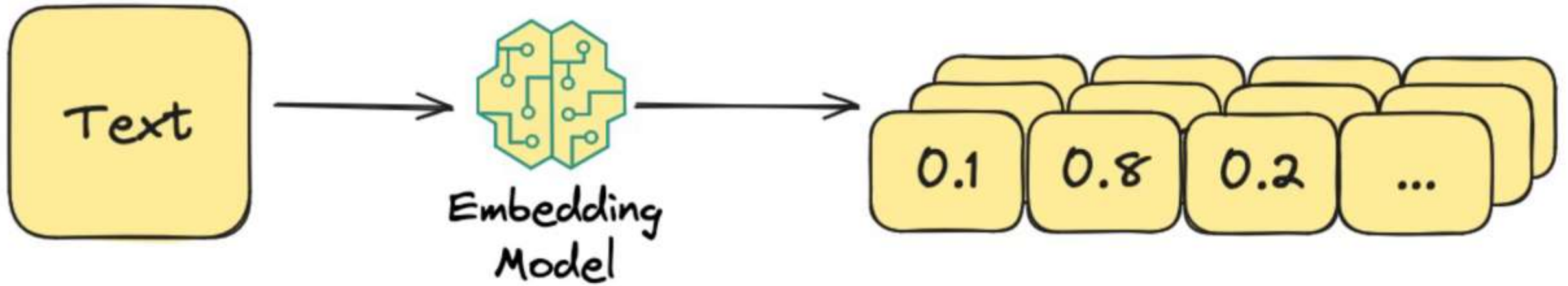
RAG Pre-processing



RAG with Vector Search



Semantic Embeddings



Amazon Bedrock ChatBot Demo

<https://github.com/aws-samples/bedrock-claude-chat>

[README](#) [Code of conduct](#) [MIT-0 license](#) [Security](#)

Bedrock Claude Chat

CDK tests passing

Tip

Multi-modal chat by [Claude v3 \(Haiku, Sonnet\)](#) is available for now. See [Release](#) for the detail.


Warning

The current version (v0.4.x) has no compatibility with ex version (~ v0.3.0) due to the change of DynamoDB table schema. Please note that UPDATE (i.e. `cdk deploy`) FROM EX VERSION TO v0.4.x WILL DESTROY ALL OF EXISTING CONVERSATIONS.

This repository is a sample chatbot using the Anthropic company's LLM [Claude](#), one of the foundational models provided by [Amazon Bedrock](#) for generative AI.

Basic Conversation


Not only text but also images are available with [Anthropic's Claude 3](#). Currently we support `Haiku` and `Sonnet`.



Packages


No packages published

Contributors 20



[+ 6 contributors](#)

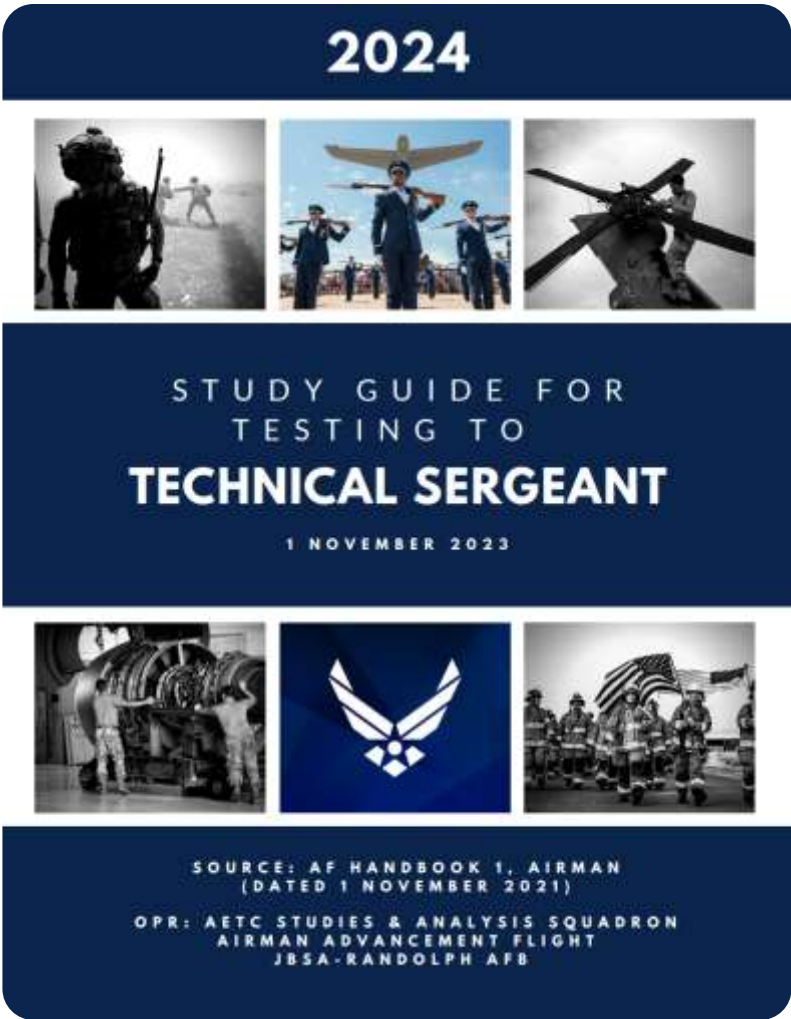
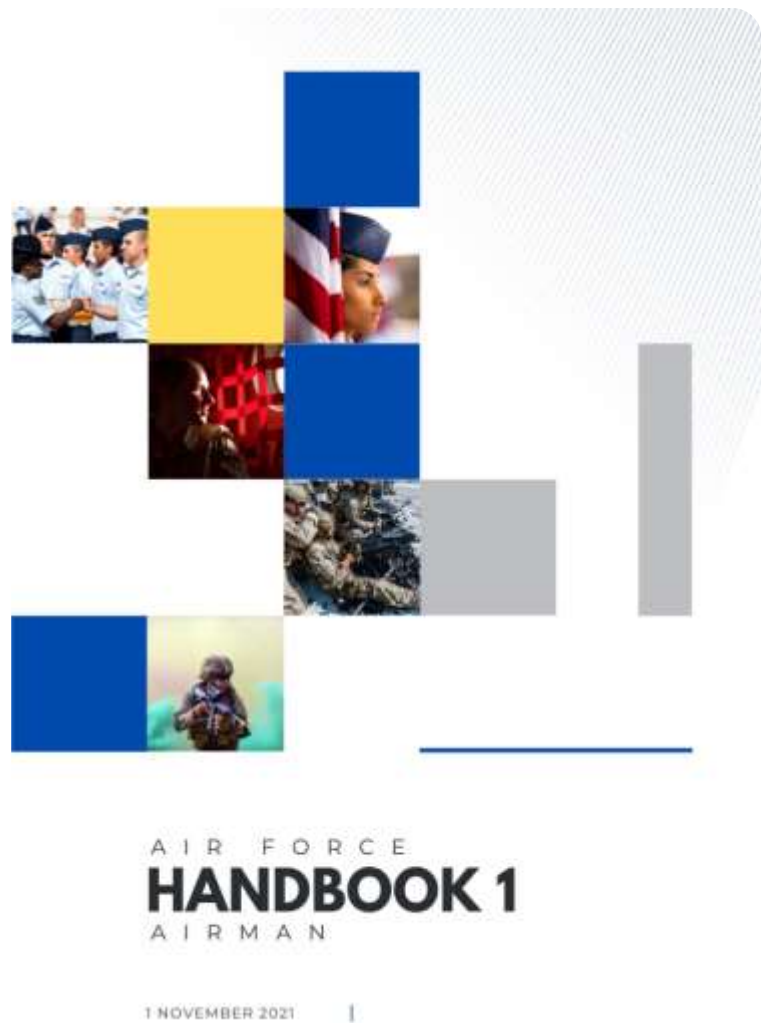
Languages



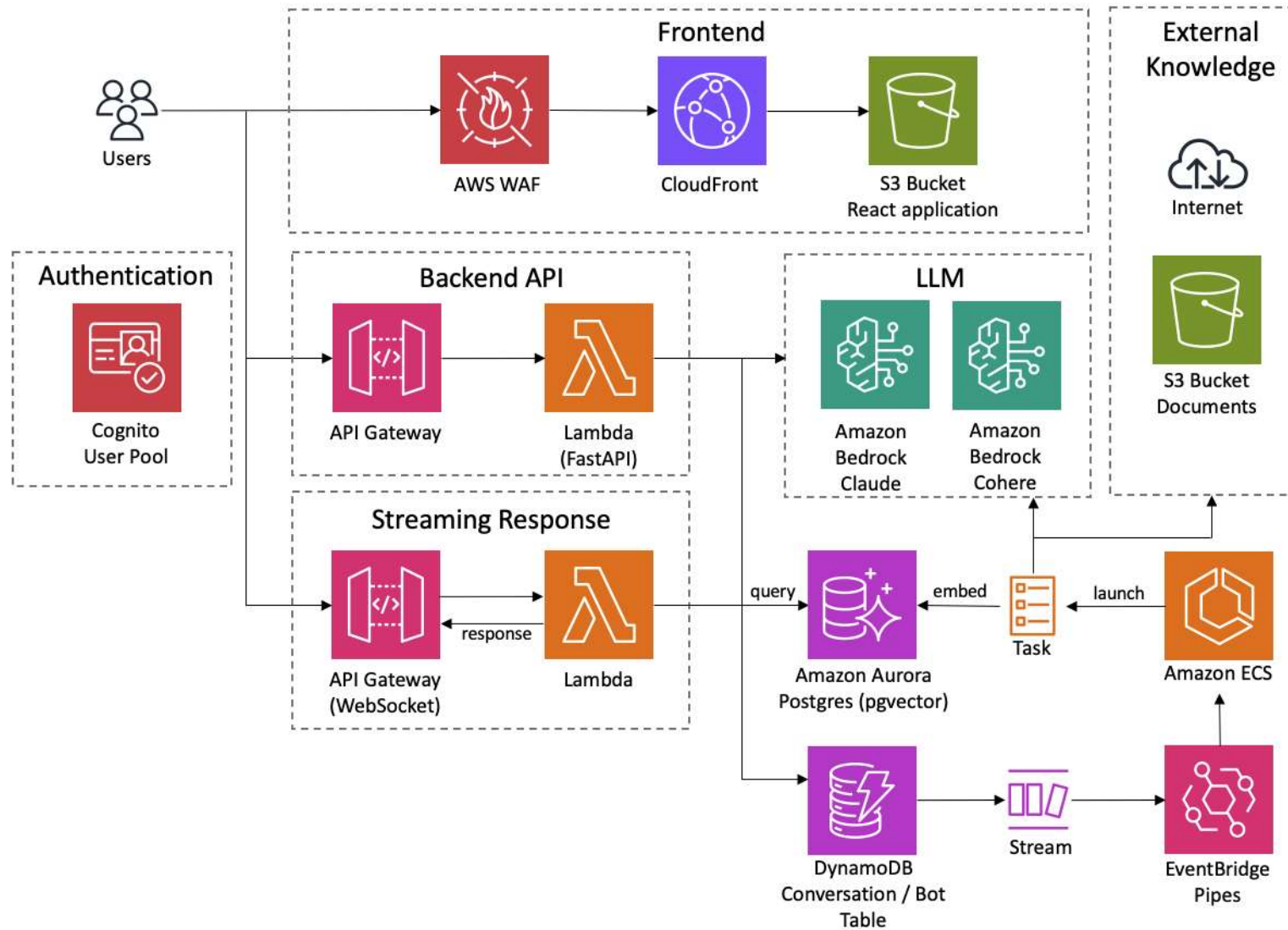
Language	Percentage
TypeScript	51.2%
Python	45.5%
JavaScript	1.3%
HTML	1.1%
Shell	0.4%
Dockerfile	0.4%
CSS	0.1%



Documents Used



Architecture



However, when it comes to implementing RAG, there are challenges...



Managing
multiple data
sources



Creating vector
embeddings for large
volumes of data



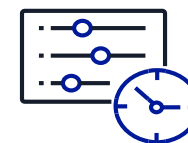
Incremental
updates to vector
store



Coding effort



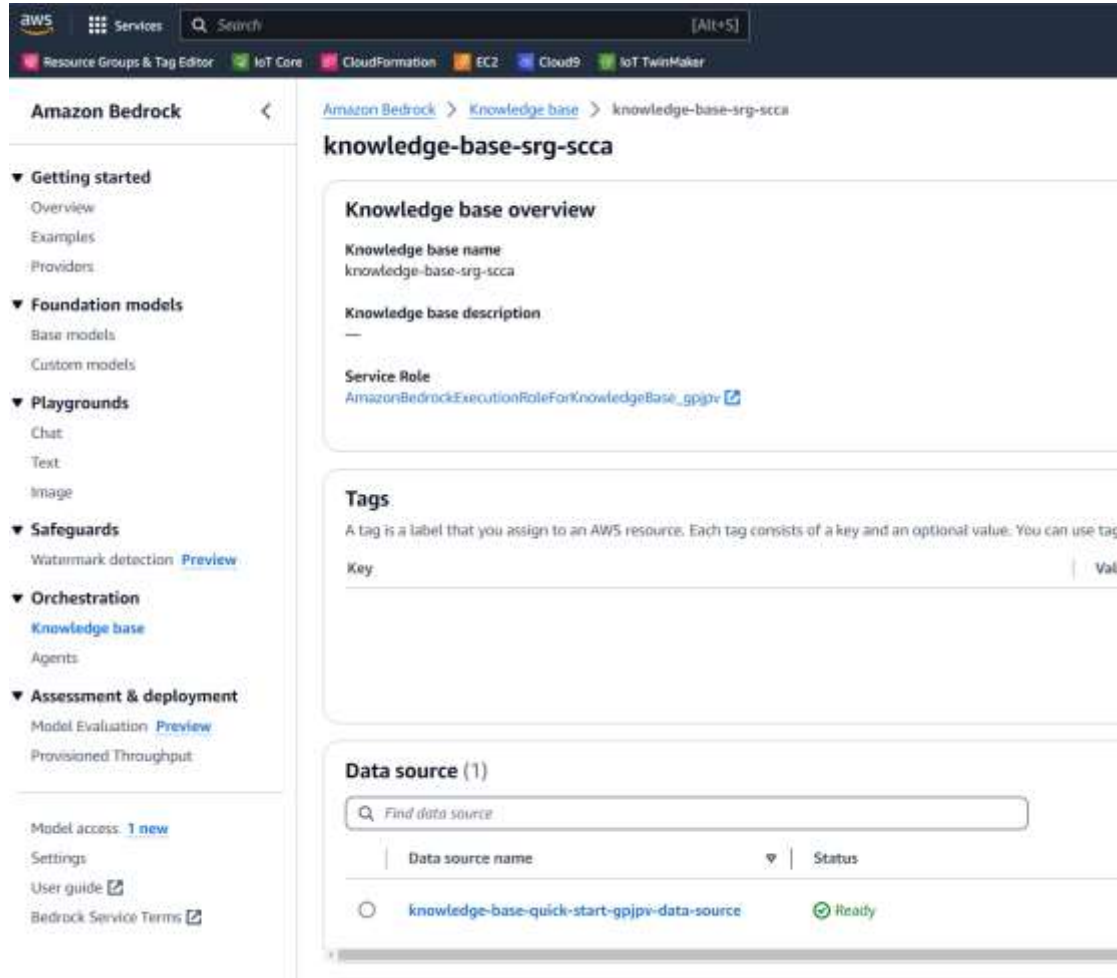
Scaling retrieval
mechanism



Orchestration

Amazon Bedrock Knowledge Bases Demo

<https://aws.amazon.com/bedrock/knowledge-bases/>



The screenshot displays the Amazon Bedrock console interface. On the left, a navigation sidebar lists various sections: Getting started, Foundation models, Playgrounds, Safeguards, Orchestration, and Assessment & deployment. The main content area shows the 'knowledge-base-srg-scca' knowledge base overview. This overview includes the knowledge base name, a description, and the service role. Below this, there is a 'Tags' section with a description of what a tag is and a table for managing tags. At the bottom, the 'Data source (1)' section shows a table with one data source named 'knowledge-base-quick-start-gpjpv-data-source' in a 'Ready' state.

Amazon Bedrock

Knowledge base overview

Knowledge base name
knowledge-base-srg-scca

Knowledge base description
—

Service Role
[AmazonBedrockExecutionRoleForKnowledgeBase_gpjpv](#)

Tags

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tag

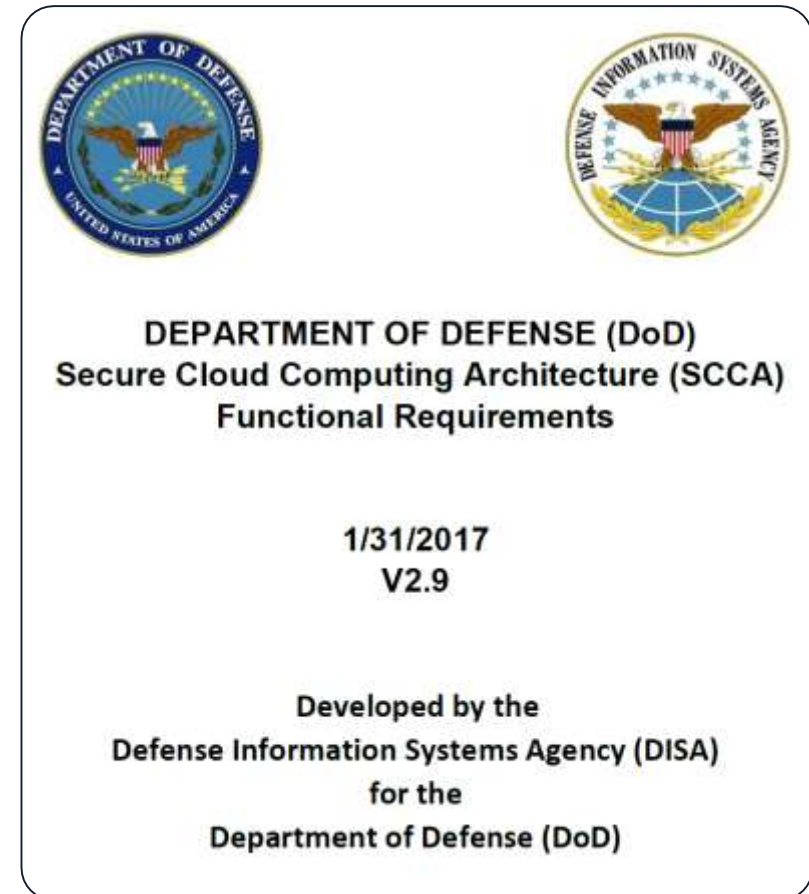
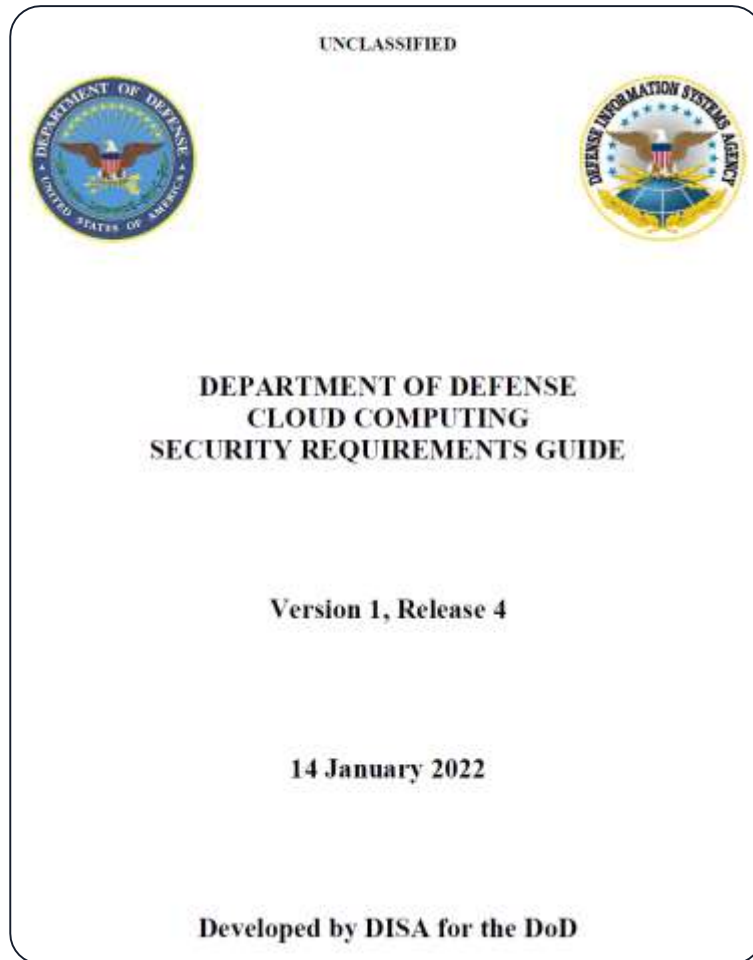
Key	Value
-----	-------

Data source (1)

Data source name	Status
knowledge-base-quick-start-gpjpv-data-source	Ready



Documents Used



Thank you!

Jeff Maruschek

Solutions Architect – DoD – Air Force

jefmarus@amazon.com

