Applied GenAI Workshop: LectureBot

Building you first LLM powered Application

Introduction – Naveen Neppalli

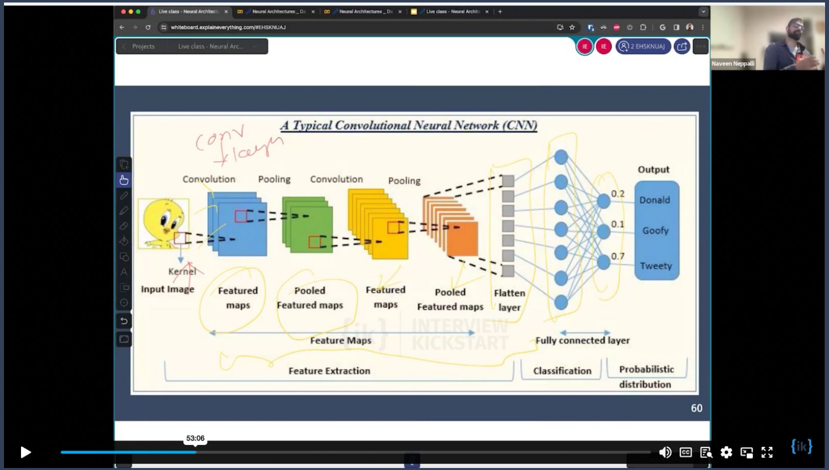
- VP of AI & Engineering at Vouched
- Head of Applied Science Engineering, Amazon Private brands Product discovery
- Lives in Seattle, WA.
- Masters in CS from Rutgers
- Experiences: NLP (LLM, Recommendations, Ads), Computer Vision (Face recognition, Object detection/segmentation)
- 17 Years in various ML roles - IC, Lead, Manager, Director, VP
- https://www.linkedin.com/in/naveen-neppalli/

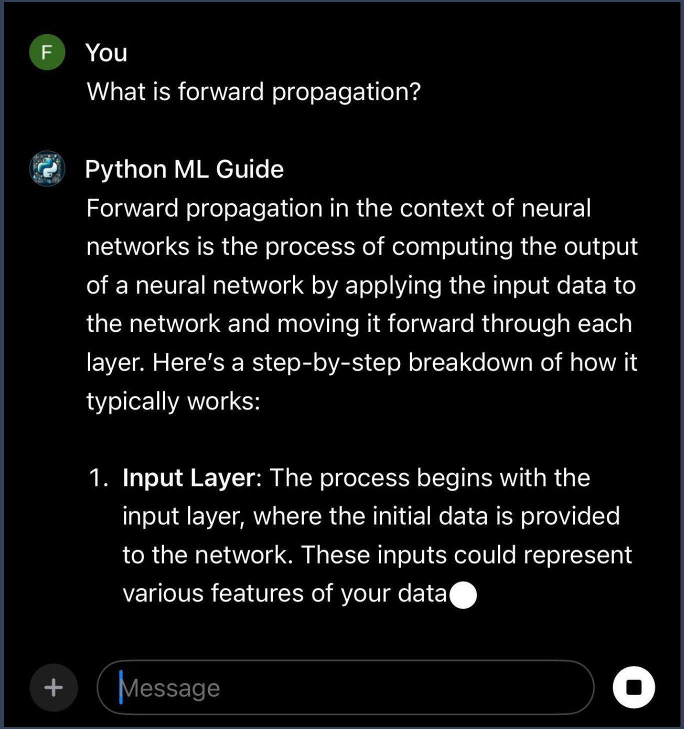
Applied Gen AI Workshop | Agenda

- ➢ Gen AI Tech Stack
- ➢ Problem Statement and Motivation
- ➢ RAG Architecture
- ➢ Why RAG?
- ➢ How to build a RAG workflow?
- ➢ Live Demo: LectureBot
- ➢ Q&A
- ➢ Overview about IK program

Gen AI Application - LectureBot | Motivation

Given a lecture video as input, can we convert it to a chabot so learners can ask questions about the lecture without watching the entire lecture?





F You

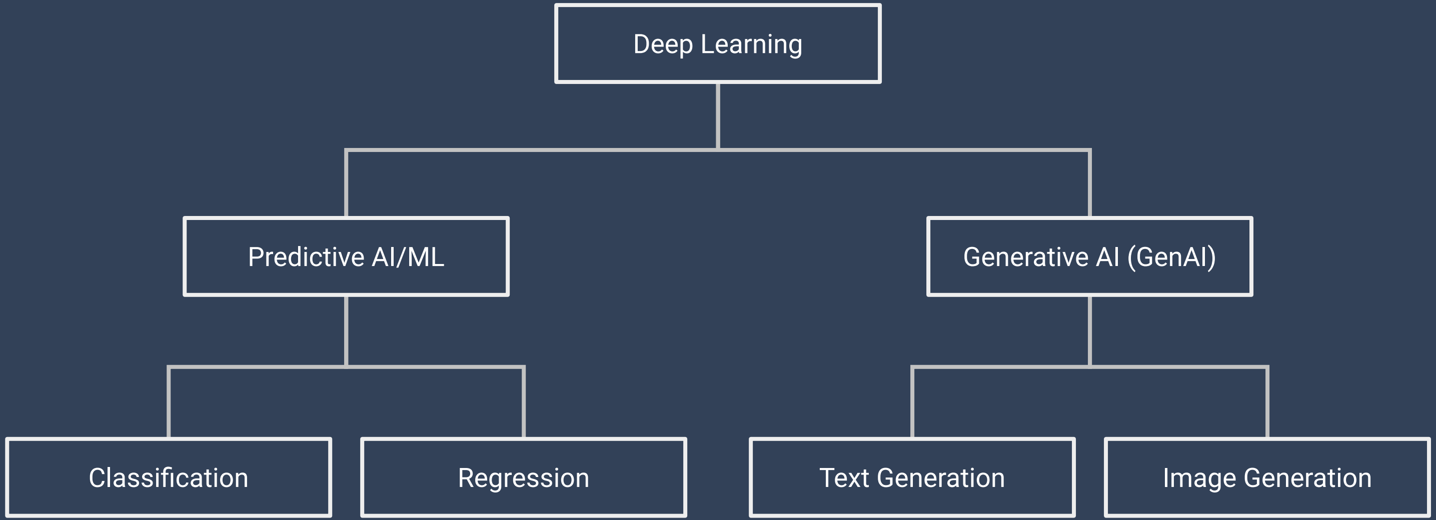
What is forward propagation?

Python ML Guide

Forward propagation in the context of neural networks is the process of computing the output of a neural network by applying the input data to the network and moving it forward through each layer. Here's a step-by-step breakdown of how it typically works:

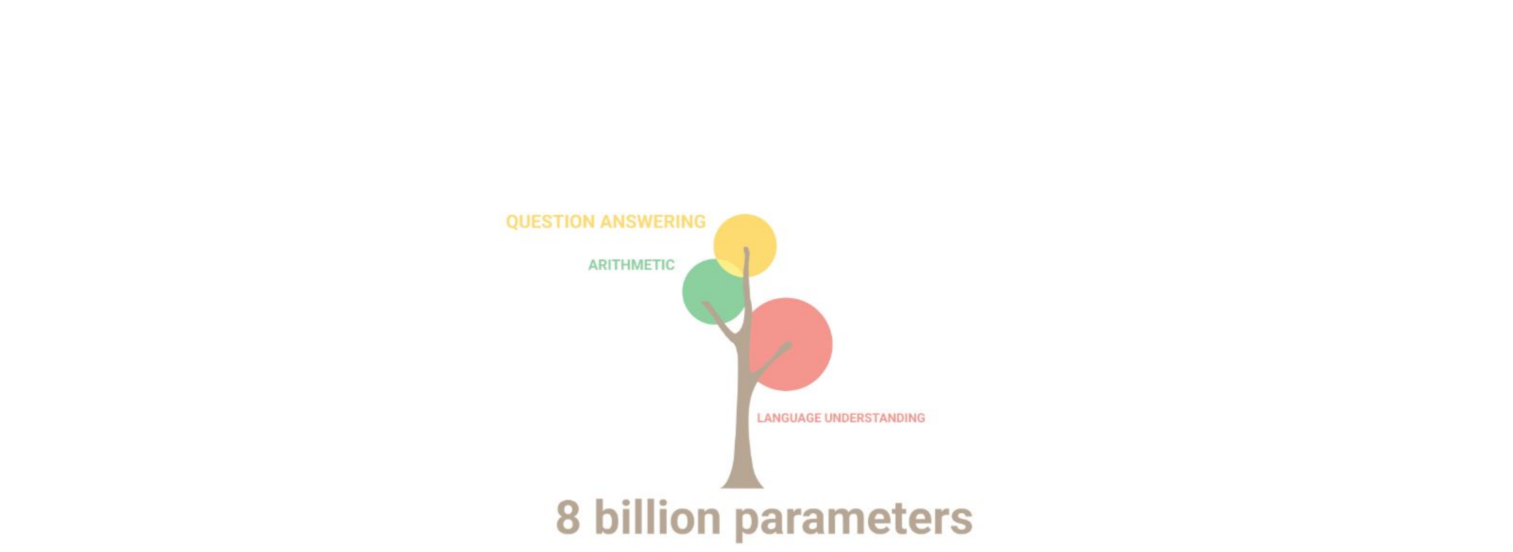
1. **Input Layer**: The process begins with the input layer, where the initial data is provided to the network. These inputs could represent various features of your data

What is GenAI? | What’s so special?



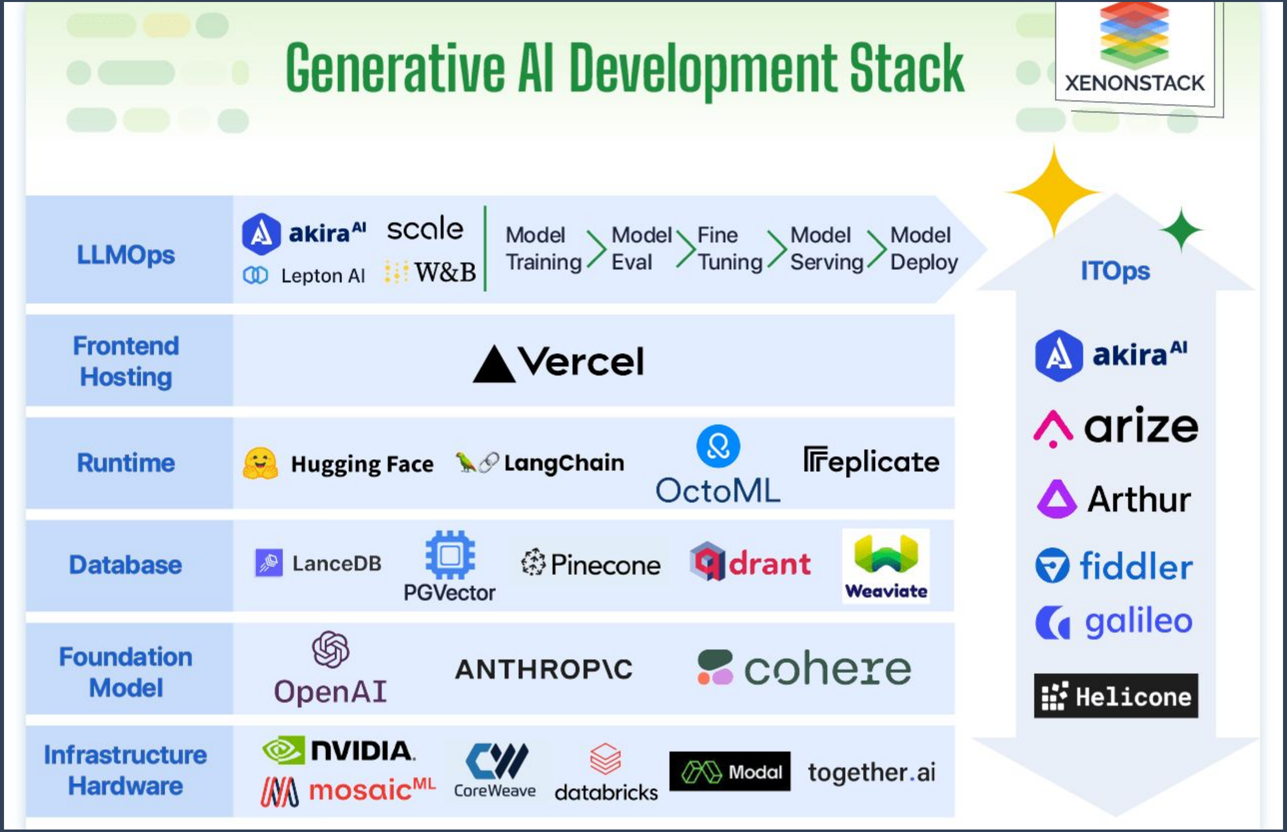
Generative AI (Gen AI): Refers to deep-learning models that can generate high-quality text, images, and other content.

GenAI - LLMs | What can they do?



Credit: superwise.ai

Gen AI Tech Stack | The new growth engine



Generative AI Development Stack

XENONSTACK

LLMOps	akira AI, scale, Lepton AI, W&B Model Training > Model Eval > Fine Tuning > Model Serving > Model Deploy
Frontend Hosting	Vercel
Runtime	Hugging Face, LangChain, OctoML, replicate
Database	LanceDB, PGVector, Pinecone, qdrant, Weaviate
Foundation Model	OpenAI, ANTHROPIC, cohere
Infrastructure Hardware	NVIDIA, mosaicML, CoreWeave, databricks, Modal, together.ai

ITOps

- akira AI
- arize
- Arthur
- fiddler
- galileo
- Helicone

Source: https://www.xenonstack.com/blog/generative-ai-tech-stack-breakdown

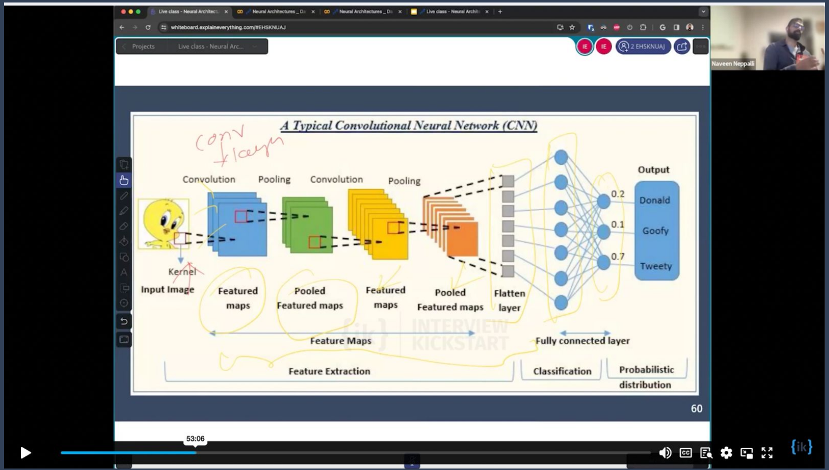
Objective

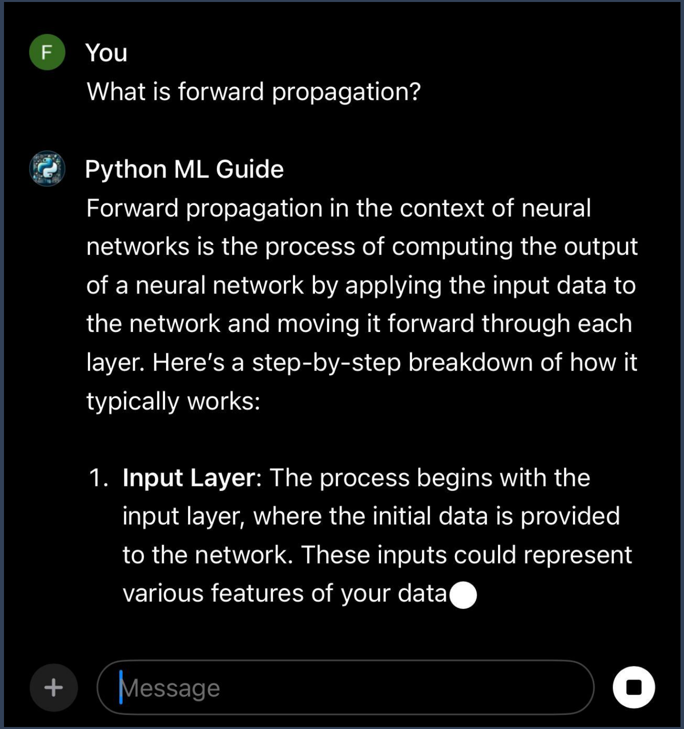
Let’s build our first GenAI Application

LectureBot

Gen AI Application - LectureBot | Motivation

Given a lecture video as input, can we convert it to a chabot so learners can ask questions about the lecture without watching the entire lecture?





F You

What is forward propagation?

Python ML Guide

Forward propagation in the context of neural networks is the process of computing the output of a neural network by applying the input data to the network and moving it forward through each layer. Here's a step-by-step breakdown of how it typically works:

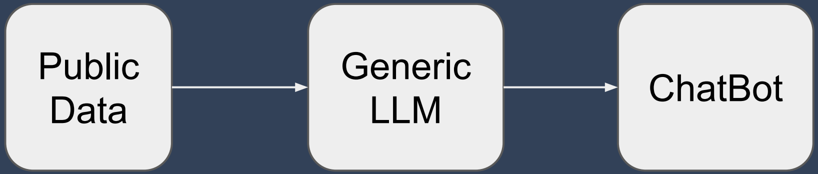
1. **Input Layer**: The process begins with the input layer, where the initial data is provided to the network. These inputs could represent various features of your data

LectureBot | Uses

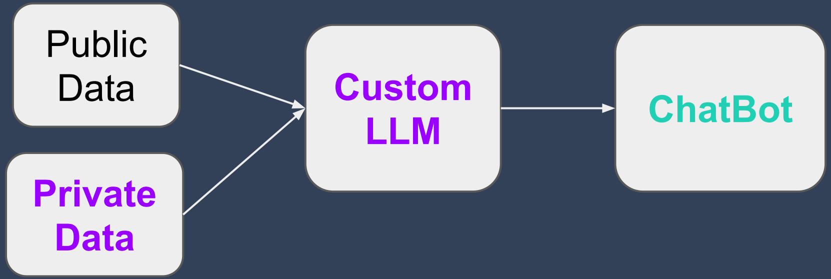
- The LectureBot should analyze and summarize class video data from private sources, and can enable learners to grasp the material using custom prompts.
- Few example prompts that a learner can use to ask questions and receive summarized answers.
  - ‘What are the key topics in the lecture?’
  - ‘Tell me more about topic XYZ’
  - ‘Provide a summarized version of the lecture in less than 3 paragraphs’

How can we build this? | Ideas

- Use a publicly available LLM (gpt4) and build a chatbot - Too generic/No private data

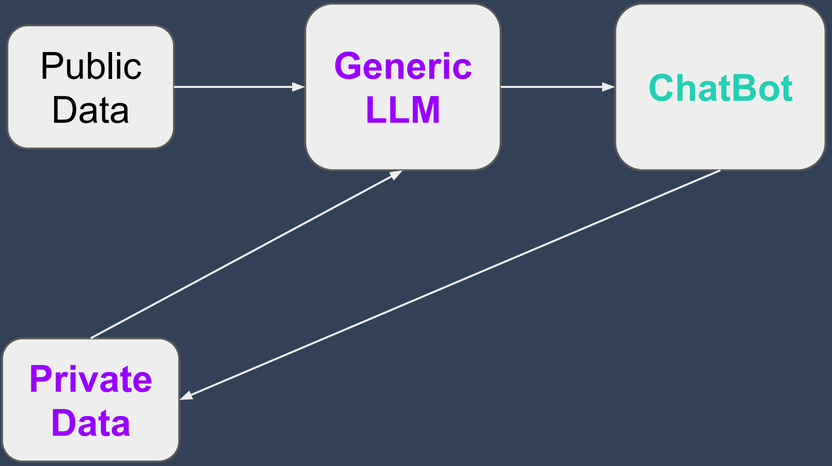


- **Finetune** a publicly available LLM on your data and then build a chatbot - Too expensive, need a lot of data

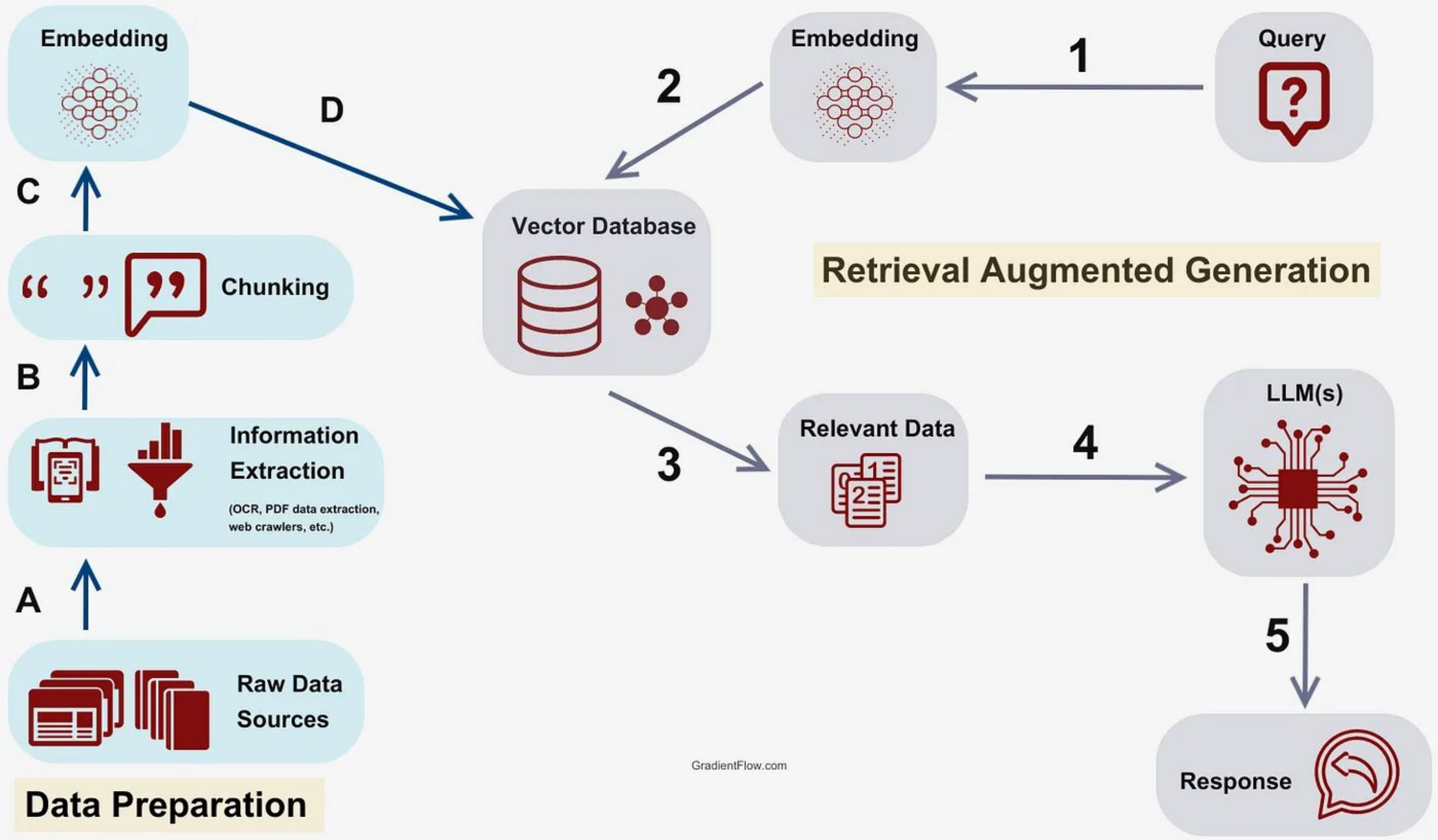


RAG comes to rescue | Retrieval Augmented Generation

What if we do not need to train a Custom LLM on private data rather we can just give it at inference time as context?



RAG Architecture in Practice



Source: https://gradientflow.com/techniques-challenges-and-future-of-augmented-language-models/

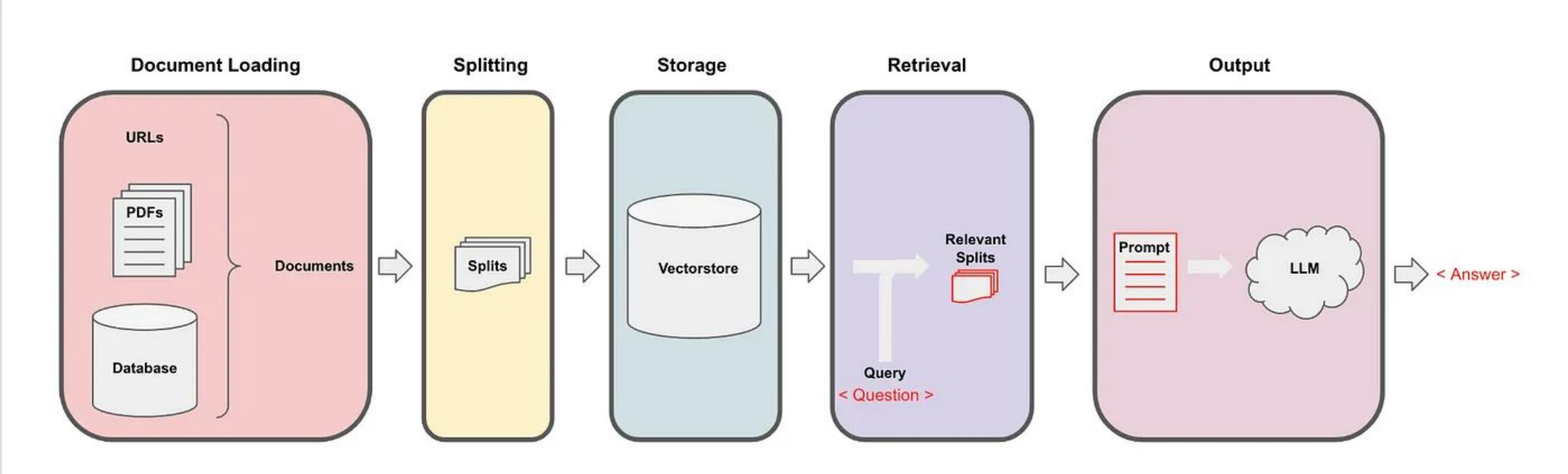
LectureBot Demo | Tools/APIs Used

- Python 3.9
- VSCode (any other code editor works fine)
- AWS EC2 for compute - chunking and generating summaries
- Langchain framework for orchestration
- llama-2-13b-chat.Q4_K_M.gguf as LLM
- Qdrant as vector store
- sentence-transformers/all-mpnet-base-v2 for embedding generation

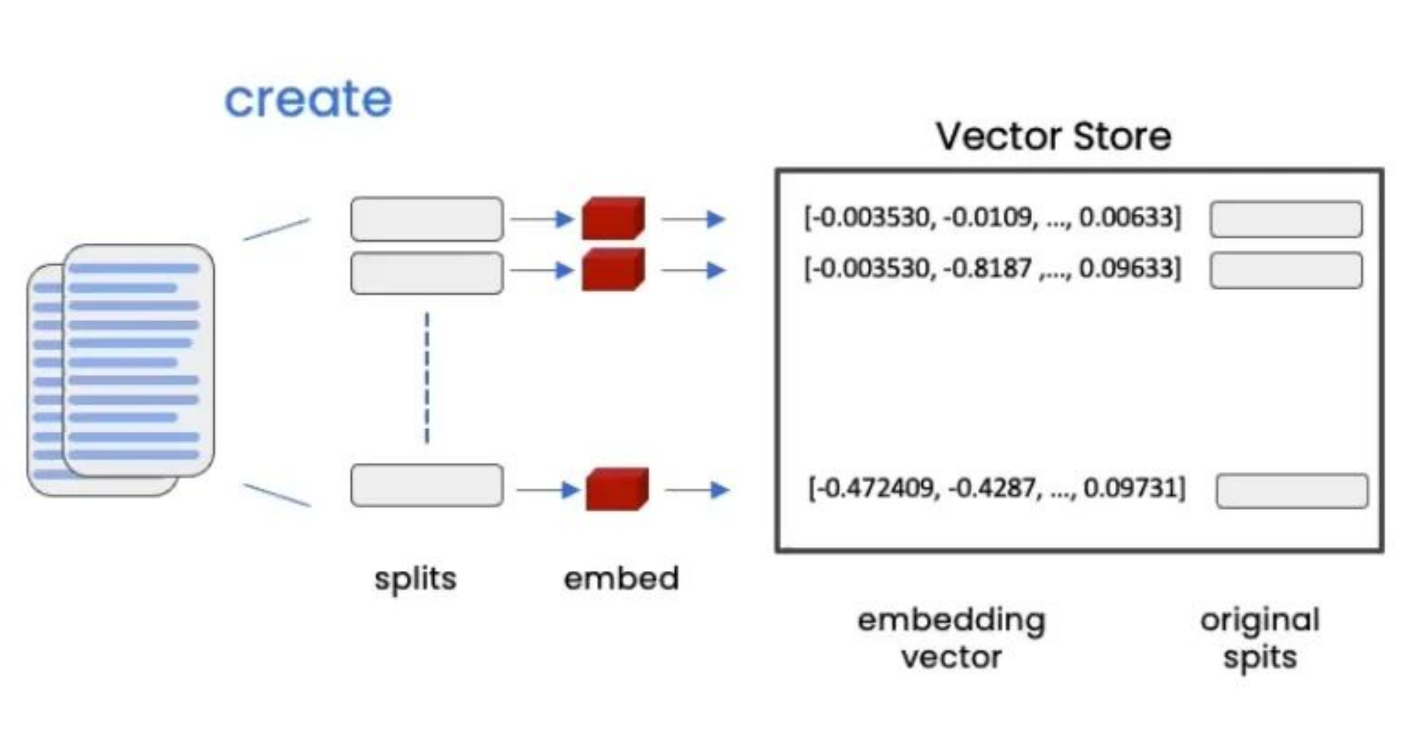
Demo
LectureBot

https://github.com/avrawat-ik/28apr_applied_genAi_workshop_LectureBot

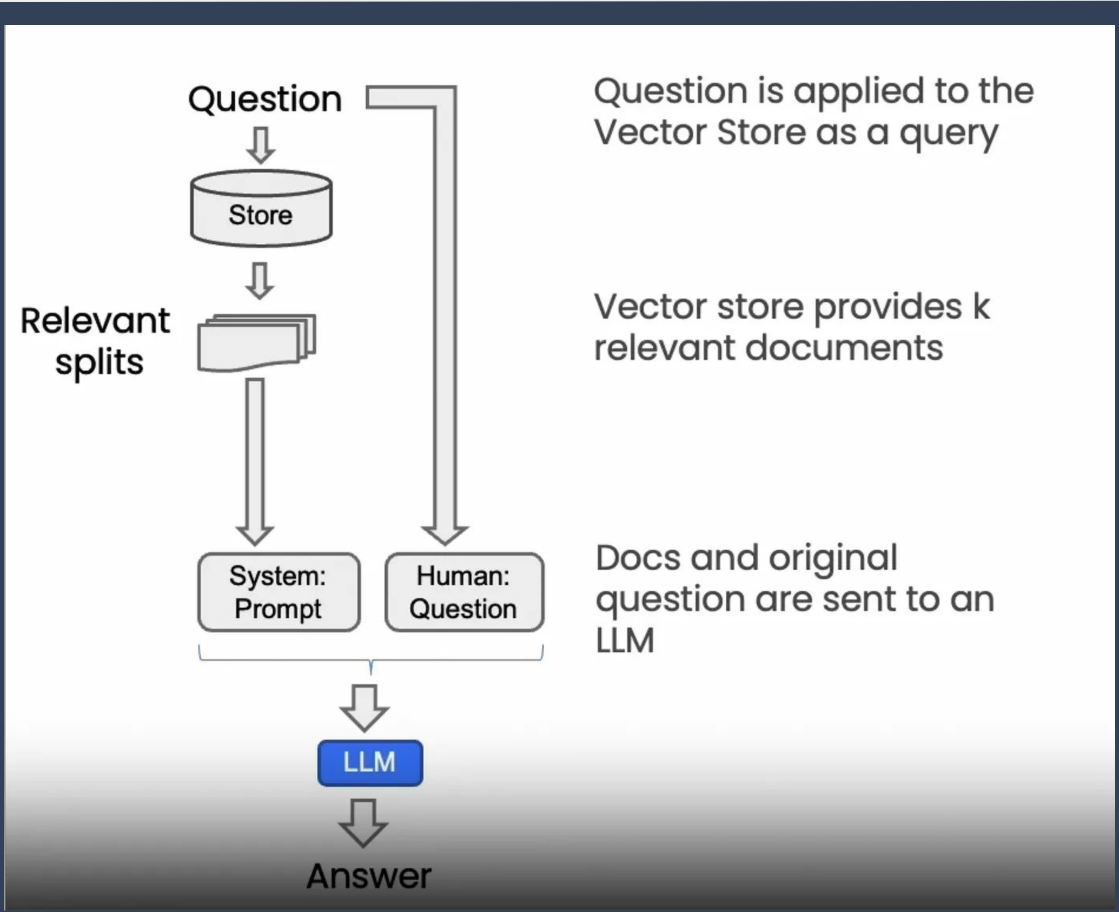
LectureBot | Big Picture



Document processing | Chunking



Document processing | Retrieval

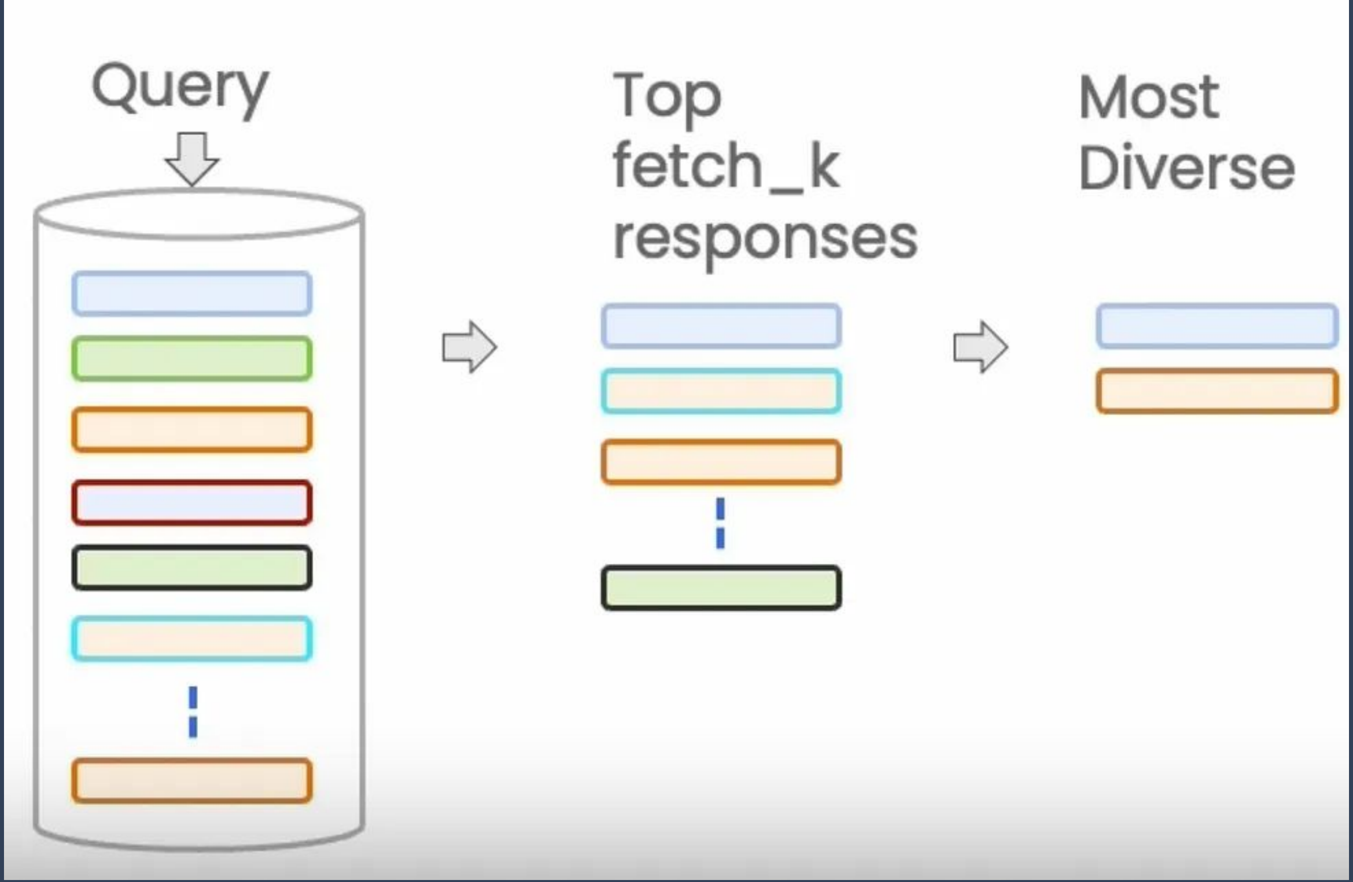


Question is applied to the Vector Store as a query

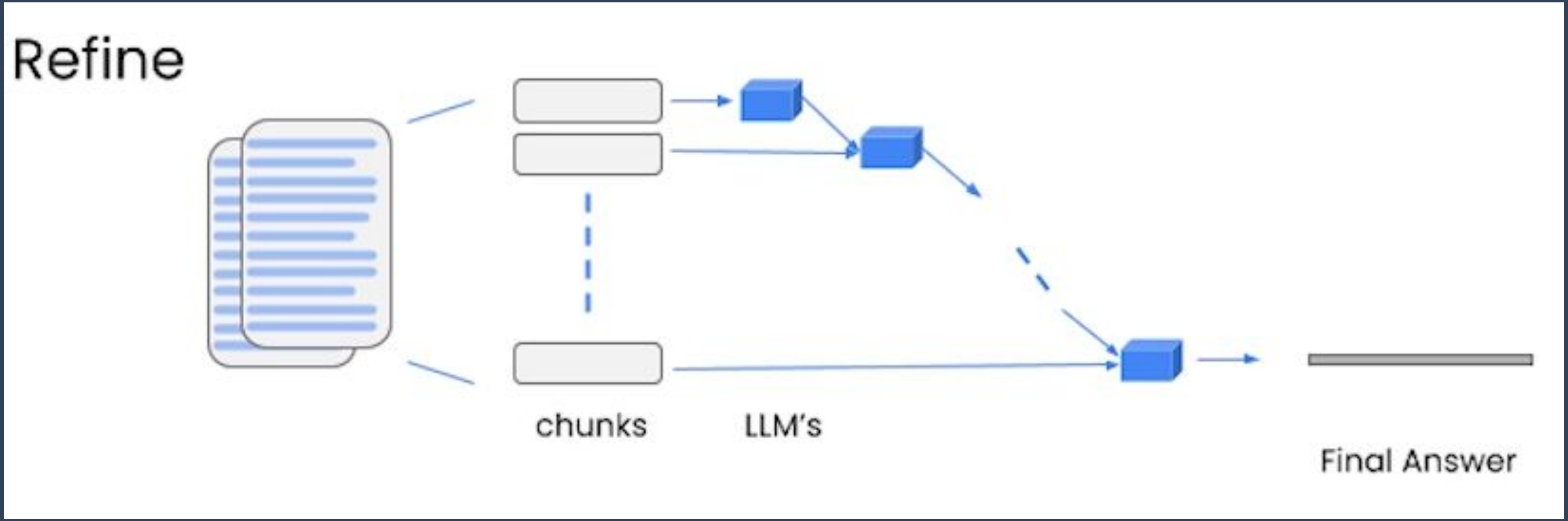
Vector store provides k relevant documents

Docs and original question are sent to an LLM

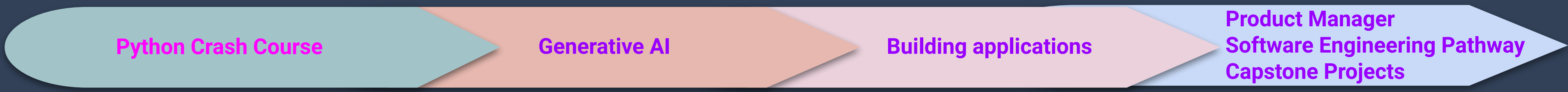
Retrieval Query| MMR Maximum Marginal Relevance



Prompting Context | Stuff Documents vs. Refine Docs



Applied GenAI | Innovate with Gen AI



2 weeks

Python Crash Course

- Python fundamentals
- Python Libraries for Machine Learning

4 weeks

Generative AI

- Hands-on with Generative AI
- Gen AI Background and Neural Networks
- Deep dive into LLMs

4 weeks

Building applications

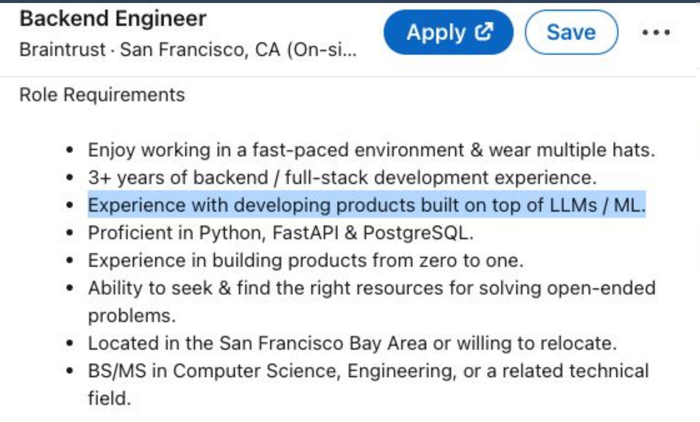
- Building Applications with LLMs
- Training LLMs
- GenAI for Images
- GenAI for Audio

2-4 weeks

Product Manager
Software Engineering Pathway
Capstone Projects

- Separate learning tracks for Product Managers & Software Engineers
- Learn generative AI applications specific to your job roles
- Appealing Capstone Projects for All domains

Software Engineers | Why should they upskill in Generative AI

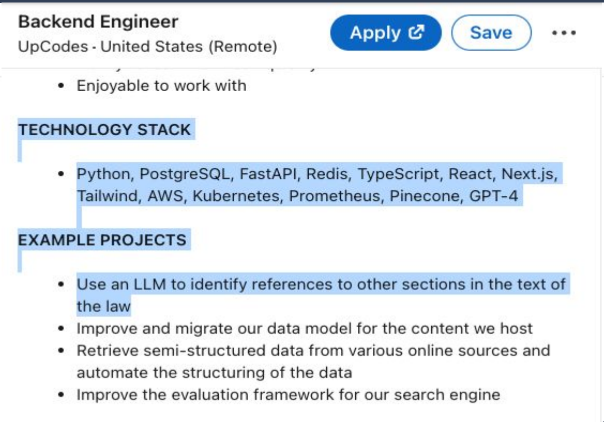


Backend Engineer

Braintrust · San Francisco, CA (On-si...

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)

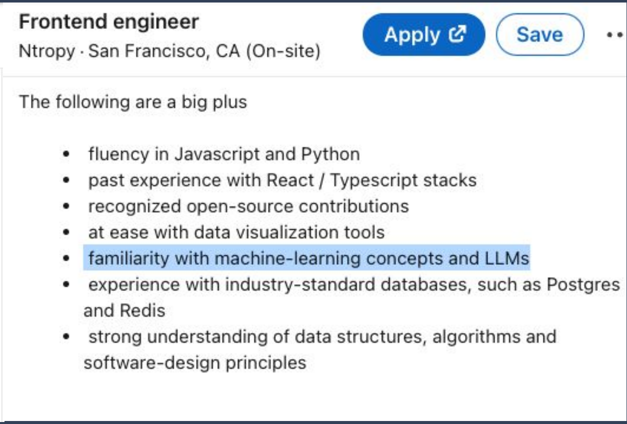
- Enjoyable to work with

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



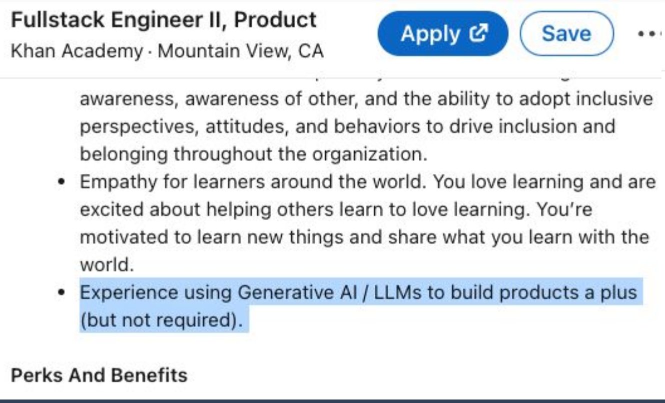
Frontend engineer

Ntropy · San Francisco, CA (On-site)

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



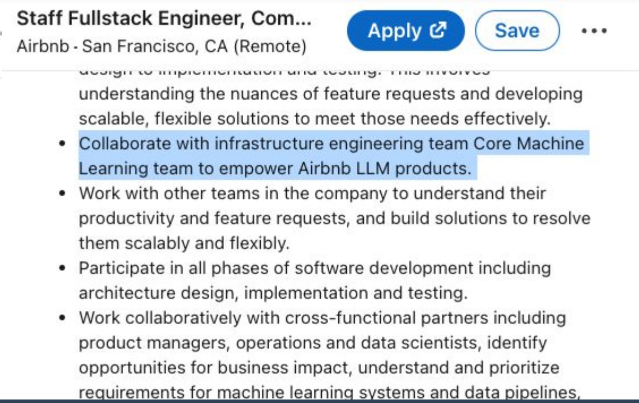
Fullstack Engineer II, Product

Khan Academy · Mountain View, CA

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

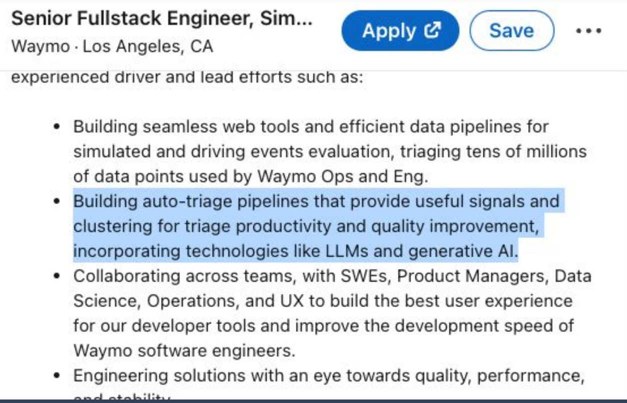


Staff Fullstack Engineer, Com...

Airbnb · San Francisco, CA (Remote)

design to 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,



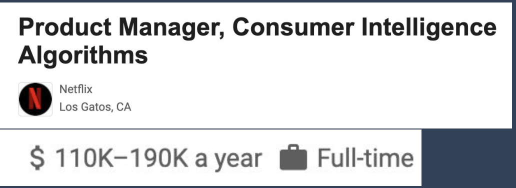
Senior Fullstack Engineer, Sim...

Waymo · Los Angeles, CA

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

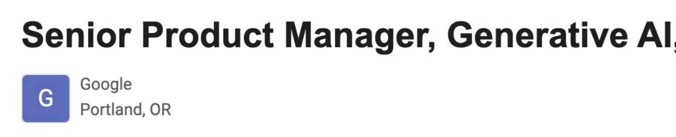
Product Managers | Why should they upskill in Generative AI



Product Manager, Consumer Intelligence Algorithms

Netflix
Los Gatos, CA

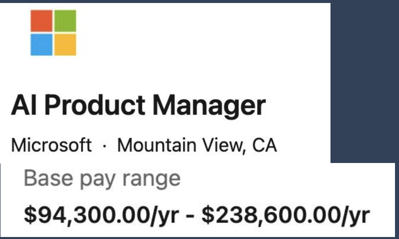
$ 110K–190K a year · Full-time



Senior Product Manager, Generative AI,

Google
Portland, OR

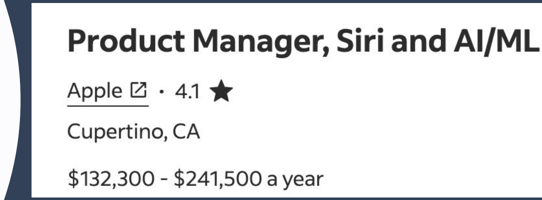
1000+ Openings for AI PM jobs



AI Product Manager

Microsoft · Mountain View, CA

Base pay range
$94,300.00/yr - $238,600.00/yr

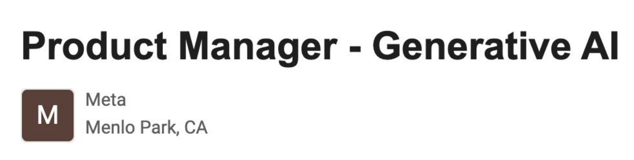


Product Manager, Siri and AI/ML

Apple · 4.1 ★

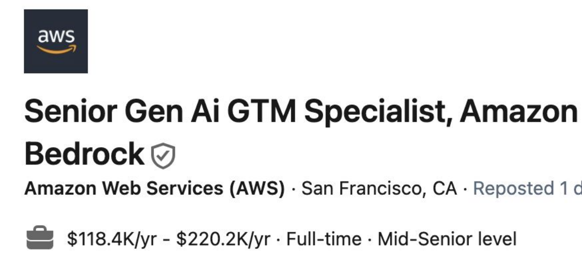
Cupertino, CA

$132,300 - $241,500 a year



Product Manager - Generative AI

Meta
Menlo Park, CA



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

Q&A

Thank you

Potentials of GenAI

- ❑ Low Resource Language - Ability to understand, generate any language especially low resource ones: historical documents

- ❑ Personalized Content Generation - Content creation that can cater to the users and individual interests at scale

- ❑ AI Tutor - Tutor to teach you any skill at your own pace

- ❑ Intelligent Assistants - Laborious and repetitive tedious tasks can be delegated to intelligent AI assistants

- ❑ Accelerating Scientific Discovery - Generative deep insights from massive datasets and design new algorithms