

Applied GenAI: NewsBot Workshop

Building you first LLM powered Application

Introduction – Sarfaraz Hussein

Senior ML Researcher at Motional Inc.

(Autonomous Driving)

PhD – Center for Research in Computer Vision @ UCF

Ex-Amazon, HomeDepot, Symantec, Siemens

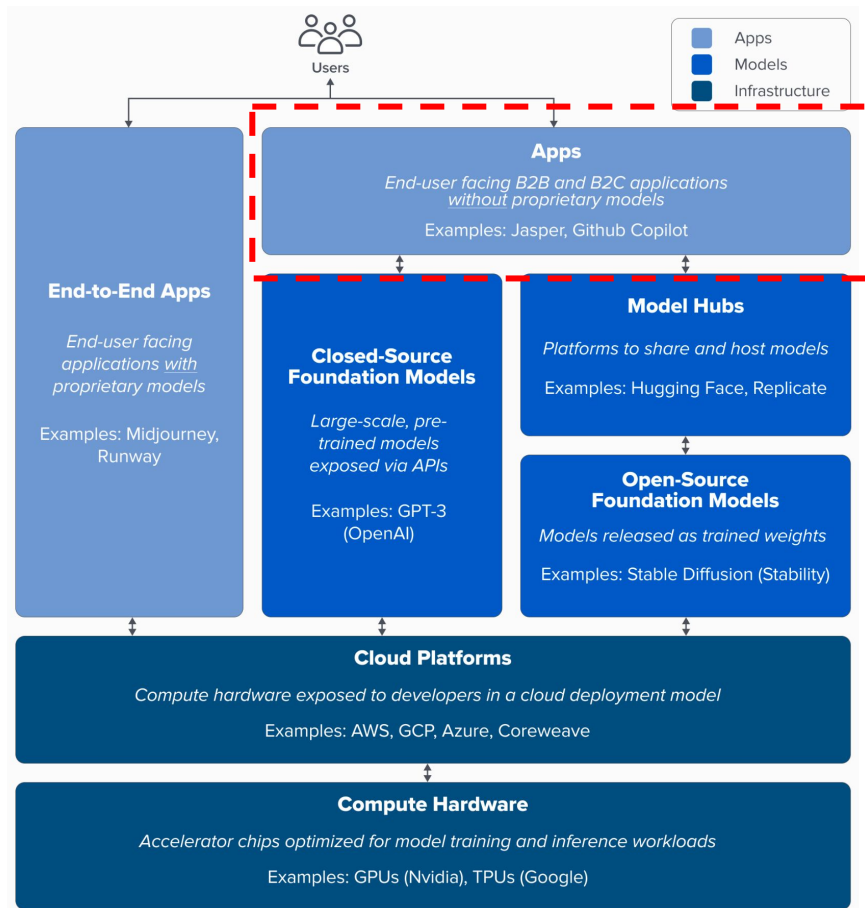
Over a decade of Teaching and Research Experience



- **Objective:** Building your first LLM Application
- **Prerequisite:** Basic Knowledge and Understanding of Python
- **Content:**
 - Problem Statement and Motivation
 - Gen AI Fundamentals
 - LLM Fundamentals
 - Prompt Engineering
 - Designing prompts for different tasks
 - Lifecycle of LLM Application Development
 - Demo #1: News Summarization (NYT API)
 - Demo #2: News Summarization - Retrieval and Embedding
 - Closing Remarks

This is NOT a workshop where we discuss the theoretical details of different LLM models

Gen AI tech stack | The new growth engine



The new growth frontiers:
Gen AI Applications

Build your first GenAI Application

News Summarizer

- By using advanced algorithms and machine learning, AI is now able to efficiently and effectively summarize news articles into concise, easy-to-read summaries that capture the essence of a story in just a few sentences.
- This technology has the potential to revolutionize the way we consume news, making it more:
 - accessible and
 - digestible for people who are short on time or simply overwhelmed by the volume of information available.

- A News Summarizer should leverage the capabilities of Artificial Intelligence (AI) to analyze and summarize news data from diverse sources, enabling users to grasp the key highlights and request custom inputs.
- Here are few prompts for an AI summarizer to aggregate news metadata and output a summary with citations to its sources.
 - *'Summarize articles with citations.'*
 - *'For every sentence, cite the article name.'*
 - *'At the end of your summary, create a sources list of each result you cited, with the article name, author, and link.'*

Generative AI (Gen AI): Refers to deep-learning models that can generate high-quality text, images, and other content based on the data they are trained on.

- No power on earth can stop an idea whose time has come!
- Several old ways of doing things have hit a brick wall and are being upended by Generative AI
- GPUs have become more powerful and all pervasive
- We have more accelerated computing at our disposal and there have been many algorithmic advances in last few years

- The pace at which research is moving is overwhelming
- How to keep up with the latest and greatest? How to approach reading papers?
- How to adapt to this changing landscape? With my current skills, where can I be most useful?
- Where to begin? Is there some method behind this madness?

- Pre-train - Train large models on internet scale data. Needs many months of training
- Finetuning / Domain Adaptation - Pre-trained models still need to be adapted to specific domains. Finetune the pre-trained models efficiently
- Alignment / Customization - Finetuned models could still make things up (hallucinate) or emit disagreeable answers. Apply RLHF / DPO to align the finetuned models to human preferences

- Data Preparation - Take internet scale data and prepare it for pre-train or take domain data for finetuning
- Training - Typically, this stage can involve distributed training over multiple nodes having multiple GPUs
- Inference Optimization - Apply techniques to optimize the latency of a single inference to support latency and bandwidth constraints
- Model health monitoring - Keep your models refreshed and factual

Backend Engineer

Braintrust · San Francisco, CA (On-site)

[Apply](#)[Save](#)[...](#)

Role Requirements

- Enjoy working in a fast-paced environment & wear multiple hats.
- 3+ years of backend / full-stack development experience.
- Experience with developing products built on top of LLMs / ML.
- Proficient in Python, FastAPI & PostgreSQL.
- Experience in building products from zero to one.
- Ability to seek & find the right resources for solving open-ended problems.
- Located in the San Francisco Bay Area or willing to relocate.
- BS/MS in Computer Science, Engineering, or a related technical field.

Backend Engineer

UpCodes · United States (Remote)

[Apply](#)[Save](#)[...](#)

- Enjoyable to work with

TECHNOLOGY STACK

- Python, PostgreSQL, FastAPI, Redis, TypeScript, React, Next.js, Tailwind, AWS, Kubernetes, Prometheus, Pinecone, GPT-4

EXAMPLE PROJECTS

- Use an LLM to identify references to other sections in the text of the law
- Improve and migrate our data model for the content we host
- Retrieve semi-structured data from various online sources and automate the structuring of the data
- Improve the evaluation framework for our search engine

Frontend engineer

Ntropy · San Francisco, CA (On-site)

[Apply](#)[Save](#)[...](#)

The following are a big plus

- fluency in Javascript and Python
- past experience with React / Typescript stacks
- recognized open-source contributions
- at ease with data visualization tools
- familiarity with machine-learning concepts and LLMs
- experience with industry-standard databases, such as Postgres and Redis
- strong understanding of data structures, algorithms and software-design principles

Generative AI skills are becoming a norm in SWE JDs

Fullstack Engineer II, Product

Khan Academy · Mountain View, CA

[Apply](#)[Save](#)[...](#)

- awareness, awareness of other, and the ability to adopt inclusive perspectives, attitudes, and behaviors to drive inclusion and belonging throughout the organization.
- Empathy for learners around the world. You love learning and are excited about helping others learn to love learning. You're motivated to learn new things and share what you learn with the world.
 - Experience using Generative AI / LLMs to build products a plus (but not required).

Perks And Benefits

Staff Fullstack Engineer, Com...

Airbnb · San Francisco, CA (Remote)

[Apply](#)[Save](#)[...](#)

- design, implementation and testing. This involves understanding the nuances of feature requests and developing scalable, flexible solutions to meet those needs effectively.
- Collaborate with infrastructure engineering team Core Machine Learning team to empower Airbnb LLM products.
 - Work with other teams in the company to understand their productivity and feature requests, and build solutions to resolve them scalably and flexibly.
 - Participate in all phases of software development including architecture design, implementation and testing.
 - Work collaboratively with cross-functional partners including product managers, operations and data scientists, identify opportunities for business impact, understand and prioritize requirements for machine learning systems and data pipelines.

Senior Fullstack Engineer, Sim...

Waymo · Los Angeles, CA

[Apply](#)[Save](#)[...](#)

experienced driver and lead efforts such as:

- Building seamless web tools and efficient data pipelines for simulated and driving events evaluation, triaging tens of millions of data points used by Waymo Ops and Eng.
- Building auto-triage pipelines that provide useful signals and clustering for triage productivity and quality improvement, incorporating technologies like LLMs and generative AI.
- Collaborating across teams, with SWEs, Product Managers, Data Science, Operations, and UX to build the best user experience for our developer tools and improve the development speed of Waymo software engineers.
- Engineering solutions with an eye towards quality, performance, and stability.



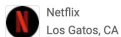
AI Product Manager

Microsoft · Mountain View, CA

Base pay range

\$94,300.00/yr - \$238,600.00/yr

Product Manager, Consumer Intelligence Algorithms



Netflix
Los Gatos, CA

\$ 110K–190K a year Full-time

Product Manager - Generative AI



Meta
Menlo Park, CA

Senior Product Manager, Generative AI



Google
Portland, OR

Product Manager, Siri and AI/ML

Apple · 4.1 ★

Cupertino, CA

\$132,300 - \$241,500 a year



Senior Gen Ai GTM Specialist, Amazon Bedrock

Amazon Web Services (AWS) · San Francisco, CA · Reposted 1 d



\$118.4K/yr - \$220.2K/yr · Full-time · Mid-Senior level



**1000+ Openings
for AI PM jobs**

AI Engineering for Software Engineers

- Overview of AI-Powered Software Development Lifecycle (SDLC)
- Discuss the design Principles for AI Engineering
- Introduction to essential AI Tech Stack: LLM frameworks, vector databases, LLM integration, front-end frameworks & guardrails to ensure model reliability.
- Develop a full-stack application using LLMs
- Overview of deployment Strategies on Cloud Platforms

Product Management with Generative AI

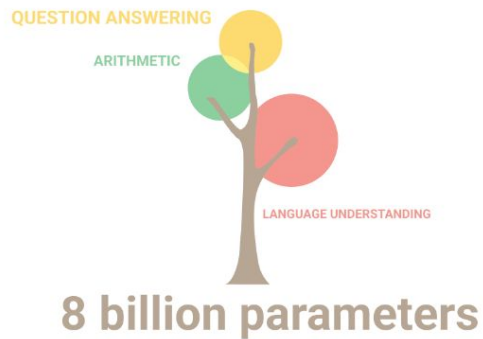
- AI Product Management
- AI Product Strategies, Roadmaps & Execution
- Leverage domain knowledge to create PRDs for AI-integration in products.
- Assignment: Develop a concept for a tool that uses generative AI to create product descriptions for e-commerce websites.

- Nature of the Task: What are you trying to generate?
- Complexity of the Task: Does it need to be broken into segments?
- Data Availability: How recent/prevalent is the data for what you are trying to do?
- Ethical Considerations: Use ethical guidelines to avoid harmful or biased content.
- Human Review/Monitoring: Human oversight is needed to ensure no errors or biases are present.
- Scalability: Assess if the task can be handled efficiently with available computational resources.

Large Language Models (LLMs)

*A large language model is a type of artificial intelligence algorithm that uses deep learning techniques and massively large data sets to **understand, summarize, generate and predict** new content.*

- **Understanding of Transformers Architecture:**
 - Encoder, Decoder, Attention, Self Attention.
 - Its better to have knowledge of Sequence- Sequence and RNN models so one can truly appreciate Transformers Architecture
- **Understanding of available models, their architecture, datasets used to train and their pro's and con's :**
 - Autoencoding models such as BERT & ROBERTA that are good for Sentiment Analysis, NER & Word Classification
 - Autoregressive models such as GPT & BLOOM are good for Text Generation and translations.
 - Seq-Seq models such has T5 and BART are good for Translation, Summarization and Q&A



- Important to consider this while selecting model so you can fine tune/custom train it within your budget, resource & time constraints.
- Ex: To train a 1B parameter model, we would need approx. (80 GB GPU RAM) if we don't use Quantization and other techniques for reducing the model size.
- Many of large LLM's are in 100B parameters size
- Size directly correlates with the computational burden. Larger models need more memory for storing parameters and intermediate computations, especially during training/fine-tuning, where backpropagation requires gradient storage. This translates to higher GPU/TPU requirements

Prompts involve instructions and context passed to a language model to achieve a desired task

What is prompt engineering?

Prompt engineering is a process of creating a set of prompts, or questions, that are used to guide the user toward a desired outcome. It is an effective tool for designers to create user experiences that are easy to use and intuitive. This method is often used in interactive design and software development, as it allows users to easily understand how to interact with a system or product..

- *Important for research, discoveries, and advancement*
- *Helps to test and evaluate the limitations of LLMs*
- *Enables all kinds of innovative applications on top of LLMs*

- The sky is

—

- Classify the text into neutral, negative or positive.
- Text: I think the economy is OK.
- Sentiment:

The odd numbers in this group add up to an even number: 4, 8, 9, 15, 12, 2, 1.

A: The answer is False.

The odd numbers in this group add up to an even number: 17, 10, 19, 4, 8, 12, 24.

A: The answer is True.

The odd numbers in this group add up to an even number: 16, 11, 14, 4, 8, 13, 24.

A: The answer is True.

The odd numbers in this group add up to an even number: 17, 9, 10, 12, 13, 4, 2.

A: The answer is False.

The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.

A: The answer is True.

Prompting | Chain-of-Thought (CoT) Prompting

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

The odd numbers in this group add up to an even number: 4, 8, 9, 15, 12, 2, 1.

A: Adding all the odd numbers (9, 15, 1) gives 25. The answer is False.

The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.

A: Adding all the odd numbers (15, 5, 13, 7, 1) gives 41. The answer is False.

- Text Summarization
- Question Answering
- Text Classification
- Role Playing
- Code Generation
- Reasoning

Prompting | Text Summarization

Antibiotics are a type of medication used to treat bacterial infections. They work by either killing the bacteria or preventing them from reproducing, allowing the body's immune system to fight off the infection.

→ Antibiotics are usually taken orally in the form of pills, capsules, or liquid solutions, or sometimes administered intravenously. They are not effective against viral infections, and using them inappropriately can lead to antibiotic resistance.

→ Explain the above in one sentence:

Antibiotics are medications used to treat bacterial infections by either killing the bacteria or stopping them from reproducing, but they are not effective against viruses and overuse can lead to antibiotic resistance.

Answer the question based on the context below. Keep the answer short and concise. Respond "Unsure about answer" if not sure about the answer.

Context: Teplizumab traces its roots to a New Jersey drug company called Ortho Pharmaceutical. There, scientists generated an early version of the antibody, dubbed OKT3. Originally sourced from mice, the molecule was able to bind to the surface of T cells and limit their cell-killing potential. In 1986, it was approved to help prevent organ rejection after kidney transplants, making it the first therapeutic antibody allowed for human use.

Question: What was OKT3 originally sourced from?

Answer: Mice.

Context obtained from: <https://www.nature.com/articles/d41586-023-00400-x>

The following is a conversation with an AI research assistant. The assistant tone is technical and scientific.

Human: Hello, who are you?

AI: Greeting! I am an AI research assistant. How can I help you today?

Human: Can you tell me about the creation of blackholes?

AI: Sure! Black holes are regions of spacetime where the gravitational force is so strong that nothing, not even light, can escape from it. They are created when a very massive star dies and its core collapses in on itself, forming a singularity of infinite density. The intense gravity of the singularity pulls in all the matter and radiation around it, creating the black hole.

```
"""
Table departments, columns = [DepartmentId,
DepartmentName]
Table students, columns = [DepartmentId, StudentId,
StudentName]
Create a MySQL query for all students in the Computer
Science Department
"""

SELECT StudentId, StudentName
FROM students
WHERE DepartmentId IN (SELECT DepartmentId FROM
departments WHERE DepartmentName = 'Computer Science');
```

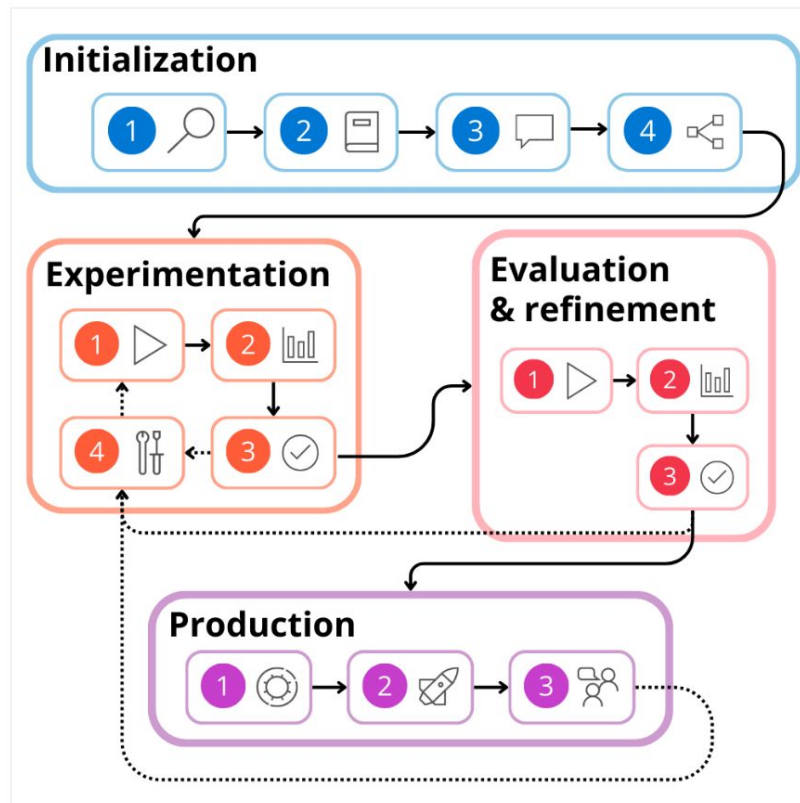

The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.

Solve by breaking the problem into steps. First, identify the odd numbers, add them, and indicate whether the result is odd or even.

Odd numbers: 15, 5, 13, 7, 1

Sum: 41

41 is an odd number.



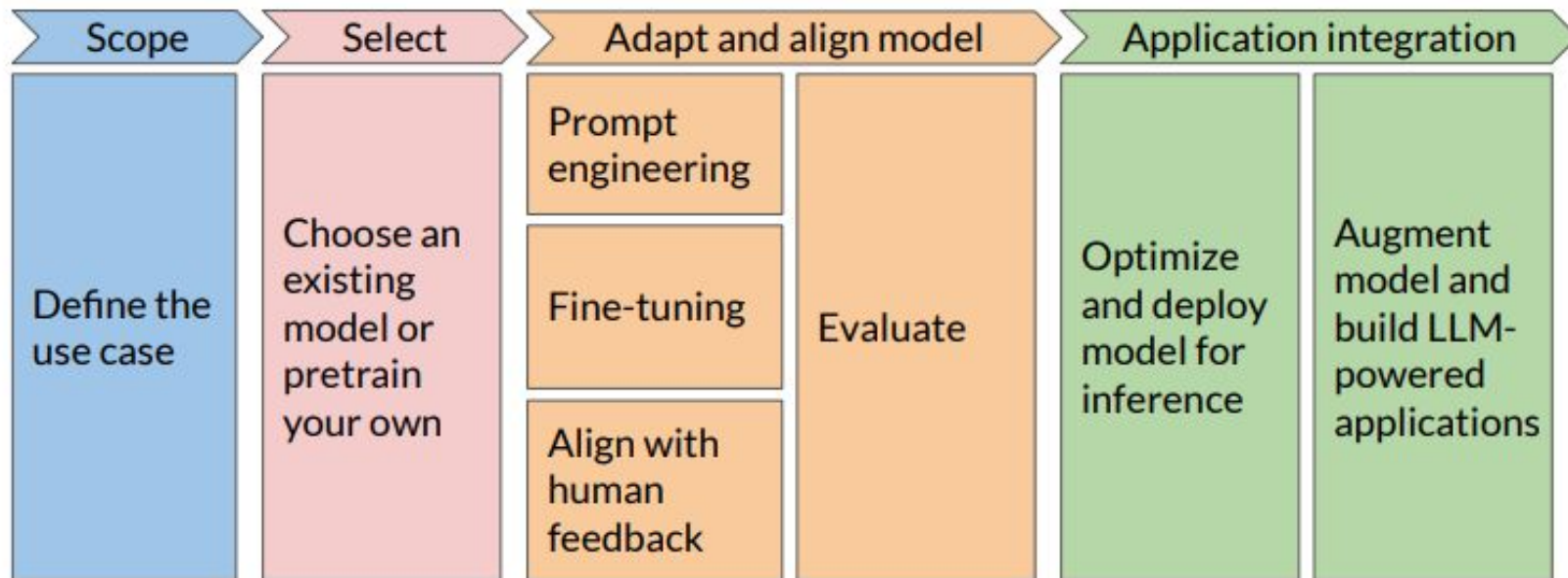
- The developer defines the objective of the LLM application, collects a sample dataset, builds a basic prompt, and designs the flow.
- The objective is the desired outcome or goal of the LLM application, such as summarizing news articles.
- The sample dataset is a small set of representative data that can be used to test the basic functionality of the LLM application.
- The prompt is the input format and structure that the LLM application expects and the output format and structure that the LLM application produces.
- The flow is the sequence of steps and components that the LLM application follows to process the input and generate the output.

- The developer tests the flow with a small dataset to evaluate its performance.
- The performance can be measured by various metrics, such as accuracy, precision, recall, or F1-score, depending on the task and the objective.
- This stage is an iterative process that involves running the flow against a sample dataset, evaluating its performance, and modifying the flow if necessary.
- The modifications can include changing the prompt, adding or removing components, adjusting parameters, or fine-tuning the LLM.

- In this stage, the developer assesses the flow with a larger dataset to evaluate how well the LLM application generalizes to new data.
- The larger dataset should cover a wider range of scenarios and variations that the LLM application might encounter in the real world.
- The Evaluation and Refinement stage is important for identifying potential bottlenecks or areas for optimization or refinement.
- For example, the developer might discover that the LLM application is biased towards certain categories, or that it fails to handle certain edge cases or exceptions.

- In this stage, the developer optimizes the flow for efficiency and effectiveness, deploys the LLM application to an endpoint, and monitors its performance.
- The optimization can include reducing the latency, increasing the throughput, or improving the quality of the LLM application.
- The deployment can involve hosting the LLM application on a cloud platform, a web server, or a mobile device.
- The Production stage is the ultimate goal of the development lifecycle, where the LLM application is ready to serve its intended purpose and audience.

Lifecycle of LLM Application Development | Flowchart

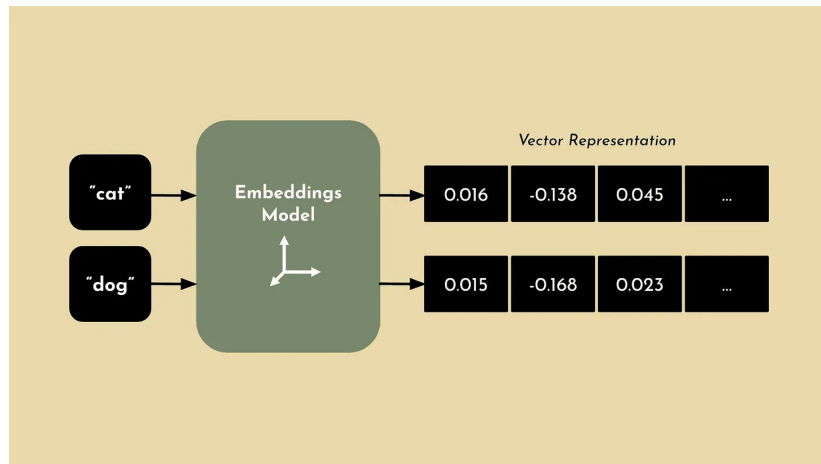


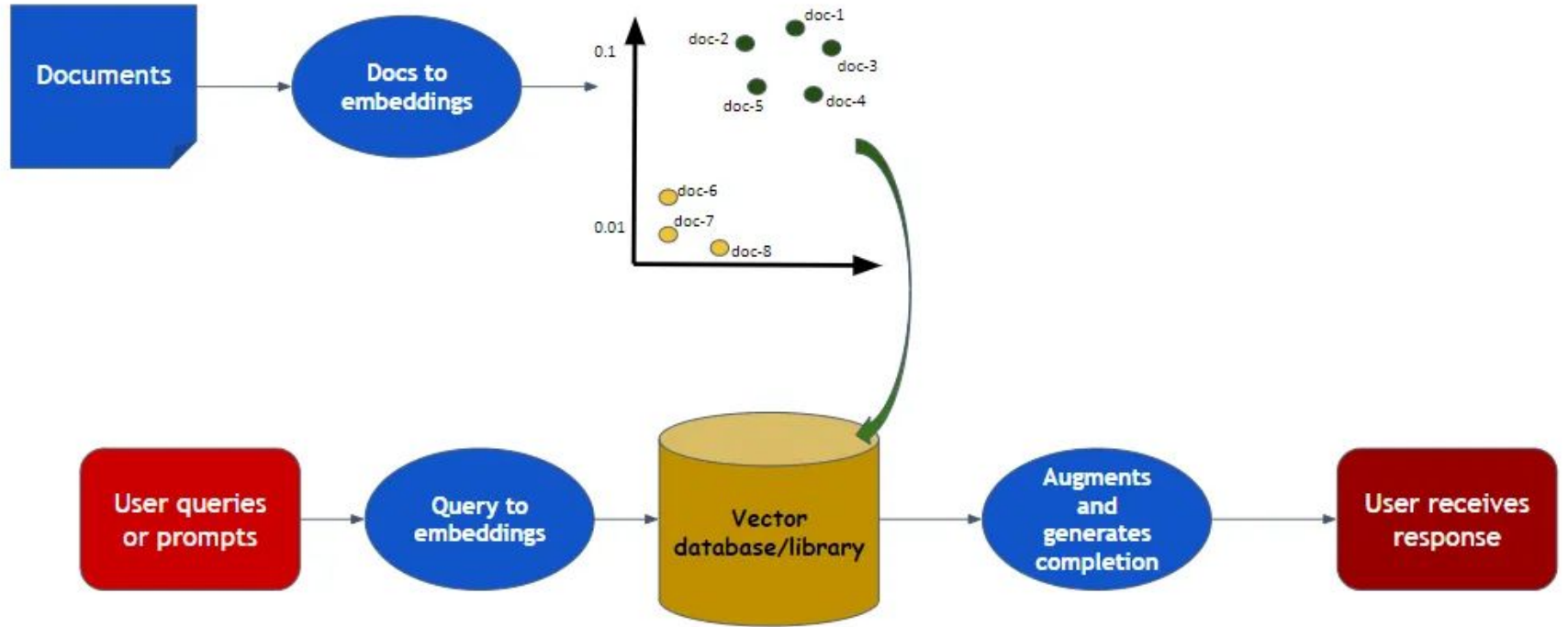
- Python
- Google Colab
- pynytimes library: <https://pypi.org/project/pynytimes/>
- NewsAPI: <http://newsapi.org>
- OpenAI API: <https://platform.openai.com/docs/api-reference/authentication>

Demo # 1

News Summarization (NYT API)

- Embeddings are vectors generated by neural networks, representing data in vector databases.
- They are numerical representation that capture the essential features and relationship of discrete objects like words or documents
- These embeddings capture the essence of the text in a format that the LLM can process—turning discrete tokens (like words or sentences) into a continuous vector space.





Demo #2

**News Summarization -
Retrieval and Embedding**

- Where not to use AI (Limitations):
 - AI is fast. But it lacks the creativity, experience, and wisdom that professional writers or designers have.
 - As a subject matter expert
 - To draw conclusions or make decisions based on complex situations — something that only humans can do at present.
 - Where hallucinations are intolerable
- Ethical considerations
 - Misinformation and Deepfakes
 - Bias and Discrimination
 - Copyright and Intellectual Property
 - Privacy and Data Security