One Stop Health

Arda Celik, Wesley Burnawan
Ontario Tech University
19 November 2024

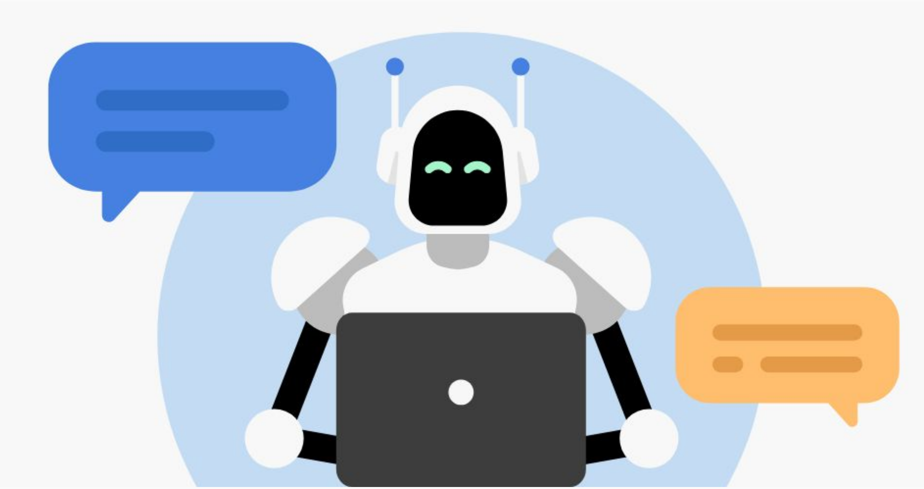
Motivation

- 22% of Canadian adults don’t have a family doctor
- 66% described Canadian healthcare system as “long wait”
- Only 26% of Canadians could get same day or next day appointment
- Google is not always reliable
- Research question: How can LLMs be optimized to provide accurate and context-aware health-related responses?

Background

1. Chatbot with prefixed query and response
2. Chatbot with NLP
3. Chatbot with LLM and sensors

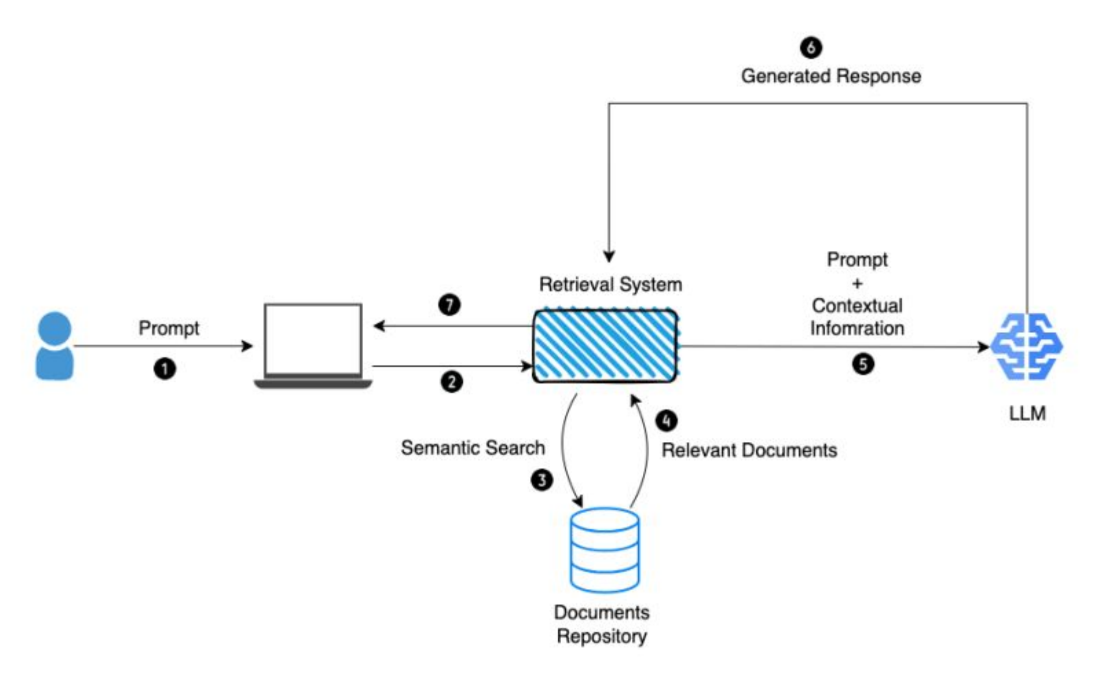
Our version: RAG Chatbot



Design and Implementation

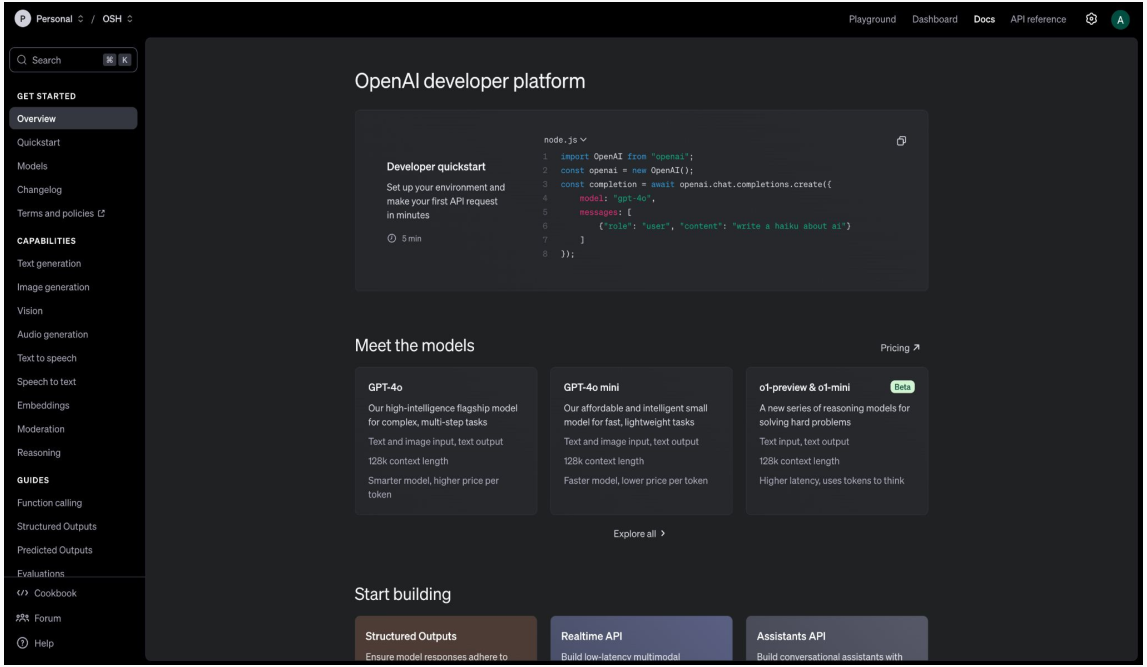
App components:

- UI
- API
- LLM
- MongoDB
- Pinecone (vector database)
- CI/CD
- Google Cloud



LLM Integration OpenAI

- OpenAI's API platform
  - Models: GPT-4o, GPT-4o-mini, text-embedding-3-small
  - Capable models
  - Easy integration



OpenAI developer platform

Developer quickstart

Set up your environment and make your first API request in minutes

5 min

node.js

```
import OpenAI from "openai";
const openai = new OpenAI();
const completion = await openai.chat.completions.create({
    model: "gpt-4o",
    messages: [
        {"role": "user", "content": "write a haiku about ai"}
    ]
});
```

Meet the models

Pricing ↗

GPT-4o

Our high-intelligence flagship model for complex, multi-step tasks

Text and image input, text output

128k context length

Smarter model, higher price per token

GPT-4o mini

Our affordable and intelligent small model for fast, lightweight tasks

Text and image input, text output

128k context length

Faster model, lower price per token

o1-preview & o1-mini Beta

A new series of reasoning models for solving hard problems

Text input, text output

128k context length

Higher latency, uses tokens to think

Explore all >

Start building

Structured Outputs

Ensure model responses adhere to

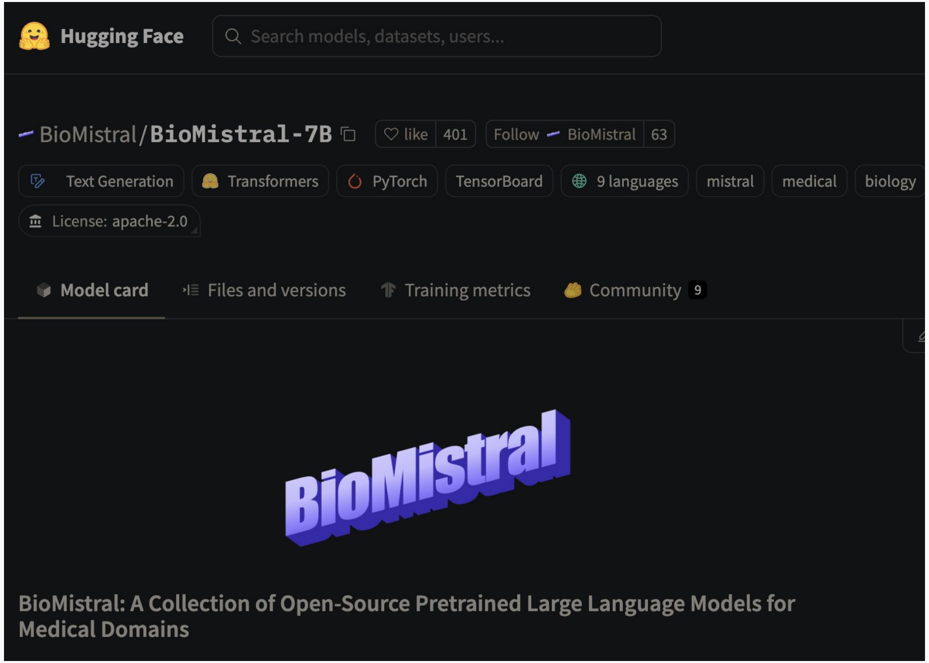
Realtime API

Build low-latency multimodal

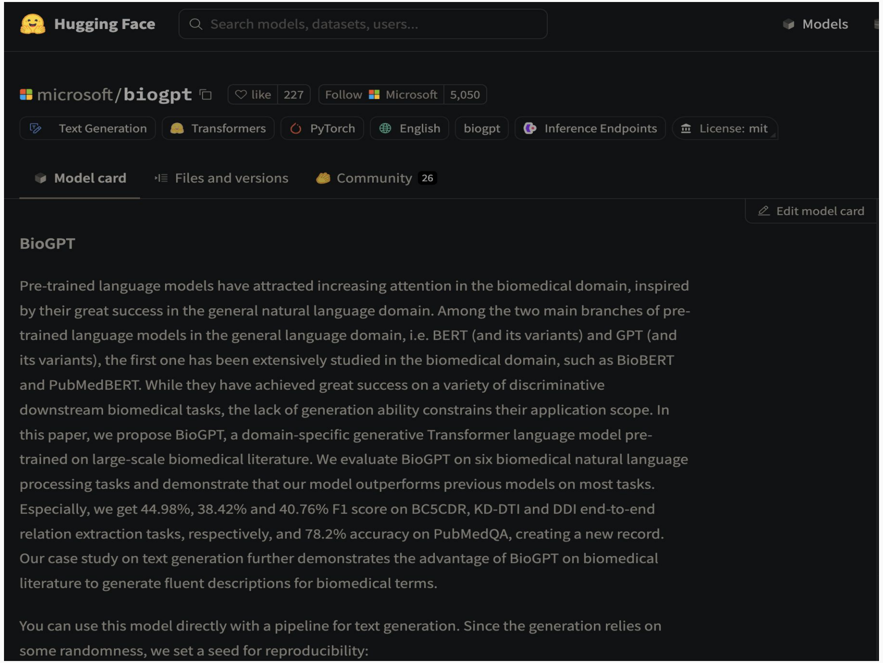
Assistants API

Build conversational assistants with

LLM Alternatives



BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains



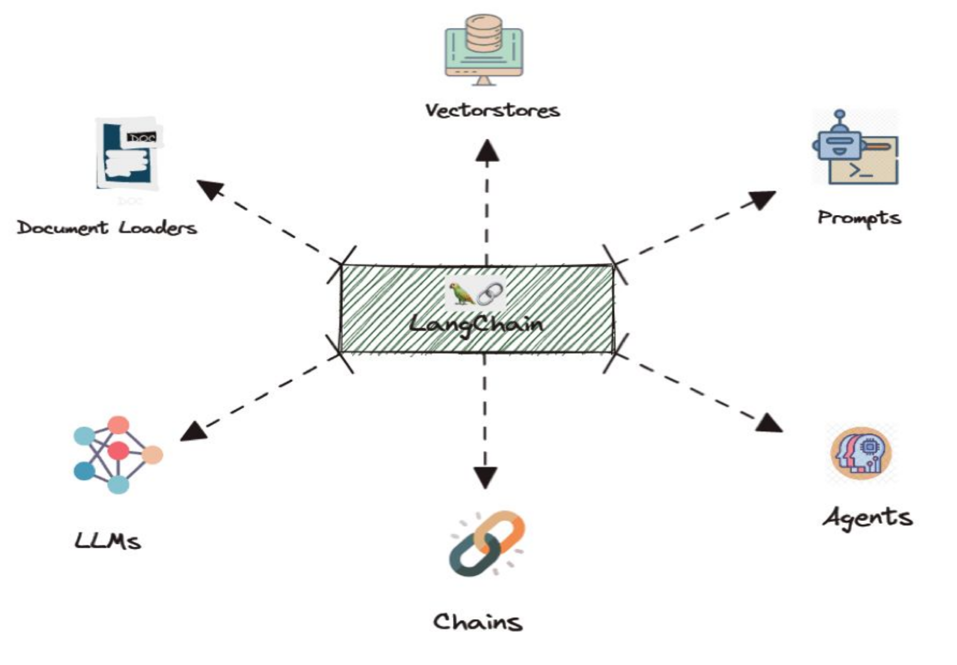
BioGPT

Pre-trained language models have attracted increasing attention in the biomedical domain, inspired by their great success in the general natural language domain. Among the two main branches of pre-trained language models in the general language domain, i.e. BERT (and its variants) and GPT (and its variants), the first one has been extensively studied in the biomedical domain, such as BioBERT and PubMedBERT. While they have achieved great success on a variety of discriminative downstream biomedical tasks, the lack of generation ability constrains their application scope. In this paper, we propose BioGPT, a domain-specific generative Transformer language model pre-trained on large-scale biomedical literature. We evaluate BioGPT on six biomedical natural language processing tasks and demonstrate that our model outperforms previous models on most tasks. Especially, we get 44.98%, 38.42% and 40.76% F1 score on BC5CDR, KD-DTI and DDI end-to-end relation extraction tasks, respectively, and 78.2% accuracy on PubMedQA, creating a new record. Our case study on text generation further demonstrates the advantage of BioGPT on biomedical literature to generate fluent descriptions for biomedical terms.

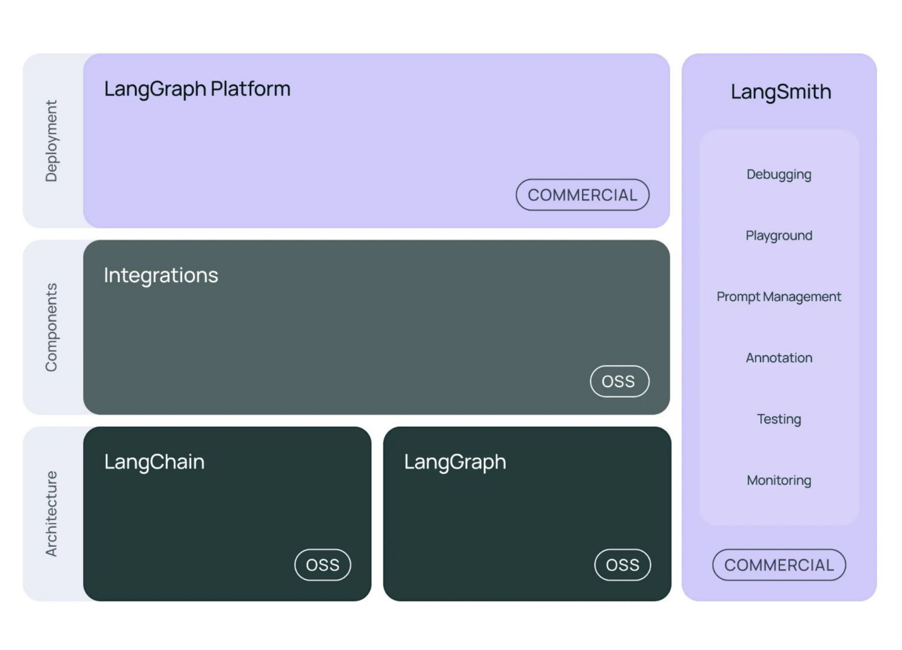
You can use this model directly with a pipeline for text generation. Since the generation relies on some randomness, we set a seed for reproducibility:

Langchain 🦜🔗

- Langchain (JS)
    - Allows easy integration with LLMs and other tools
    - LLM invocation, retrieval, chat memory

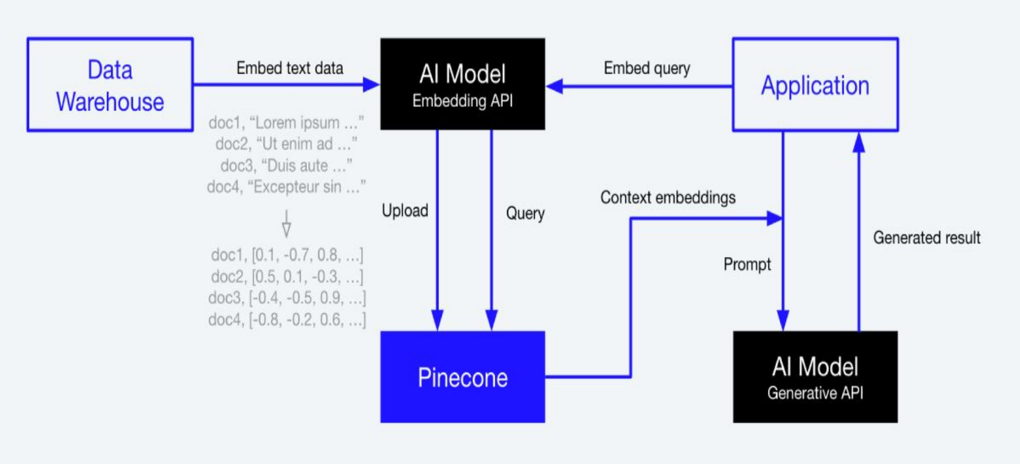


Langchain 🦜🔗


Deployment
LangGraph Platform
COMMERCIAL
Components
Integrations
OSS
Architecture
LangChain
OSS
LangGraph
OSS
LangSmith
Debugging
Playground
Prompt Management
Annotation
Testing
Monitoring
COMMERCIAL

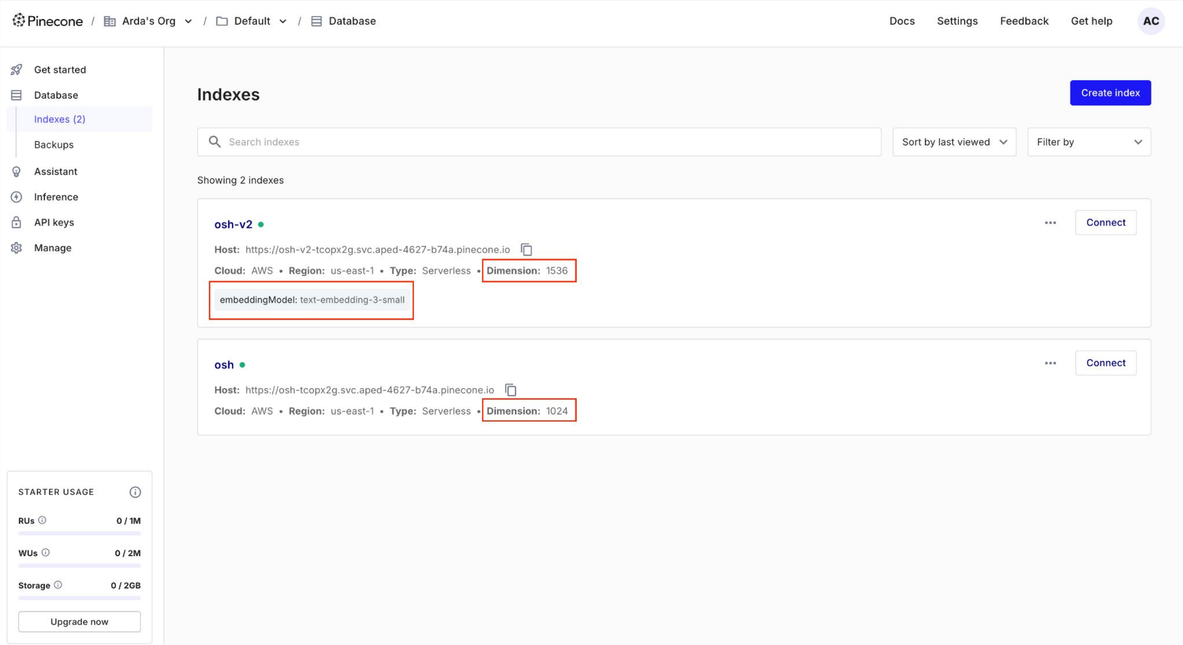
Pinecone Pinecone

- Serves as additional data source for RAG applications
- Data is stored as vectors in indexes
- Data needs to be turned into embeddings
- Embedding size is important!



Pinecone

Pinecone



Pinecone / Arda's Org / Default / Database

Docs Settings Feedback Get help AC

Indexes

Showing 2 indexes

osh-v2 ●

Host: https://osh-v2-tcopx2g.svc.aped-4627-b74a.pinecone.io

Cloud: AWS • **Region:** us-east-1 • **Type:** Serverless • **Dimension:** 1536

embeddingModel: text-embedding-3-small

osh ●

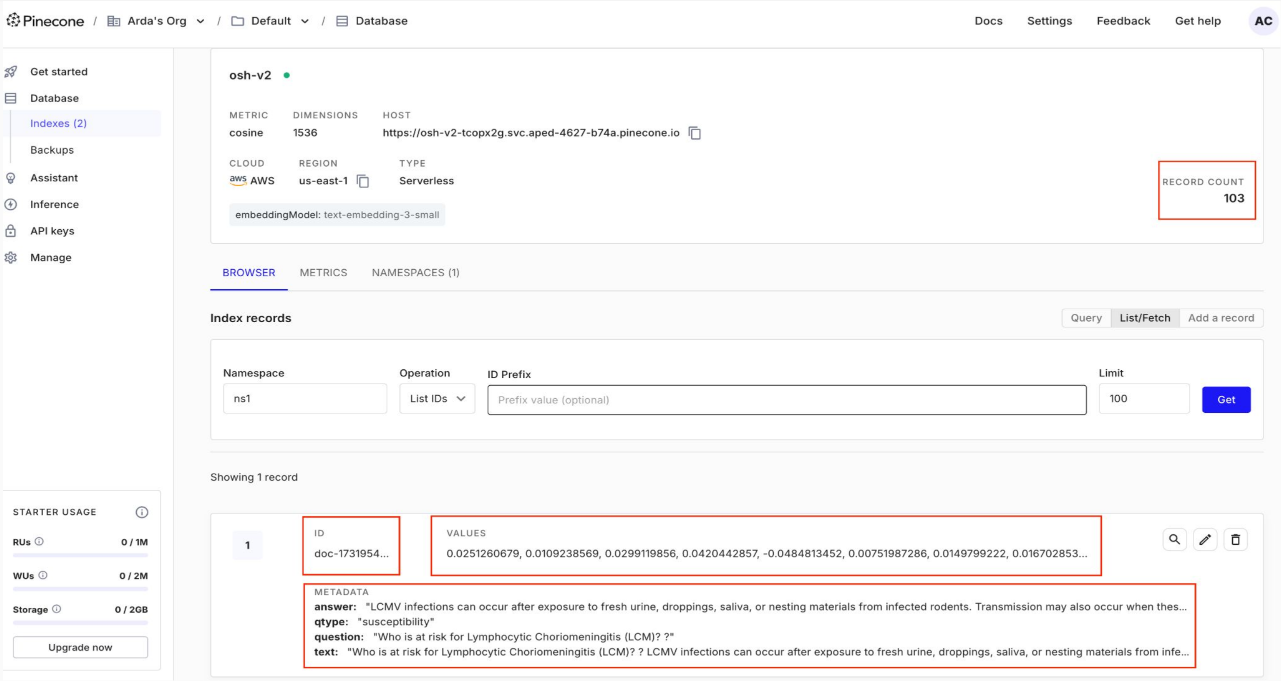
Host: https://osh-tcopx2g.svc.aped-4627-b74a.pinecone.io

Cloud: AWS • **Region:** us-east-1 • **Type:** Serverless • **Dimension:** 1024

STARTER USAGE

RUs	0 / 1M
WUs	0 / 2M
Storage	0 / 2GB

Pinecone



osh-v2 ●

METRIC: cosine
DIMENSIONS: 1536
HOST: https://osh-v2-tcopx2g.svc.aped-4627-b74a.pinecone.io

CLOUD: AWS
REGION: us-east-1
TYPE: Serverless

embeddingModel: text-embedding-3-small

RECORD COUNT
103

Index records

Namespace: ns1
Operation: List IDs
ID Prefix: Prefix value (optional)
Limit: 100

Showing 1 record

1

ID
doc-1731954...

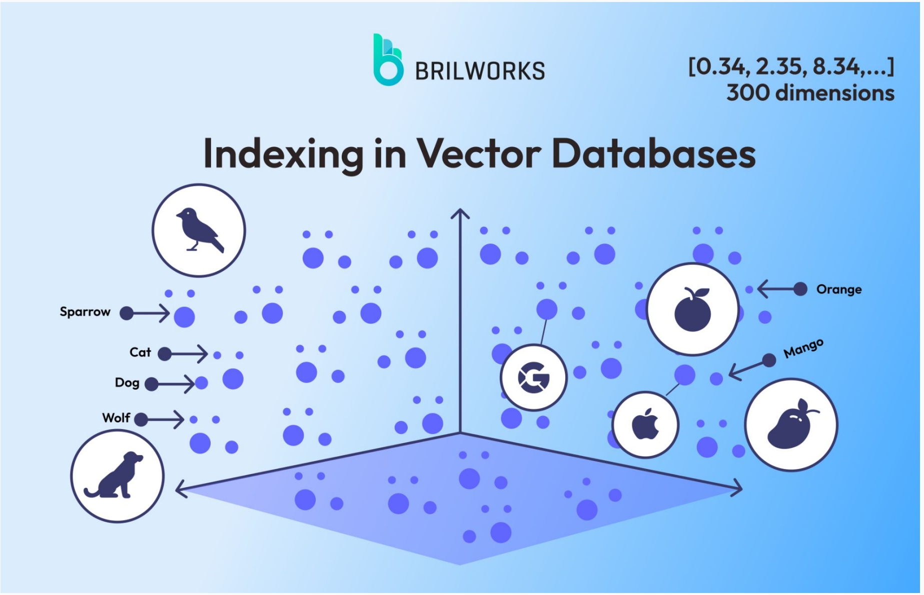
VALUES
0.0251260679, 0.0109238569, 0.0299119856, 0.0420442857, -0.0484813452, 0.00751987286, 0.0149799222, 0.016702853...

METADATA
answer: "LCMV infections can occur after exposure to fresh urine, droppings, saliva, or nesting materials from infected rodents. Transmission may also occur when thes...
qtype: "susceptibility"
question: "Who is at risk for Lymphocytic Choriomeningitis (LCM)? ?"
text: "Who is at risk for Lymphocytic Choriomeningitis (LCM)? ? LCMV infections can occur after exposure to fresh urine, droppings, saliva, or nesting materials from infe...

STARTER USAGE
RUs 0 / 1M
WUs 0 / 2M
Storage 0 / 2GB

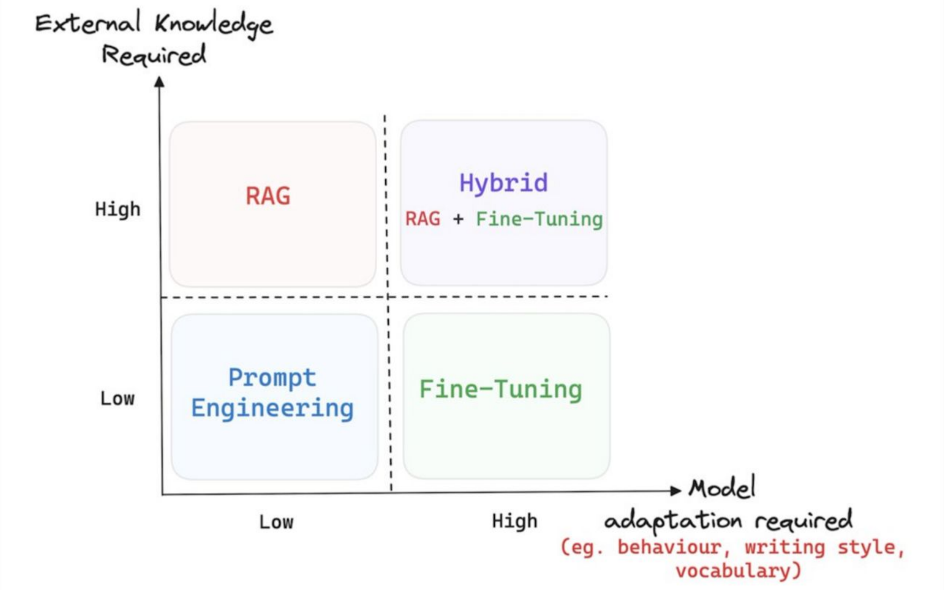
Pinecone

Pinecone



RAG vs Fine Tuning

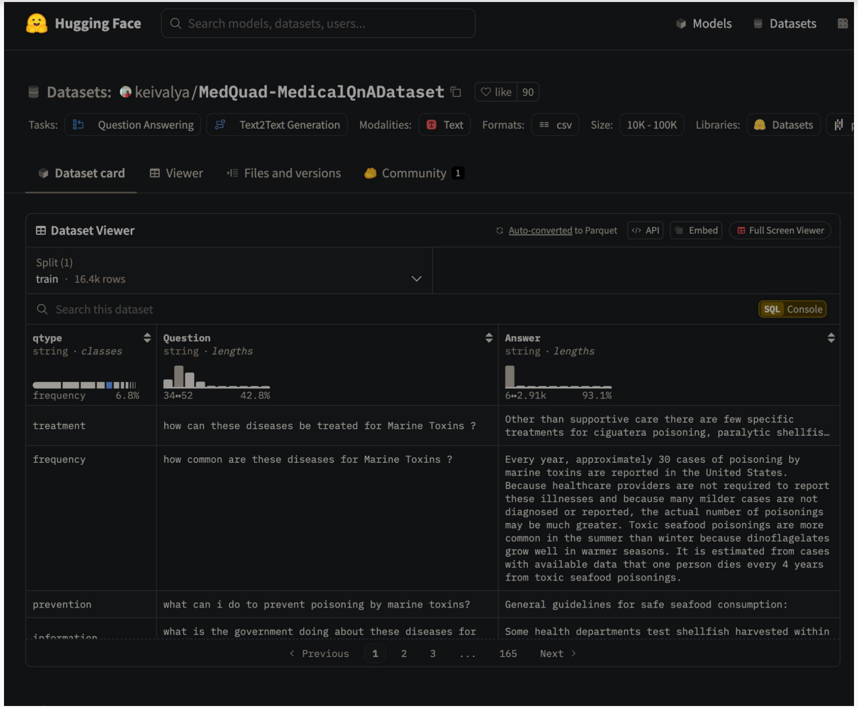
- RAG
  - Real time data
  - Ability to store data from different sources (research paper)
  - Factual context
- Fine tuning:
  - Availability of a dataset in a specific format
  - Changes in the “tone”


External Knowledge Required
High
Low
RAG
Hybrid
RAG + Fine-Tuning
Prompt Engineering
Fine-Tuning
Low
High
Model adaptation required
(eg. behaviour, writing style, vocabulary)

“LLMs are frozen in time”

Dataset

- MedQuad-MedicalQnADataset
  - Q&A pairs
  - 16k + rows



qtype string · *classes*	Question string · *lengths*	Answer string · *lengths*
frequency 6.8%	34↔52 42.8%	6↔2.91k 93.1%
treatment	how can these diseases be treated for Marine Toxins ?	Other than supportive care there are few specific treatments for ciguatera poisoning, paralytic shellfis…
frequency	how common are these diseases for Marine Toxins ?	Every year, approximately 30 cases of poisoning by marine toxins are reported in the United States. Because healthcare providers are not required to report these illnesses and because many milder cases are not diagnosed or reported, the actual number of poisonings may be much greater. Toxic seafood poisonings are more common in the summer than winter because dinoflagelates grow well in warmer seasons. It is estimated from cases with available data that one person dies every 4 years from toxic seafood poisonings.
prevention	what can i do to prevent poisoning by marine toxins?	General guidelines for safe seafood consumption:
information	what is the government doing about these diseases for	Some health departments test shellfish harvested within

Demo

Results

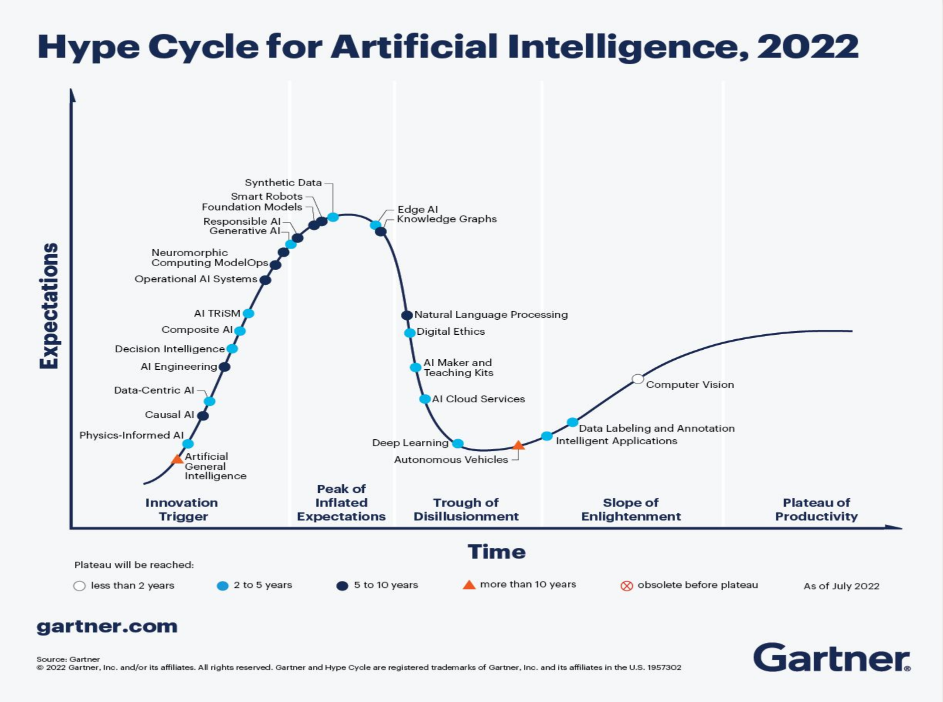
- Verified that the RAG pipeline works ✔️
- LLM has memory ✔️
- Comprehensive evaluation ❌

Challenges

- Finding the right dataset (Scale and type)
- Designing user interactions (multi-turn conversations)
- Conditional context retrieval (not every user query requires context)
- Evaluation methods (What is a successful interaction? What do we consider accurate?)
- Additional features like appointment scheduling, speech recognition etc.
- Familiarity with the tech stack
- Unit testing

Conclusion

Exciting time to be a developer :)



Hype Cycle for Artificial Intelligence, 2022

gartner.com

Source: Gartner

Thank you!