

One Stop Health

Arda Celik, Wesley Burnawan
Ontario Tech University
19 November 2024

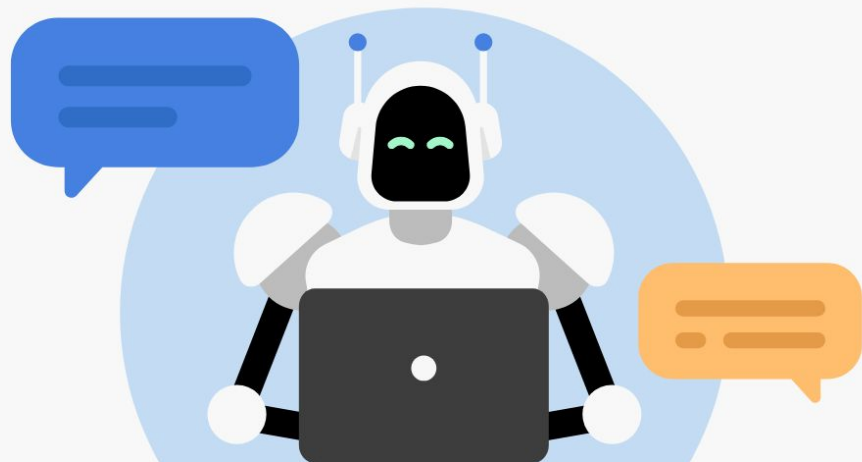
Motivation

- 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

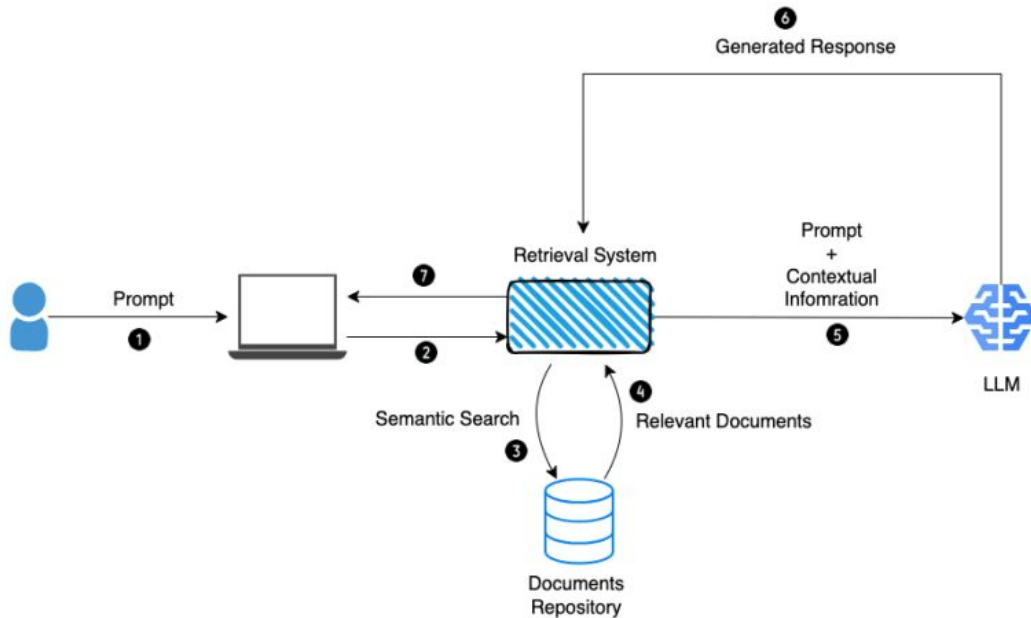
Our version: RAG Chatbot



Design and Implementation

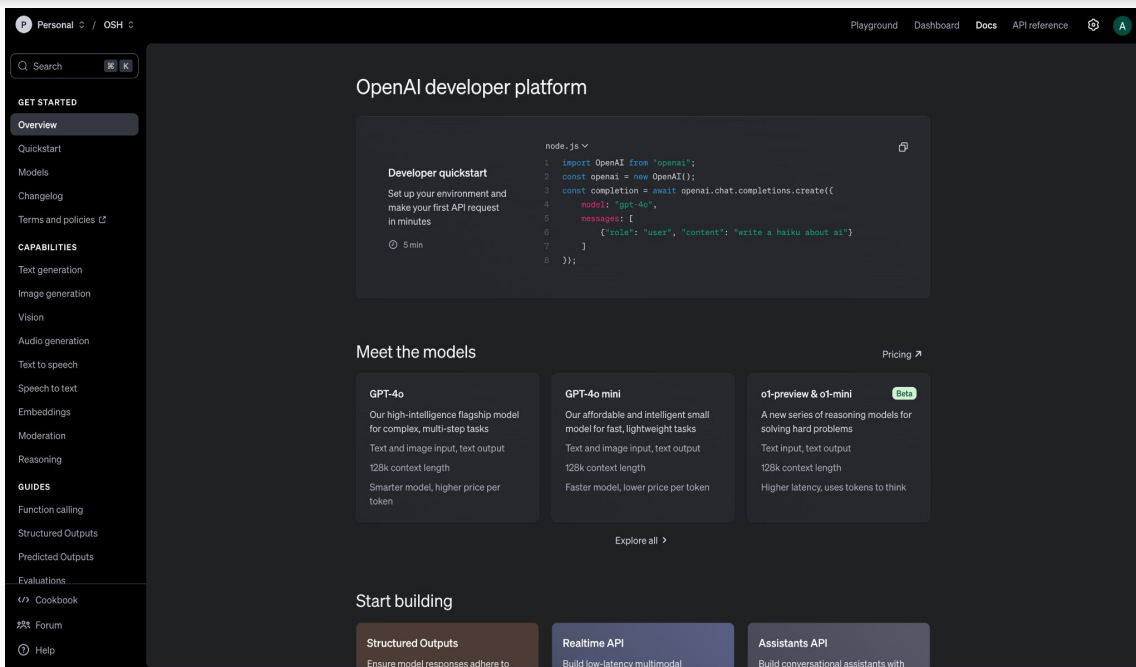
App components:

- 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



The screenshot shows the OpenAI developer platform dashboard. The left sidebar contains navigation links for 'GET STARTED' (Overview, Quickstart, Models, Changelog, Terms and policies), 'CAPABILITIES' (Text generation, Image generation, Vision, Audio generation, Text to speech, Speech to text, Embeddings, Moderation, Reasoning), and 'GUIDES' (Function calling, Structured Outputs, Predicted Outputs, Evaluations, Cookbook, Forum, Help). The main content area is titled 'OpenAI developer platform' and features a 'Developer quickstart' section with a code snippet for using the OpenAI API in Node.js. Below this is the 'Meet the models' section, which lists GPT-4o, GPT-4o mini, and o1-preview & o1-mini with their respective capabilities and pricing. At the bottom, the 'Start building' section offers options for 'Structured Outputs', 'Realtime API', and 'Assistants API'.

Personal / OSH

Playground Dashboard Docs API reference

Search

GET STARTED

Overview

Quickstart

Models

Changelog

Terms and policies

CAPABILITIES

Text generation

Image generation

Vision

Audio generation

Text to speech

Speech to text

Embeddings

Moderation

Reasoning

GUIDES

Function calling

Structured Outputs

Predicted Outputs

Evaluations

Cookbook

Forum

Help

OpenAI developer platform

Developer quickstart

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

5 min

```
node.js
1 import OpenAI from 'openai';
2 const openai = new OpenAI();
3 const completion = await openai.chat.completions.create({
4   model: 'gpt-4o',
5   messages: [
6     { "role": "user", "content": "write a haiku about ai" }
7   ]
8 });
```

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

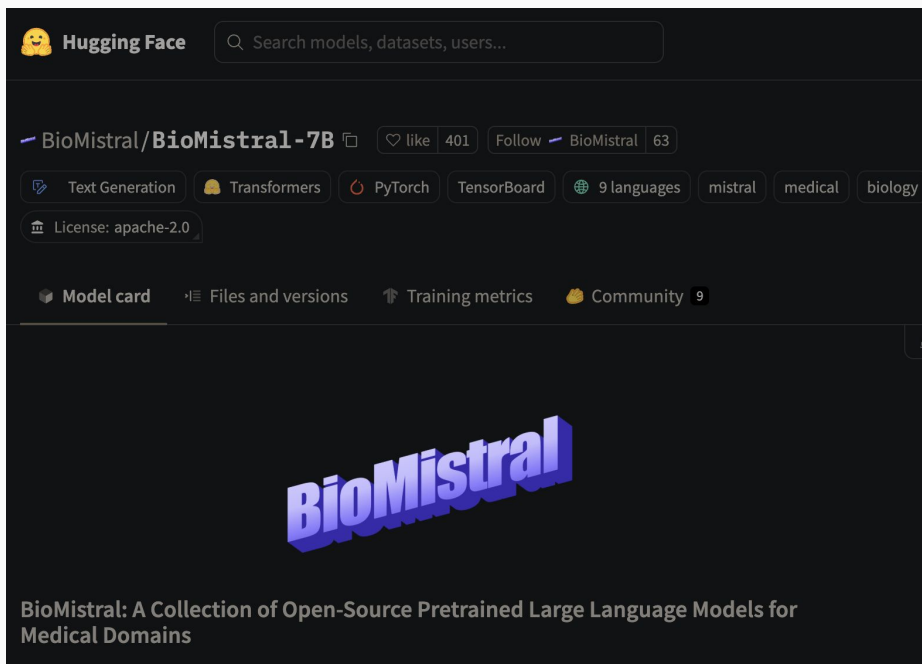
Realtime API

Build low-latency multimodal

Assistants API

Build conversational assistants with

LLM Alternatives



Hugging Face Search models, datasets, users...

BioMistral/BioMistral-7B like 401 Follow BioMistral 63

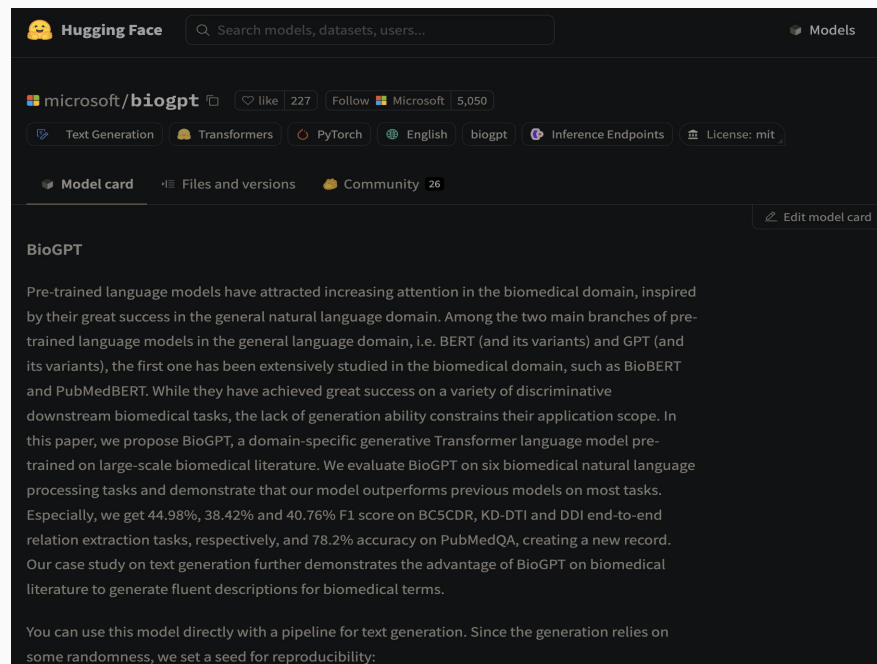
Text Generation Transformers PyTorch TensorBoard 9 languages mistral medical biology

License: apache-2.0

Model card Files and versions Training metrics Community 9

BioMistral

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



Hugging Face Search models, datasets, users... Models

microsoft/biogpt like 227 Follow Microsoft 5,050

Text Generation Transformers PyTorch English biogpt Inference Endpoints License: mit

Model card Files and versions Community 26

BioGPT

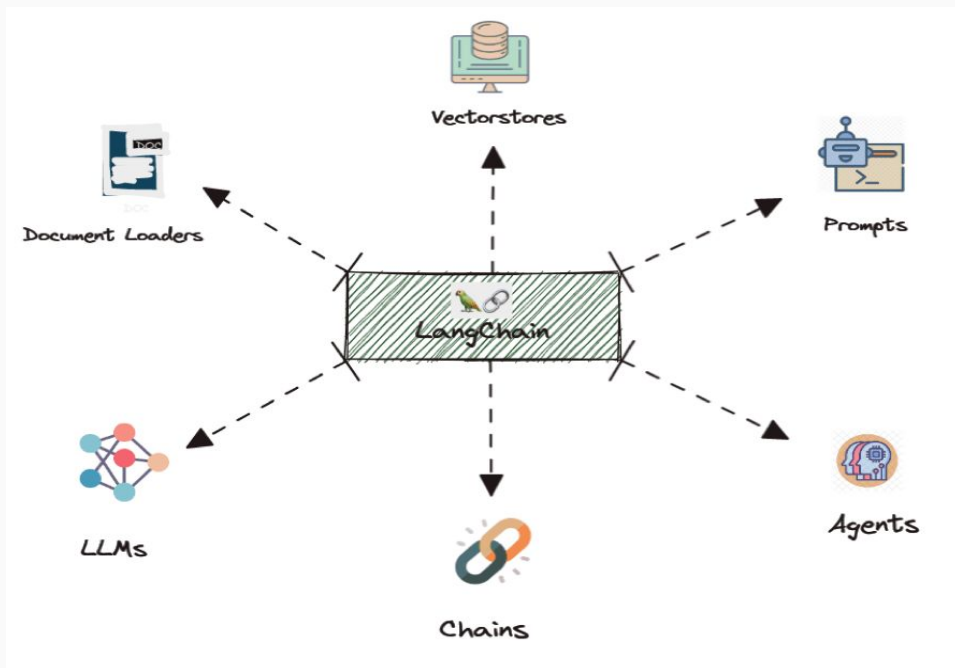
Pre-trained language models have attracted increasing attention in the biomedical domain, inspired by their great success in the general natural language domain. Among the two main branches of pre-trained language models in the general language domain, i.e. BERT (and its variants) and GPT (and its variants), 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:

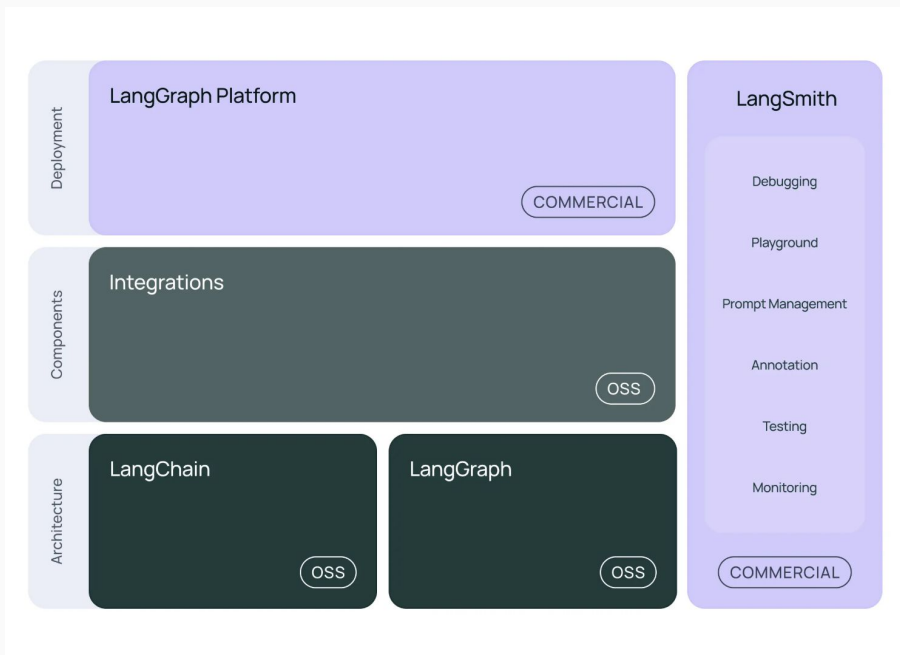
Langchain



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

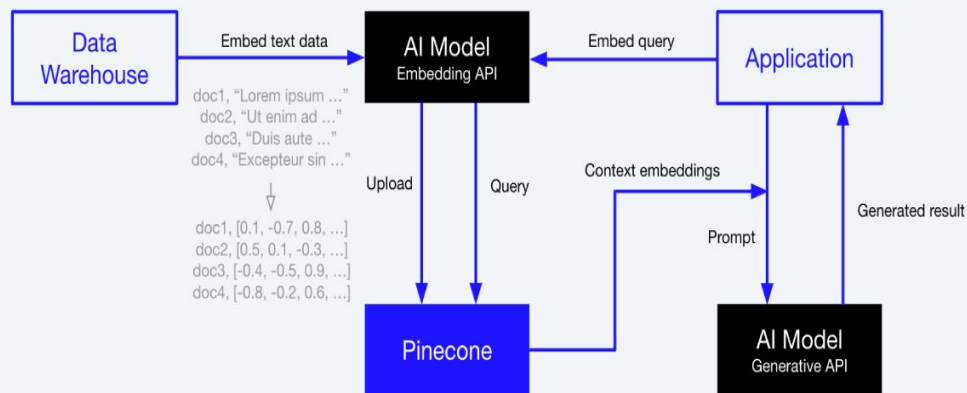


Langchain



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

Arda's Org

Default

Database

Docs

Settings

Feedback

Get help

AC

Get started

Database

Indexes (2)

Backups

Assistant

Inference

API keys

Manage

STARTER USAGE

RUs

0 / 1M

WUs

0 / 2M

Storage

0 / 2GB

Upgrade now

Indexes

Create index

Search indexes

Sort by last viewed

Filter by

Showing 2 indexes

osh-v2

...

Connect

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

...

Connect

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

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

Pinecone



Pinecone

Pinecone

Arda's Org

Default

Database

Docs

Settings

Feedback

Get help

AC

Get started

Database

Indexes (2)

Backups

Assistant

Inference

API keys

Manage

osh-v2

METRIC

DIMENSIONS

HOST

cosine

1536

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

CLOUD

REGION

TYPE

AWS

us-east-1

Serverless

embeddingModel: text-embedding-3-small

RECORD COUNT
103

BROWSER

METRICS

NAMESPACES (1)

Index records

QueryList/FetchAdd a record

Namespace

Operation

ID Prefix

Limit

ns1

List IDs

Prefix value (optional)

100

Get

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 the...

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

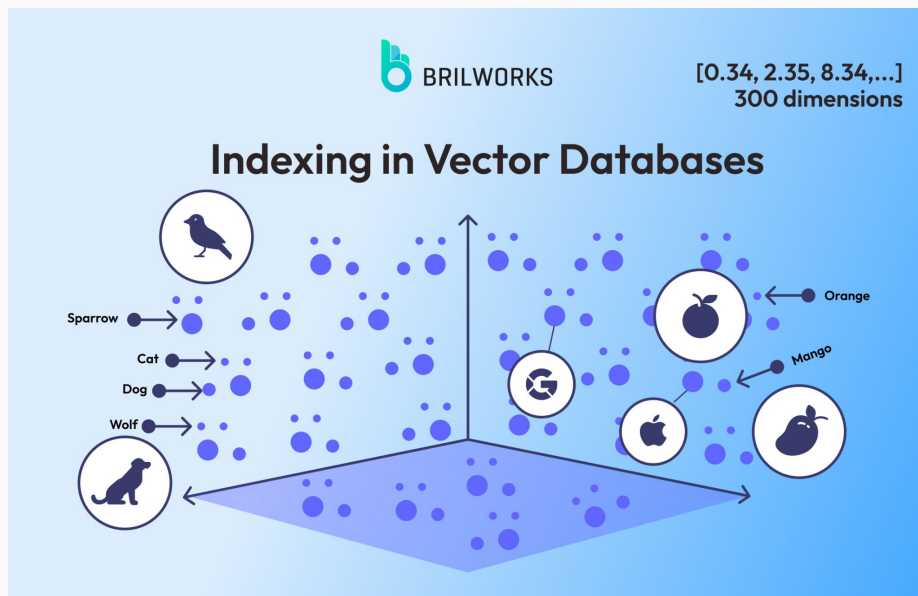
WUs

0 / 2M

Storage

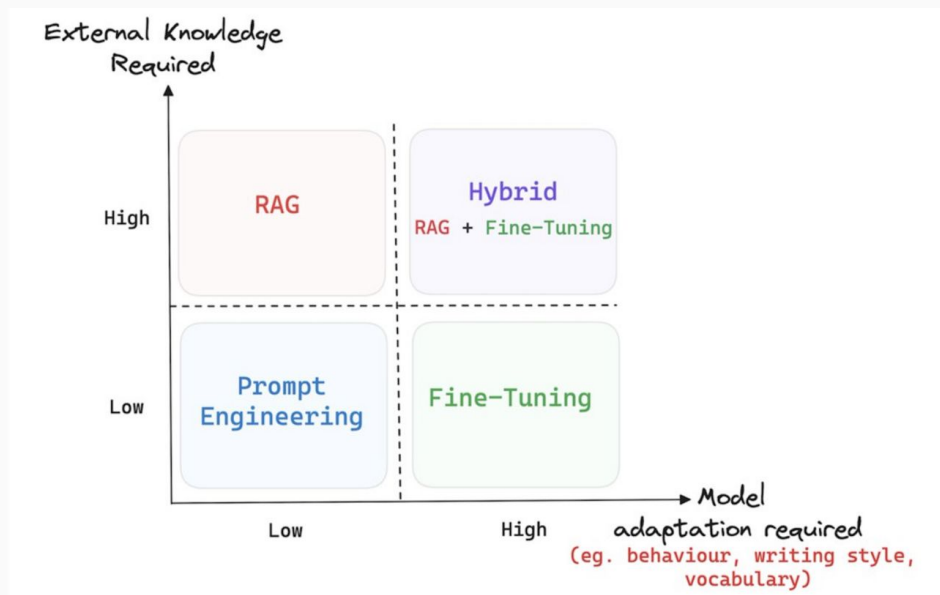
0 / 2GB

Upgrade now



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”



“LLMs are frozen in time”



Dataset

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

The screenshot shows the Hugging Face interface for the MedQuad-MedicalQnADataset. The page header includes the Hugging Face logo, a search bar, and navigation links for Models, Datasets, and a profile icon. The dataset name 'MedQuad-MedicalQnADataset' is prominently displayed, along with a 'like' button showing 90 likes. Below the name, there are tabs for 'Question Answering', 'Text2Text Generation', 'Modalities: Text', 'Formats: csv', 'Size: 10K - 100K', and 'Libraries: Datasets'. The 'Dataset card' tab is selected, showing a 'Dataset Viewer' section. This section includes a search bar, a 'SQL Console' button, and three histograms for 'qtype', 'Question', and 'Answer'. The 'qtype' histogram shows a frequency of 6.8% for 'string' and 0.8% for 'classes'. The 'Question' histogram shows a frequency of 34.5% for 'string' and 42.8% for 'lengths'. The 'Answer' histogram shows a frequency of 6.2% for 'string' and 93.1% for 'lengths'. Below the histograms, there is a table with four rows of data: 'treatment', 'frequency', 'prevention', and 'information'. Each row contains a question and an answer. The 'treatment' row asks 'how can these diseases be treated for Marine Toxins ?' and the answer is 'Other than supportive care there are few specific treatments for ciguatera poisoning, paralytic shellfish...'. The 'frequency' row asks 'how common are these diseases for Marine Toxins ?' and the answer is '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 dinoflagellates grow well in warmer seasons. It is estimated from cases with available data that one person dies every 4 years from toxic seafood poisonings..'. The 'prevention' row asks 'what can i do to prevent poisoning by marine toxins?' and the answer is 'General guidelines for safe seafood consumption:'. The 'information' row asks 'what is the government doing about these diseases for' and the answer is 'Some health departments test shellfish harvested within'. At the bottom of the table, there are navigation links for 'Previous', '1', '2', '3', and 'Next', along with a page number '165'.

Hugging Face

Search models, datasets, users...

Models Datasets

Datasets: keivalya/MedQuad-MedicalQnADataset like 90

Tasks: Question Answering Text2Text Generation Modalities: Text Formats: csv Size: 10K - 100K Libraries: Datasets

Dataset card Viewer Files and versions Community 1

Dataset Viewer

Split (1)
train · 16.4k rows

Search this dataset SQL Console

qtype	Question	Answer
string · classes	string · lengths	string · lengths
frequency 6.8%	34.5% 42.8%	6.2% 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 shellfish...
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 dinoflagellates 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

< Previous 1 2 3 ... 165 Next >

Demo

Results

- Verified that the RAG pipeline works ✓
- LLM has memory ✓
- Comprehensive evaluation ✗

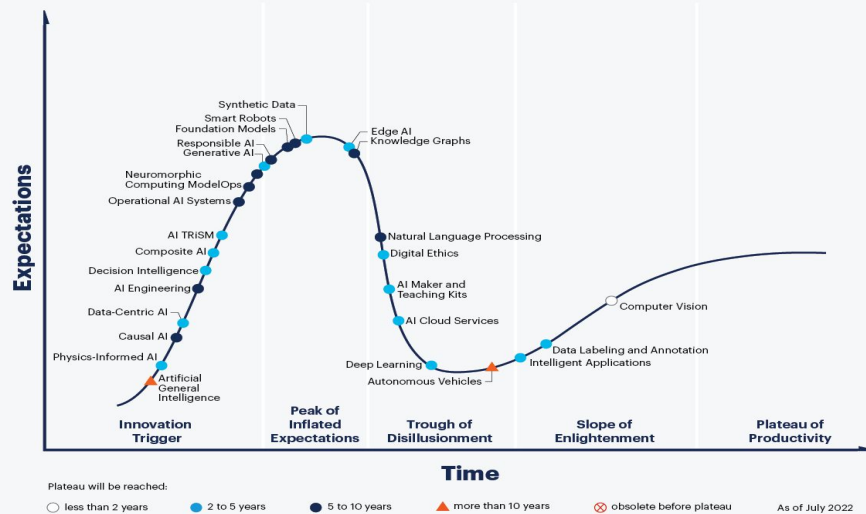
Challenges

- 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](https://www.gartner.com)

Source: Gartner
© 2022 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and Hype Cycle are registered trademarks of Gartner, Inc. and its affiliates in the U.S. 1957302

Gartner.

Thank you!

