Juan David Hernández Giraldo

- Full Stack AI Engineer
- Co founder & Co organizer Python Medellín.
- Co organizer Pycon Colombia.
- in /juandhernandez
- davoshack

Building software on top of Large Language Models

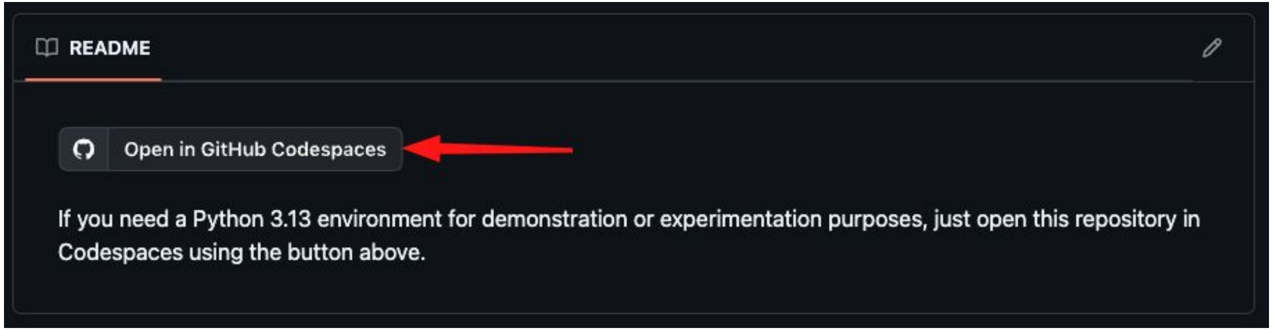
AI for Full stack Developers

What topics do we cover?

- A review of the best currently available models
- Using multi-modal LLMS to analyze images, audio and video
- Use-cases that LLMs can be effectively applied to
- How to access the most capable models via their various APIs
- Prompt engineering
- Retrieval Augmented Generation (RAG)
- LLM tool usage
- Effective LLM Workflows

If you're going to be using Codespaces…

https://github.com/pamelafox/python-3.13-playground



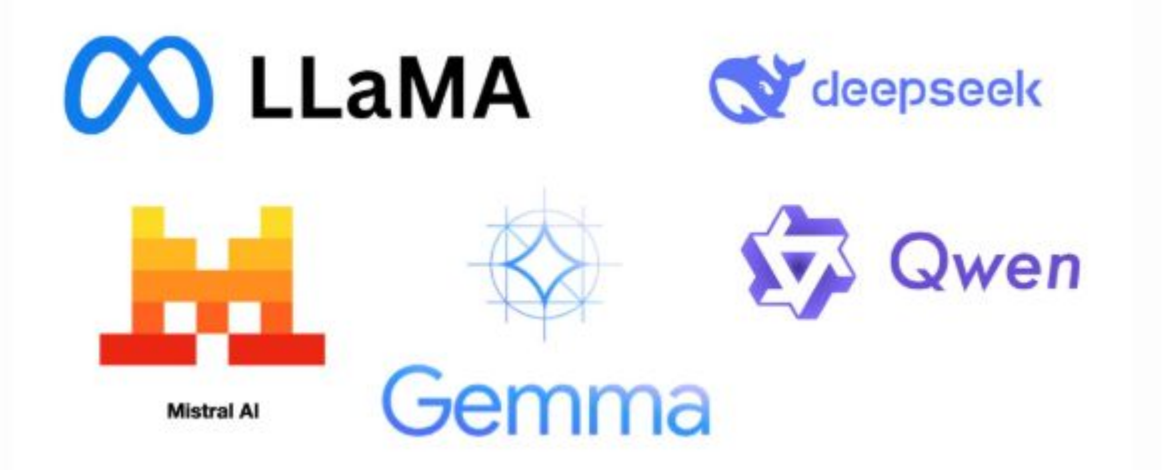
If you need a Python 3.13 environment for demonstration or experimentation purposes, just open this repository in Codespaces using the button above.

Today's LLM landscape

The big three

OpenAI
Gemini
ANTHROP\C

Open weights



At least 18 labs have released a GPT-4 equivalent model

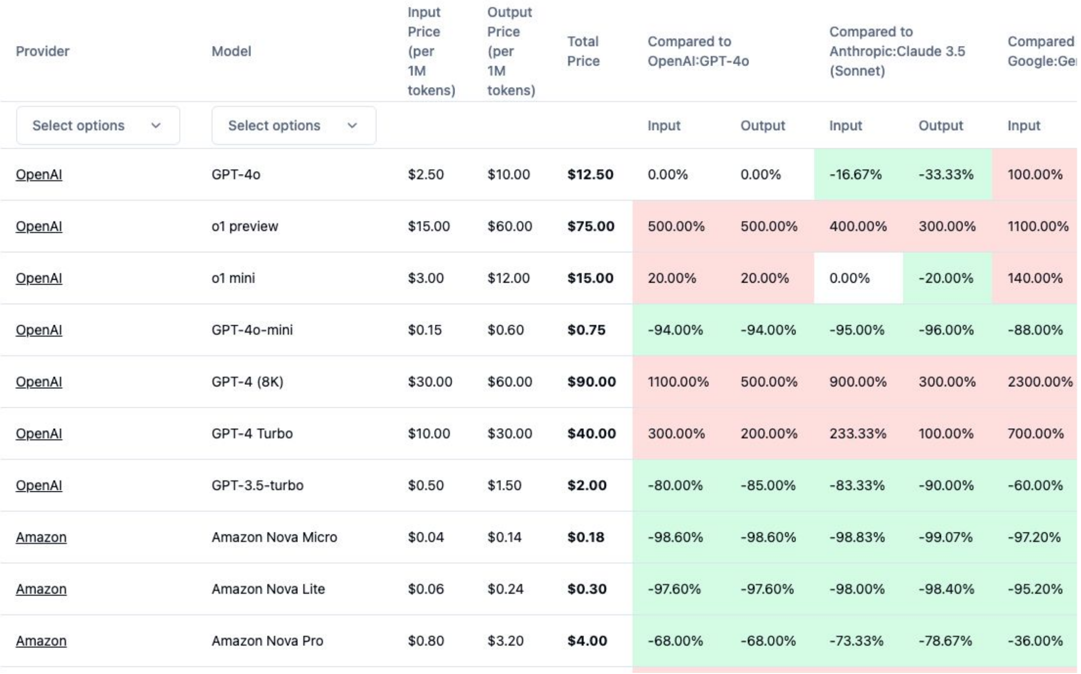
Google, OpenAI, Alibaba (Qwen), Anthropic, Meta, Reka AI, 01 AI, Amazon, Cohere, DeepSeek, Nvidia, Mistral, NexusFlow, Zhipu AI, xAI, AI21 Labs, Princeton and Tencent

Multi-modal has been a big theme over the past ~18 months

Image/audio/video input, and increasingly audio/image output as well

We're spoiled for choice

Prices keep dropping



Provider	Model	Input Price (per 1M tokens)	Output Price (per 1M tokens)	Total Price	Compared to OpenAI:GPT-4o		Compared to Anthropic:Claude 3.5 (Sonnet)		Compared Google:Ge
Select options ⌄	Select options ⌄				Input	Output	Input	Output	Input
OpenAI	GPT-4o	$2.50	$10.00	**$12.50**	0.00%	0.00%	-16.67%	-33.33%	100.00%
OpenAI	o1 preview	$15.00	$60.00	**$75.00**	500.00%	500.00%	400.00%	300.00%	1100.00%
OpenAI	o1 mini	$3.00	$12.00	**$15.00**	20.00%	20.00%	0.00%	-20.00%	140.00%
OpenAI	GPT-4o-mini	$0.15	$0.60	**$0.75**	-94.00%	-94.00%	-95.00%	-96.00%	-88.00%
OpenAI	GPT-4 (8K)	$30.00	$60.00	**$90.00**	1100.00%	500.00%	900.00%	300.00%	2300.00%
OpenAI	GPT-4 Turbo	$10.00	$30.00	**$40.00**	300.00%	200.00%	233.33%	100.00%	700.00%
OpenAI	GPT-3.5-turbo	$0.50	$1.50	**$2.00**	-80.00%	-85.00%	-83.33%	-90.00%	-60.00%
Amazon	Amazon Nova Micro	$0.04	$0.14	**$0.18**	-98.60%	-98.60%	-98.83%	-99.07%	-97.20%
Amazon	Amazon Nova Lite	$0.06	$0.24	**$0.30**	-97.60%	-97.60%	-98.00%	-98.40%	-95.20%
Amazon	Amazon Nova Pro	$0.80	$3.20	**$4.00**	-68.00%	-68.00%	-73.33%	-78.67%	-36.00%

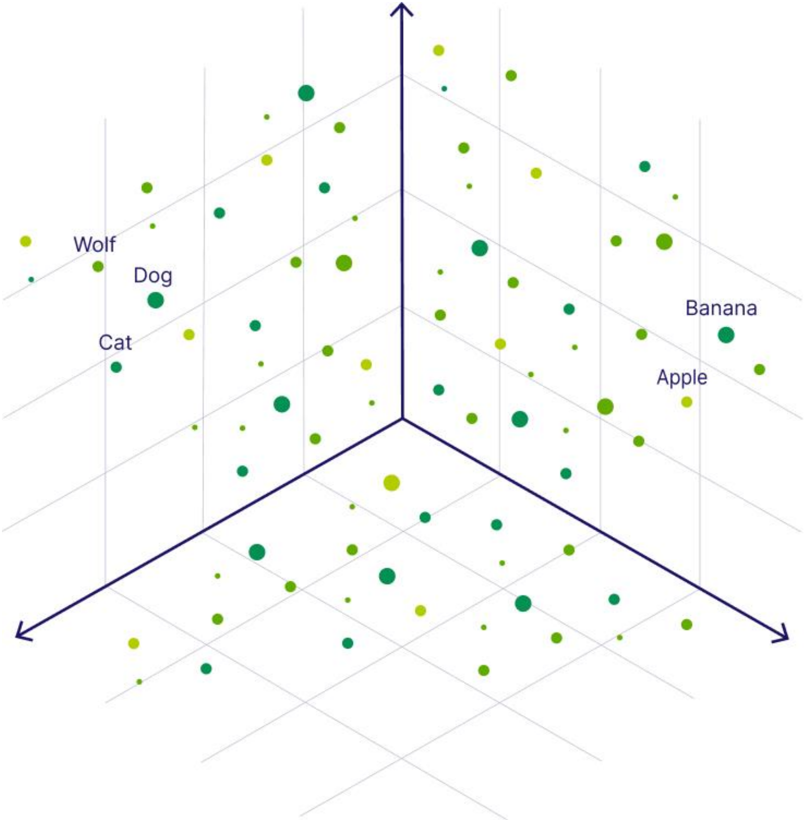
LLMs suffer from a "jagged frontier" - they are great at some things, terrible at others and it's surprisingly hard to figure out which

The best thing to do is play with them, a lot, and keep notes of your experiments …

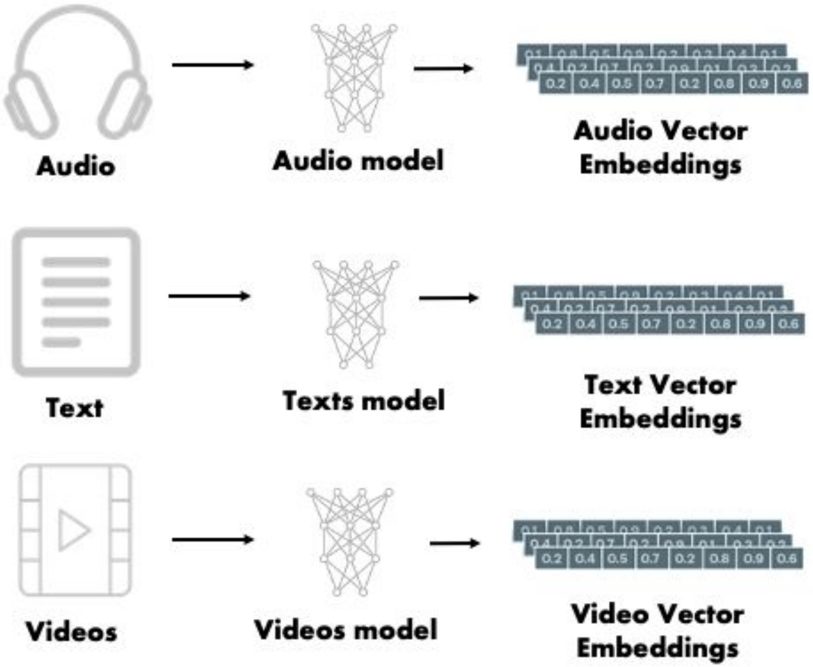
and be ready to switch between them

Let's start prompting

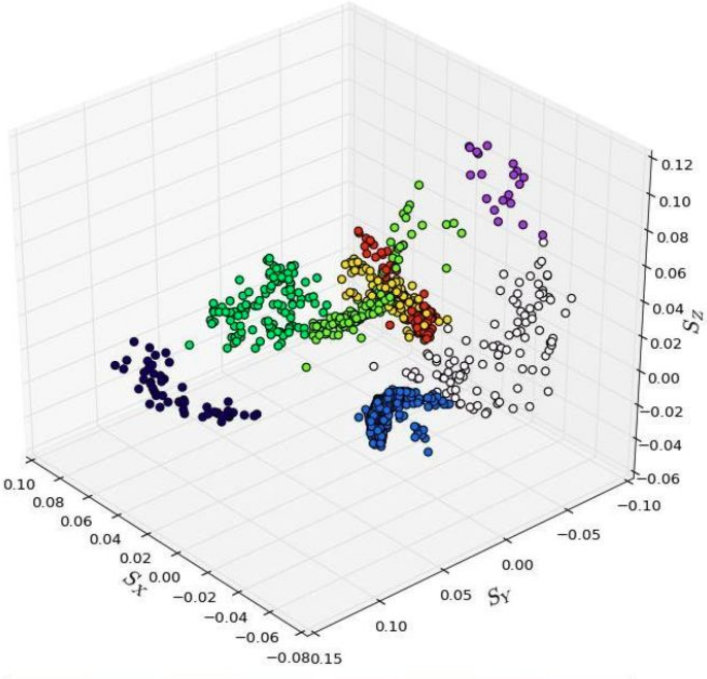
Embeddings



What are Embeddings?

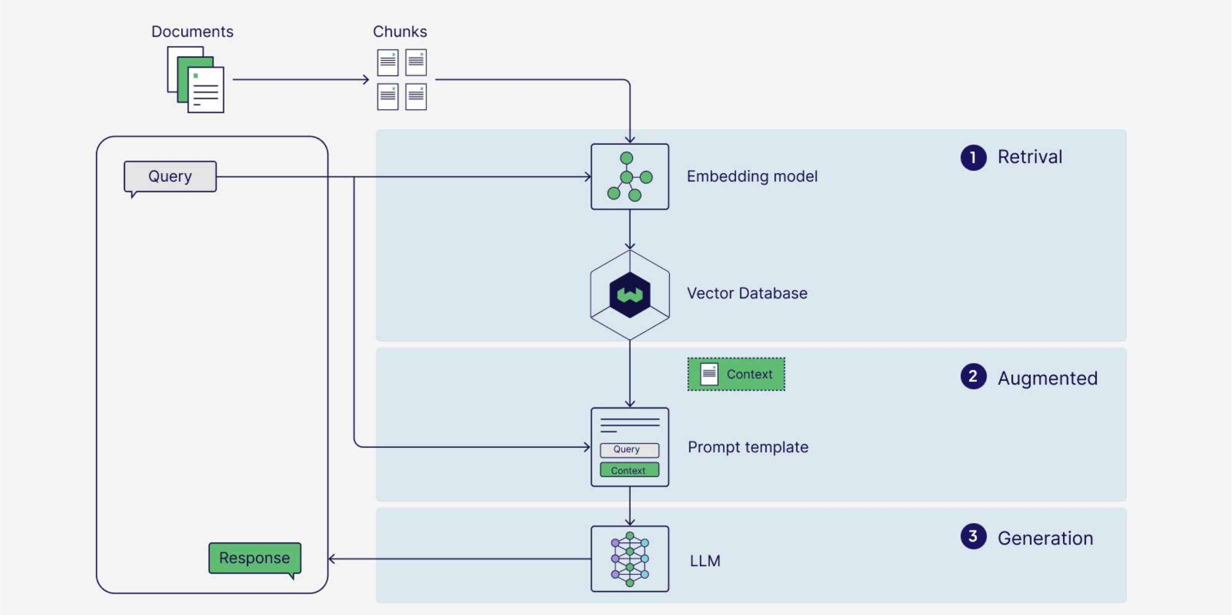


A location in many-multi dimensional space

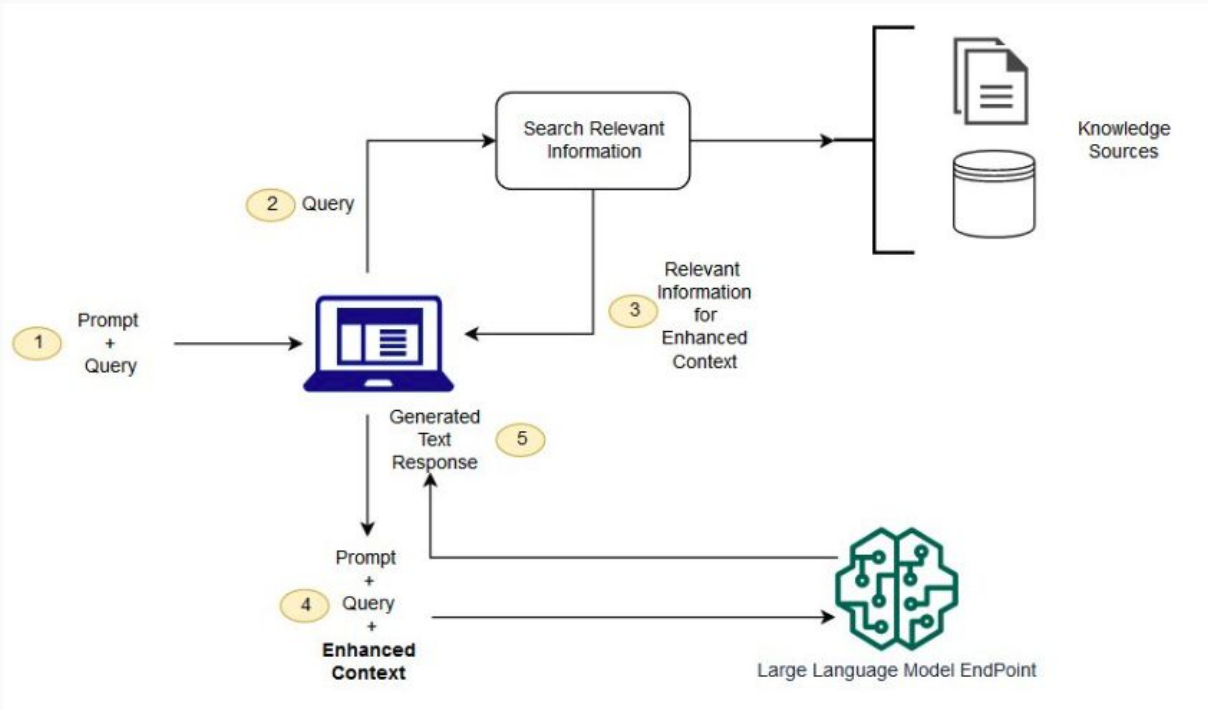


Semantic Search

Embed the user's question, find related documents.



Semantic Search and RAG


Search Relevant Information
Knowledge Sources
2 Query
1 Prompt + Query
3 Relevant Information for Enhanced Context
Generated Text Response 5
4 Prompt + Query + Enhanced Context
Large Language Model EndPoint

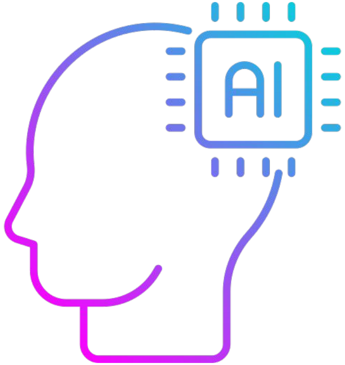
Evals

LLM as a judge

Questions with a "right" answer

- Your AI Product Needs Evals
- Creating a LLM-as-a-Judge That Drives Business Results
- Ragas

Building effective agents



Resources

- AI Agents for Beginners
- Anthropic Cookbook
- Prompt Engineering Guide
- LLM Engineer Handbook