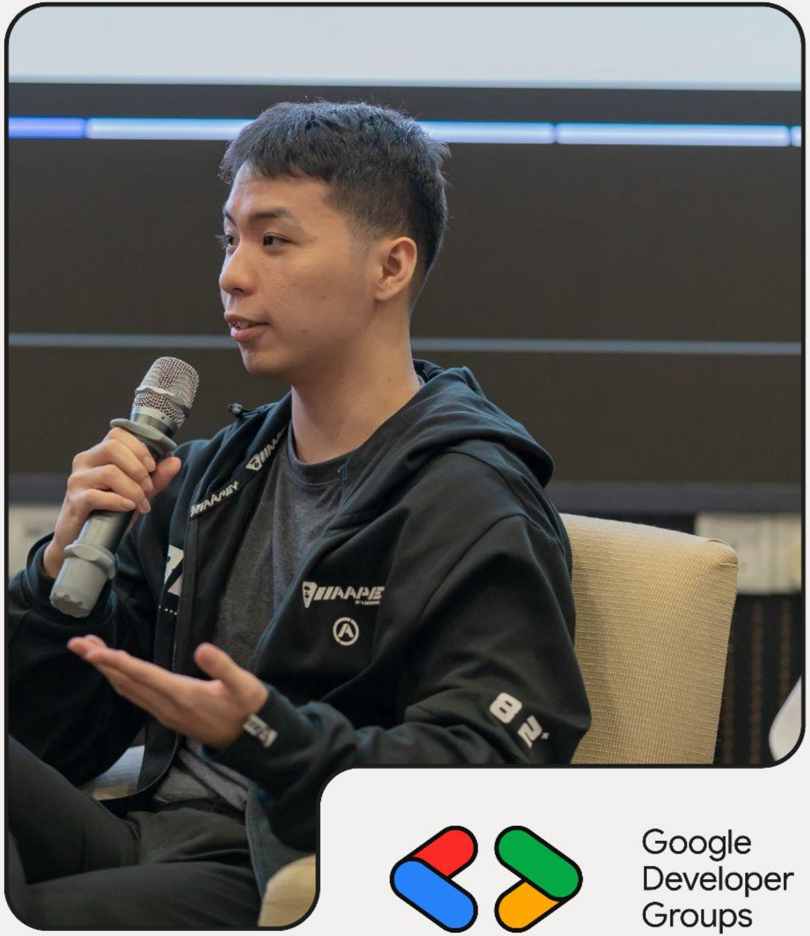
DevFest

GDG Cloud KL

Building RAG with MongoDB & Gemma

Lai Kai Yong (Vandyck)



Google Developer Groups

Agenda

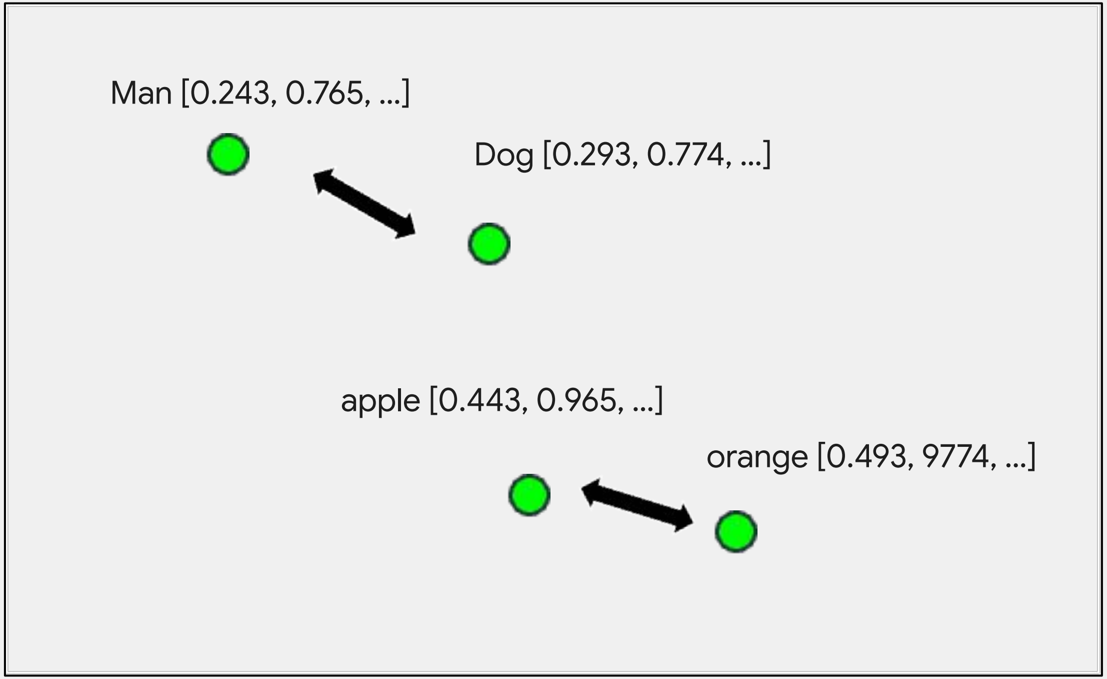
1. RAG Concepts
2. Why MongoDB
3. Models in Google
4. Why Google Models
5. Workshop

Vectors in ML are **lists of numbers** that represent attributes.

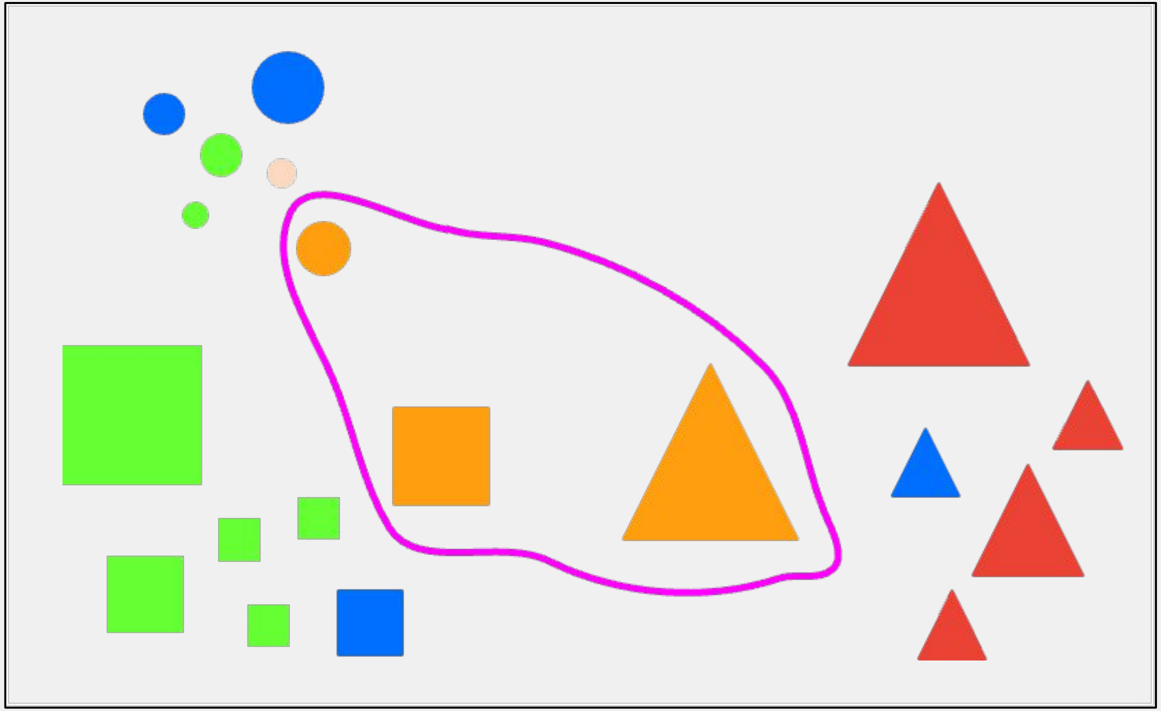
[0.743, 0.720, -0.325, 0.195, 0.835, -0.945, ...]

Each number represents a feature
(or a property of a data object)

Similar vectors plotted in space will be **near one another**

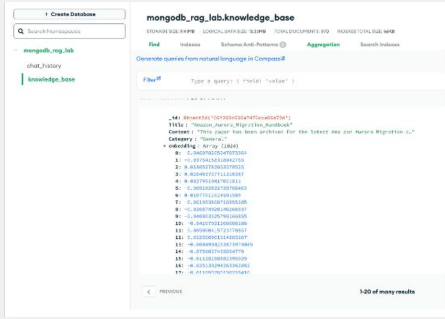


We can then take **search queries** and use **algorithms** to find **clusters** in high-dimensional space



Why RAG in **MongoDB**

Embedding as Attributes



+ Create Database

Search Namespaces

mongodb_rag_lab
- chat_history
- knowledge_base

mongodb_rag_lab.knowledge_base

STORAGE SIZE: 9.41MB LOGICAL DATA SIZE: 18.33MB TOTAL DOCUMENTS: 970 INDEXES TOTAL SIZE: 56KB

Find | Indexes | Schema Anti-Patterns | Aggregation | Search Indexes

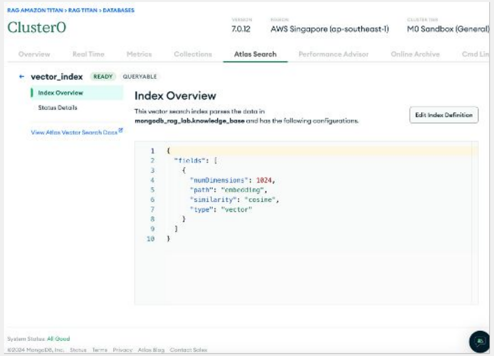
Generate queries from natural language in Compass

Filter: Type a query: { field: 'value' }

```
_id: ObjectId('66f2b0c61ca747ccca06a70d')
Title : "Amazon_Aurora_Migration_Handbook"
Content : "This paper has been archived For the latest Ama zon Aurora Migration c…"
Category : "General"
▾ embedding : Array (1024)
   0: 0.046978205047073364
   1: -0.03754158318842755
   2: 0.018652762863378525
   3: 0.026463737711318387
   4: 0.00279510417821811
   5: -0.009282822720786463
   6: 0.01077012624591589
   7: -0.001953608716055105
   8: -0.026874928146266937
   9: -0.046093325790166856
   10: -0.04207021268608186
   11: 0.005080415723778657
   12: 0.012306001014263167
   13: -0.0004994253873974865
   14: -0.0756037435054779
   15: -0.01128298582395029
   16: -0.025139294263362865
   17: -0.011093782190229416
```

PREVIOUS   1-20 of many results

Vector Search Index



RAG AMAZON TITAN > RAG TITAN > DATABASES

Cluster0

VERSION 7.0.12 | REGION AWS Singapore (ap-southeast-1) | CLUSTER TIER M0 Sandbox (General)

Overview | Real Time | Metrics | Collections | Atlas Search | Performance Advisor | Online Archive | Cmd Lin

← vector_index READY QUERYABLE

- Index Overview
- Status Details

View Atlas Vector Search Docs

Index Overview

This vector search index parses the data in **mongodb_rag_lab.knowledge_base** and has the following configurations.

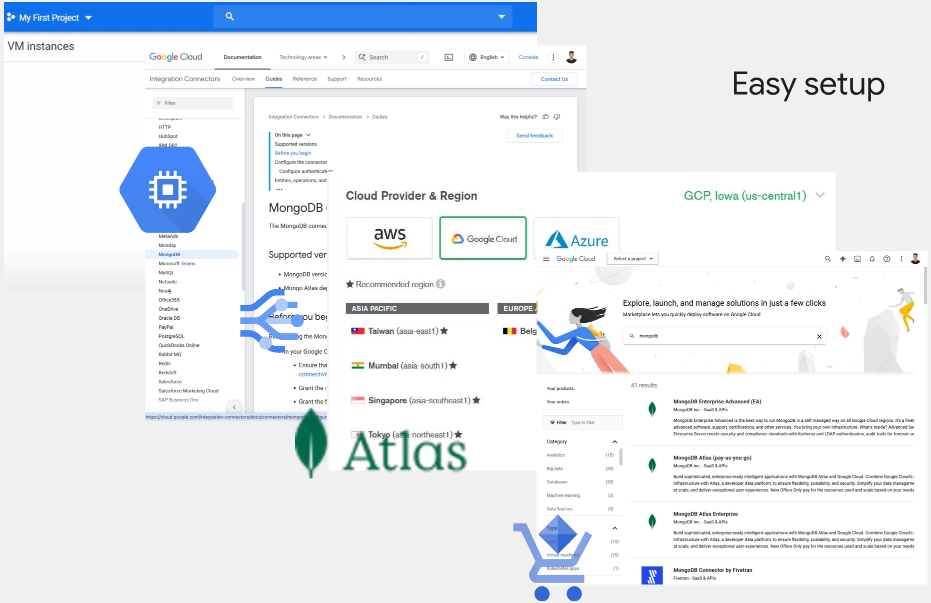
Edit Index Definition

```
{
  "fields": [
    {
      "numDimensions": 1024,
      "path": "embedding",
      "similarity": "cosine",
      "type": "vector"
    }
  ]
}
```

System Status: All Good

©2024 MongoDB, Inc. Status Terms Privacy Atlas Blog Contact Sales

Easy setup



My First Project

VM instances

Google Cloud | Documentation | Technology areas | Search | English | Console

Integration Connectors | Overview | Guides | Reference | Support | Resources | Contact Us

Filter

HTTP, HubSpot, IBM DB2, MetaAds, Monday, MongoDB, Microsoft Teams, MySQL, Netsuite, Neo4j, Office365, OneDrive, Oracle DB, PayPal, PostgreSQL, QuickBooks Online, Rabbit MQ, Redis, Redshift, Salesforce, Salesforce Marketing Cloud, SAP Business One

Integration Connectors > Documentation > Guides   Was this helpful?   Send feedback

On this page: Supported versions; Before you begin; Configure the connector; Configure authentication; Entities, operations, and …

MongoDB …

The MongoDB connec…

Supported ver…
- MongoDB versio…
- Mongo Atlas dep…

Before you beg…

…ng the Mong…
…in your Google C…
- Ensure tha… connectivi…
- Grant the …
- Grant the f…

https://cloud.google.com/integration-connectors/docs/connectors/mongod…figur…

Cloud Provider & Region   GCP, Iowa (us-central1)

aws | Google Cloud | Azure

★ Recommended region

ASIA PACIFIC
- Taiwan (asia-east1) ★
- Mumbai (asia-south1) ★
- Singapore (asia-southeast1) ★
- Tokyo (asia-northeast1) ★

EUROPE
- Belg…

Atlas

Google Cloud | Select a project

Explore, launch, and manage solutions in just a few clicks

Marketplace lets you quickly deploy software on Google Cloud

mongodb

Your products
Your orders

Filter Type to filter

Category
- Analytics (15)
- Big data (20)
- Databases (28)
- Machine learning (2)
- Data Sources (3)

Type
- SaaS (15)
- Virtual machines (25)
- Kubernetes apps (1)

41 results

MongoDB Enterprise Advanced (EA)
MongoDB Inc. - SaaS & APIs
MongoDB Enterprise Advanced is the best way to run MongoDB in a self managed way on all Google Cloud regions. It's a finely advanced software, support, certifications, and other services. You bring your own infrastructure. What's Inside? Advanced Se… Enterprise Server meets security and compliance standards with Kerberos and LDAP authentication, audit trails for forensic ar…

MongoDB Atlas (pay-as-you-go)
MongoDB Inc. - SaaS & APIs
Build sophisticated, enterprise-ready intelligent applications with MongoDB Atlas and Google Cloud. Combine Google Cloud's … infrastructure with Atlas, a developer data platform, to ensure flexibility, scalability, and security. Simplify your data manageme… at scale, and deliver exceptional user experiences. New Offers Only pay for the resources used and scale based on your needs

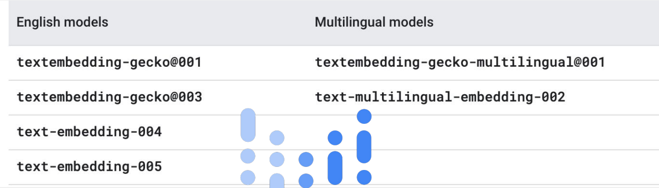
MongoDB Atlas Enterprise
MongoDB Inc. - SaaS & APIs
Build sophisticated, enterprise-ready intelligent applications with MongoDB Atlas and Google Cloud. Combine Google Cloud's … infrastructure with Atlas, a developer data platform, to ensure flexibility, scalability, and security. Simplify your data manageme… at scale, and deliver exceptional user experiences. New Offers Only pay for the resources used and scale based on your needs

MongoDB Connector by Fivetran
Fivetran - SaaS & APIs

Chunking

Vertex AI Agent Builder

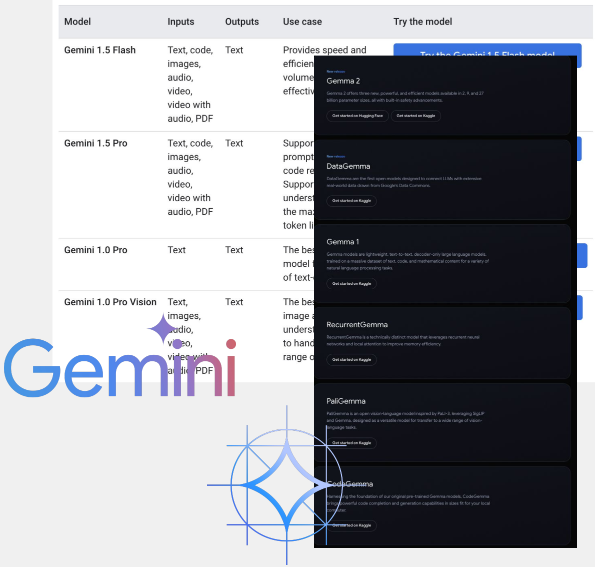
Embedding



English models	Multilingual models
`textembedding-gecko@001`	`textembedding-gecko-multilingual@001`
`textembedding-gecko@003`	`text-multilingual-embedding-002`
`text-embedding-004`	
`text-embedding-005`	

Vertex AI API

Generation



Model	Inputs	Outputs	Use case	Try the model
Gemini 1.5 Flash	Text, code, images, audio, video, video with audio, PDF	Text	Provides speed and efficien volume effectiv	Try the Gemini 1.5 Flash model
Gemini 1.5 Pro	Text, code, images, audio, video, video with audio, PDF	Text	Suppor prompt code re Suppor underst the max token li	
Gemini 1.0 Pro	Text	Text	The bes model f of text-	
Gemini 1.0 Pro Vision	Text, images, audio, video, video with audio, PDF	Text	The bes image a underst to hand range o	

New release

Gemma 2

Gemma 2 offers three new, powerful, and efficient models available in 2, 9, and 27 billion parameter sizes, all with built-in safety advancements.

New release

DataGemma

DataGemma are the first open models designed to connect LLMs with extensive real-world data drawn from Google's Data Commons.

Gemma 1

Gemma models are lightweight, text-to-text, decoder-only large language models, trained on a massive dataset of text, code, and mathematical content for a variety of natural language processing tasks.

RecurrentGemma

RecurrentGemma is a technically distinct model that leverages recurrent neural networks and local attention to improve memory efficiency.

PaliGemma

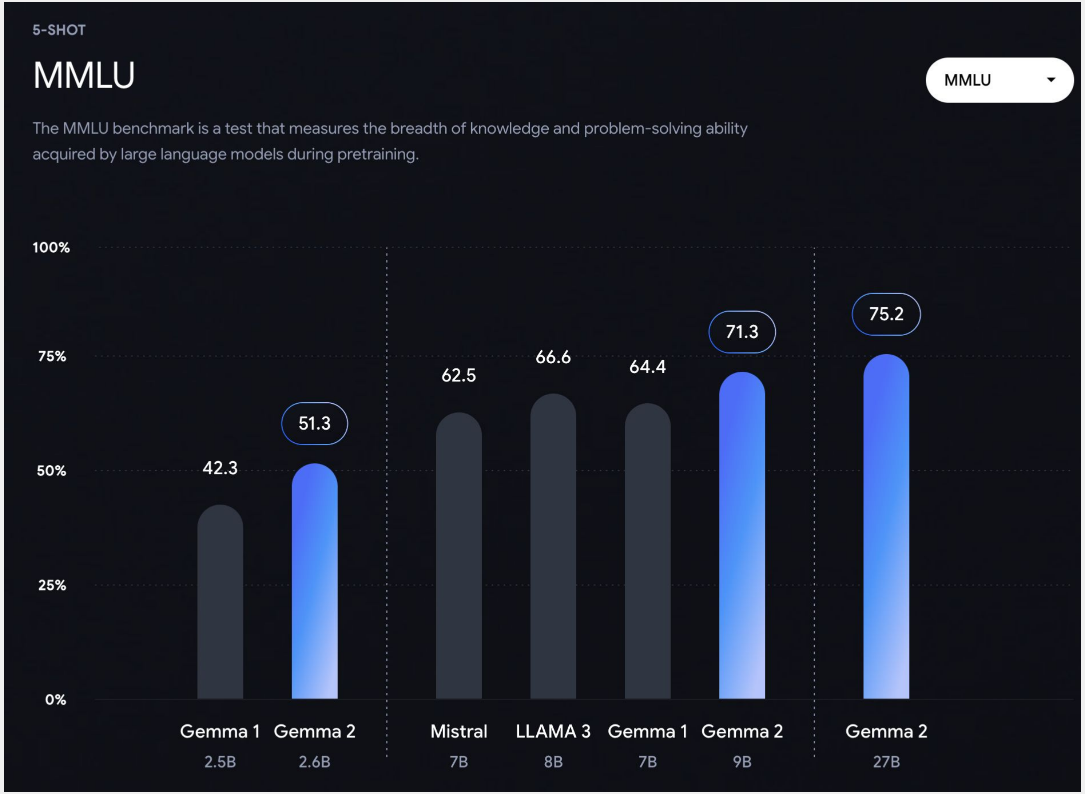
PaliGemma is an open vision-language model inspired by PaLI-3, leveraging SigLIP and Gemma, designed as a versatile model for transfer to a wide range of vision-language tasks.

CodeGemma

Harnessing the foundation of our original pre-trained Gemma models, CodeGemma brings powerful code completion and generation capabilities in sizes fit for your local computer.

Gemini / Gemma

Why **Gemma**



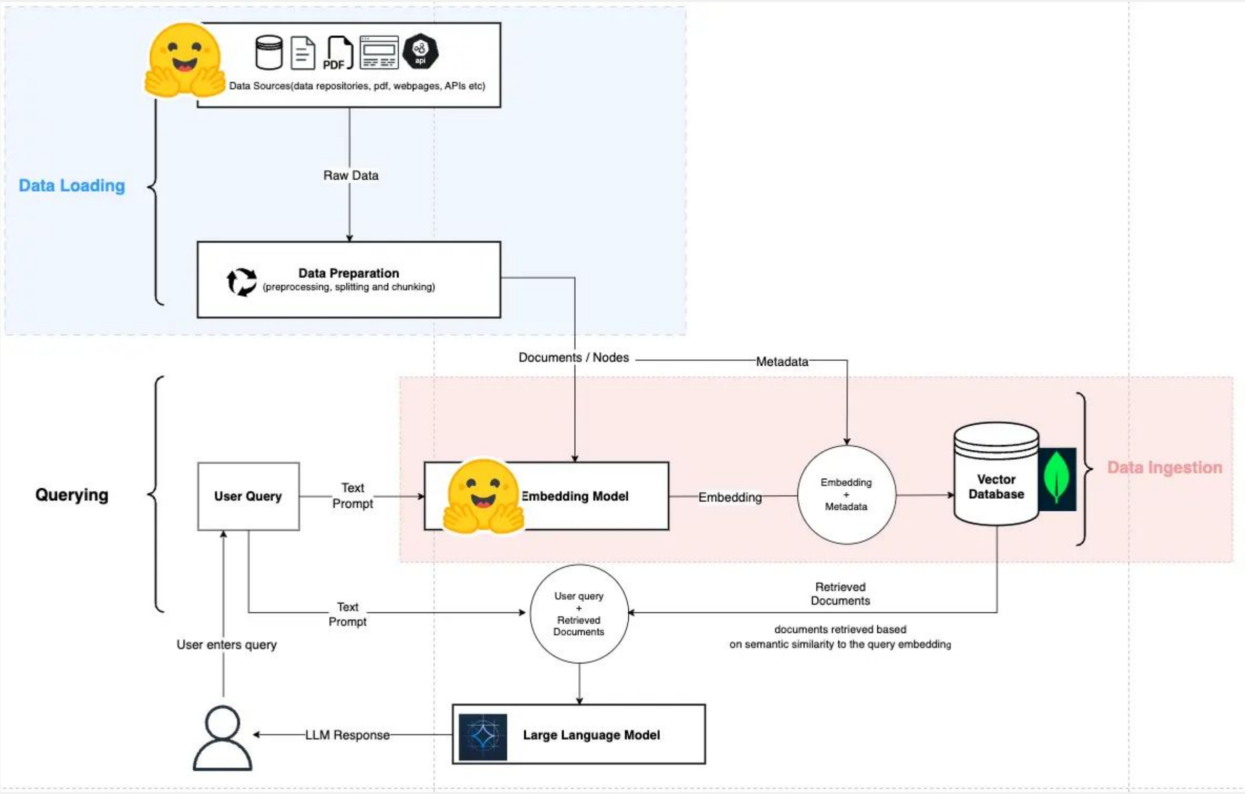
5-SHOT

MMLU

The MMLU benchmark is a test that measures the breadth of knowledge and problem-solving ability acquired by large language models during pretraining.

Context Length	8192
Strength	- Sliding window attention (low memory low time) - Knowledge Distillation (Learning from sensei model) - Model Merging
General Performance	75.2 % MMLU Benchmark

RAG Flow


Data Loading
Data Sources(data repositories, pdf, webpages, APIs etc)
PDF
api
Raw Data
Data Preparation
(preprocessing, splitting and chunking)
Documents / Nodes
Metadata
Querying
User Query
Text Prompt
Embedding Model
Embedding
Embedding + Metadata
Vector Database
Data Ingestion
Retrieved Documents
documents retrieved based on semantic similarity to the query embedding
Text Prompt
User query + Retrieved Documents
User enters query
LLM Response
Large Language Model