

RAG Hands-on Workshop:

Understanding and Leveraging AI Retrieval in Your Projects

Introduction – Sarfaraz Hussein

Senior ML Researcher at Motional Inc.
(Autonomous Driving)

PhD – Center for Research in Computer Vision @ UCF

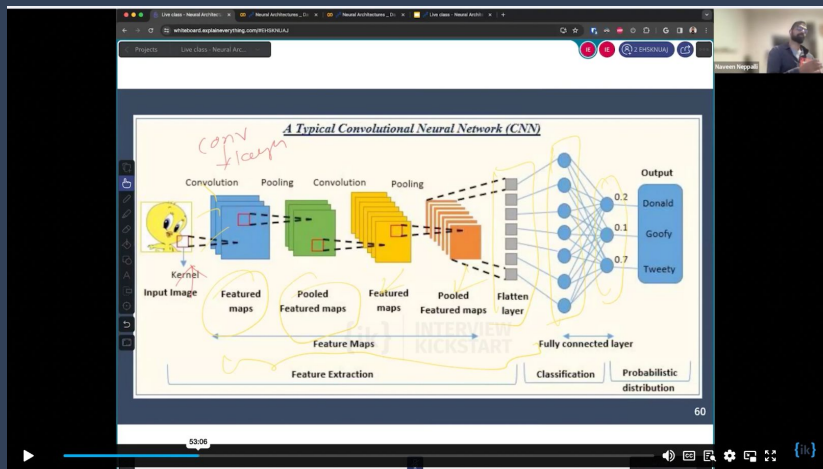
Ex-Amazon, HomeDepot, Symantec, Siemens

Over a decade of Teaching and Research Experience




Gen AI Application - LectureBot | Motivation

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



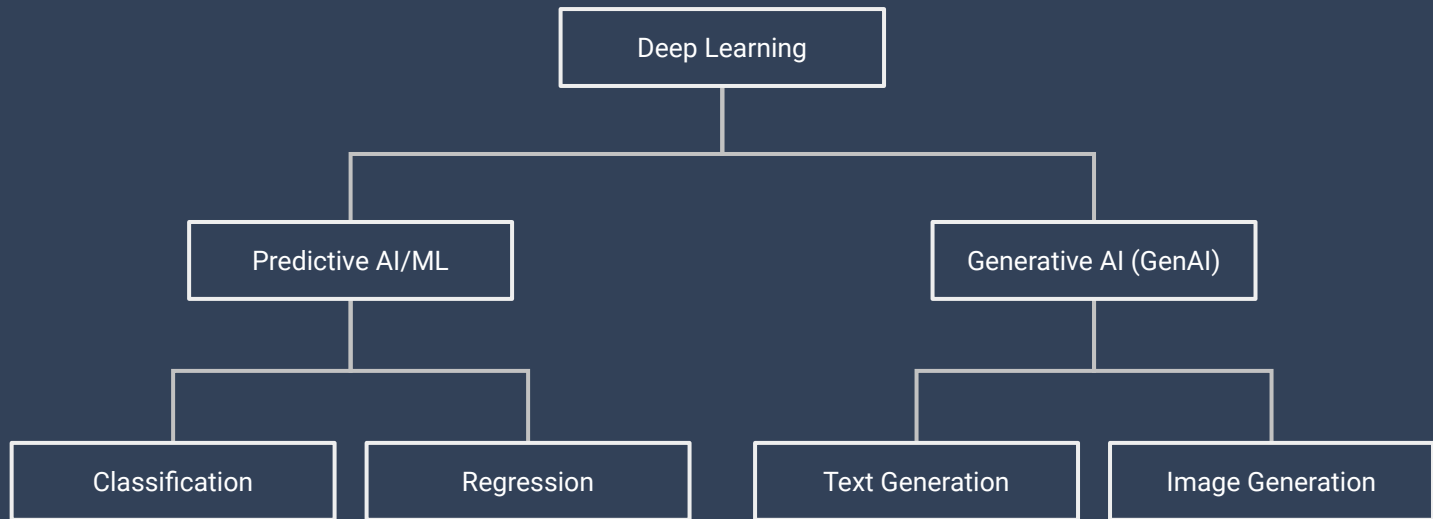
F You
What is forward propagation?

 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 ●

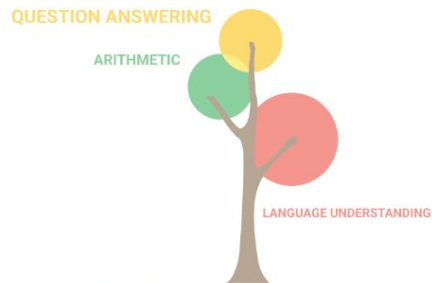
+ Message

What is GenAI? | What's so special?



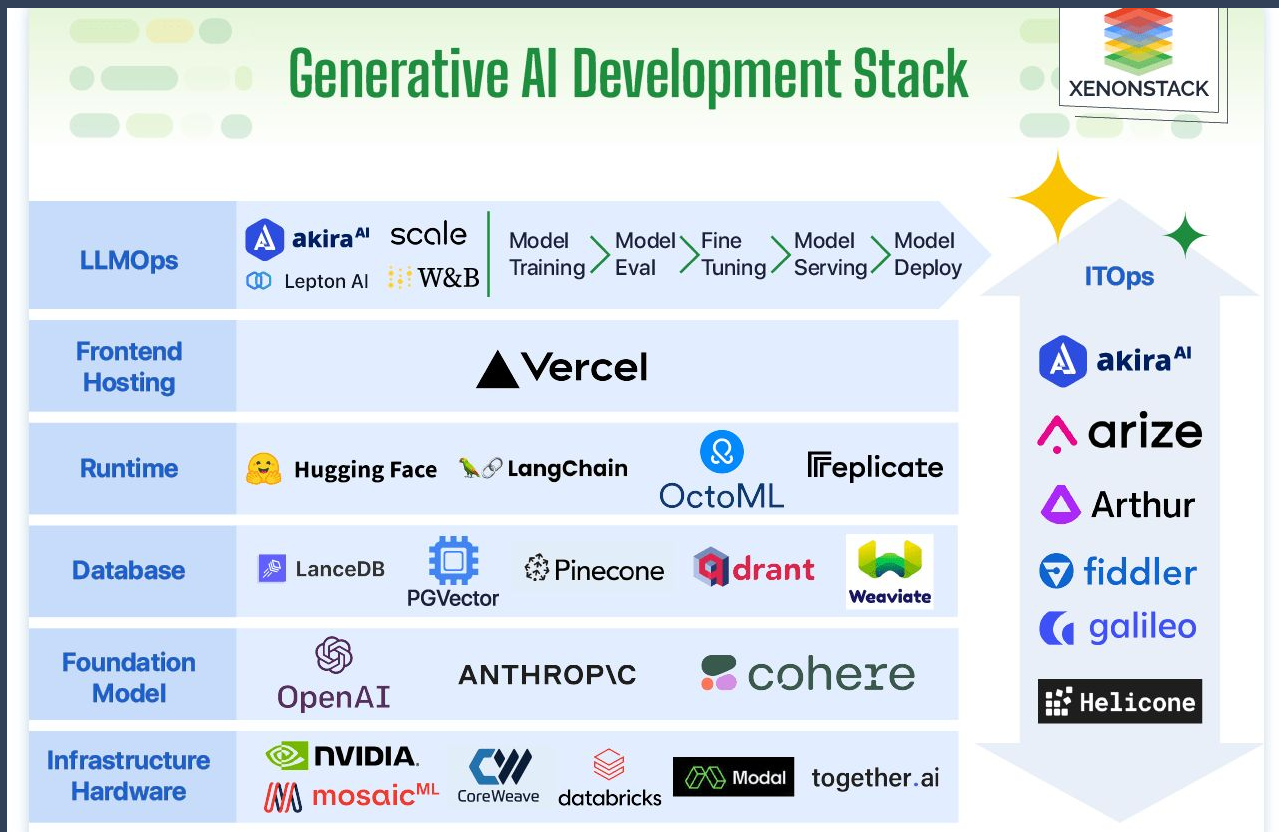
Generative AI (Gen AI): Refers to deep-learning models that can generate high-quality text, images, and other content.

GenAI - LLMs | What can they do?



8 billion parameters

Gen AI Tech Stack | The new growth engine

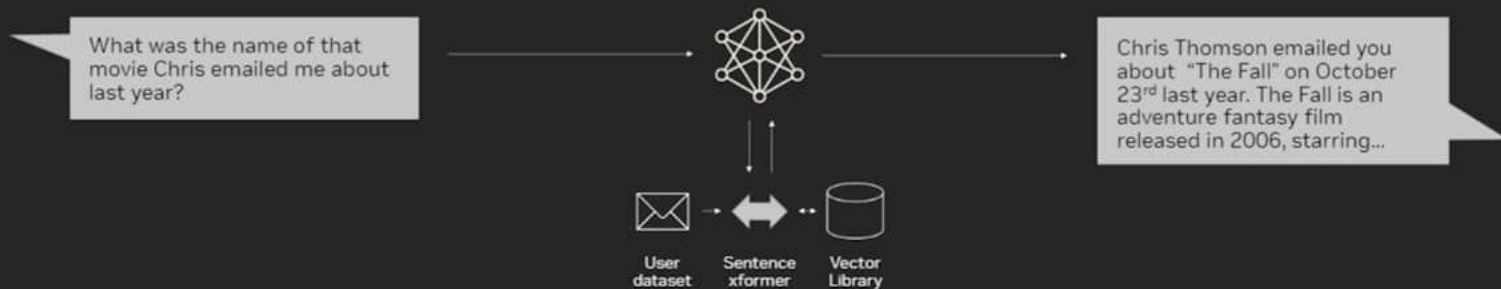


Retrieval-Augmented Generation (RAG)

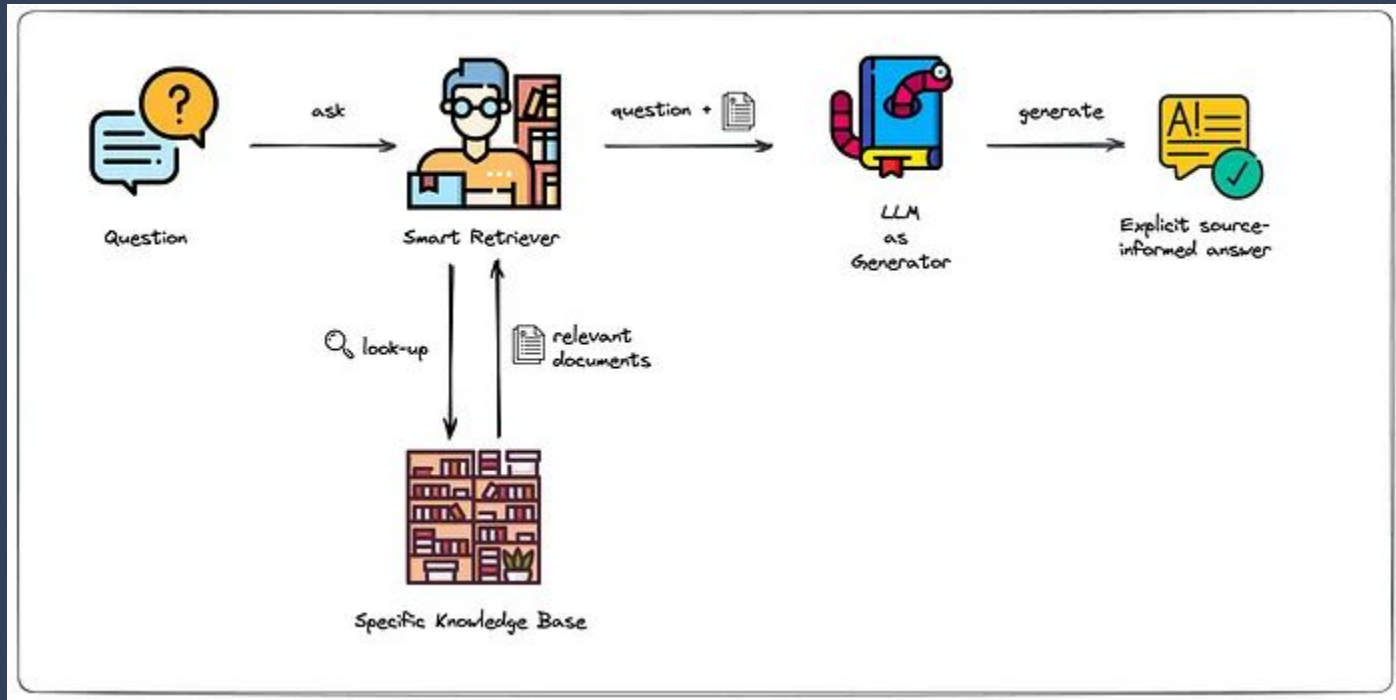
BASE MODEL



BASE MODEL + USER DATASET



Retrieval-Augmented Generation

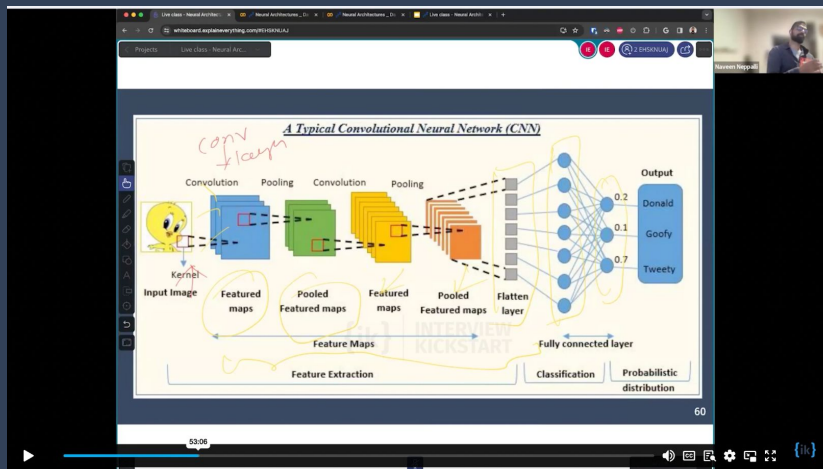


Let's build our first GenAI Application


LectureBot

Gen AI Application - LectureBot | Motivation

Given a lecture video as input, can we convert it to a chatbot 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 ●



Message

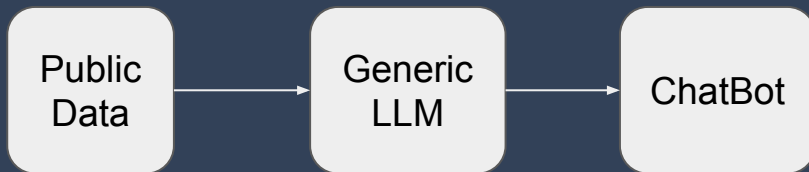


LectureBot | Uses

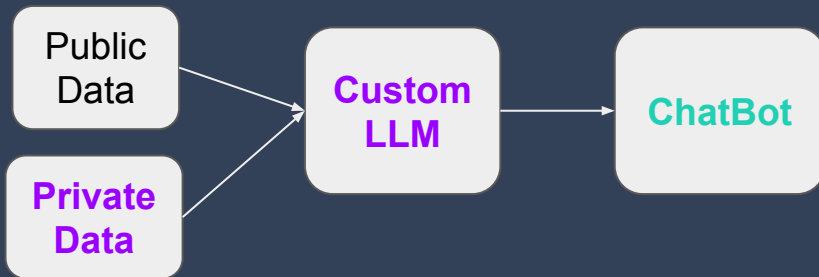
- The LectureBot should analyze and summarize class video data from private sources, and can enable learners to grasp the material using custom prompts.
- Few example prompts that a learner can use to ask questions and receive summarized answers.
 - ‘What are the key topics in the lecture?’
 - ‘Tell 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

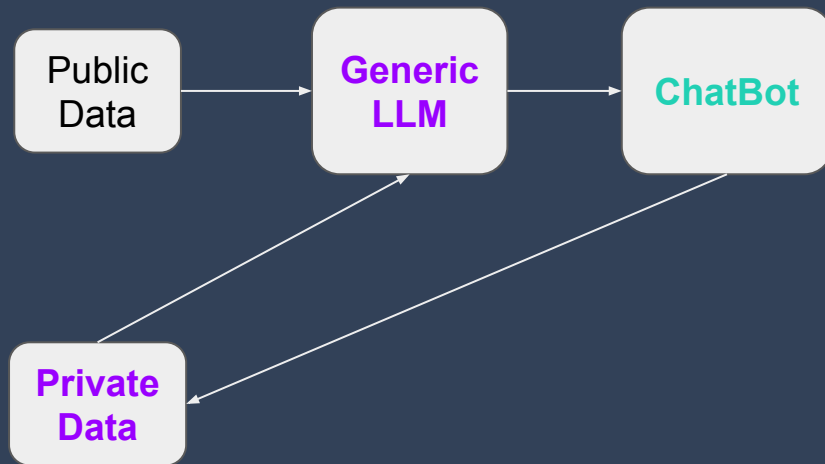


- **Finetune** a publicly available LLM on your data and then build a chatbot - Too expensive, need a lot of data

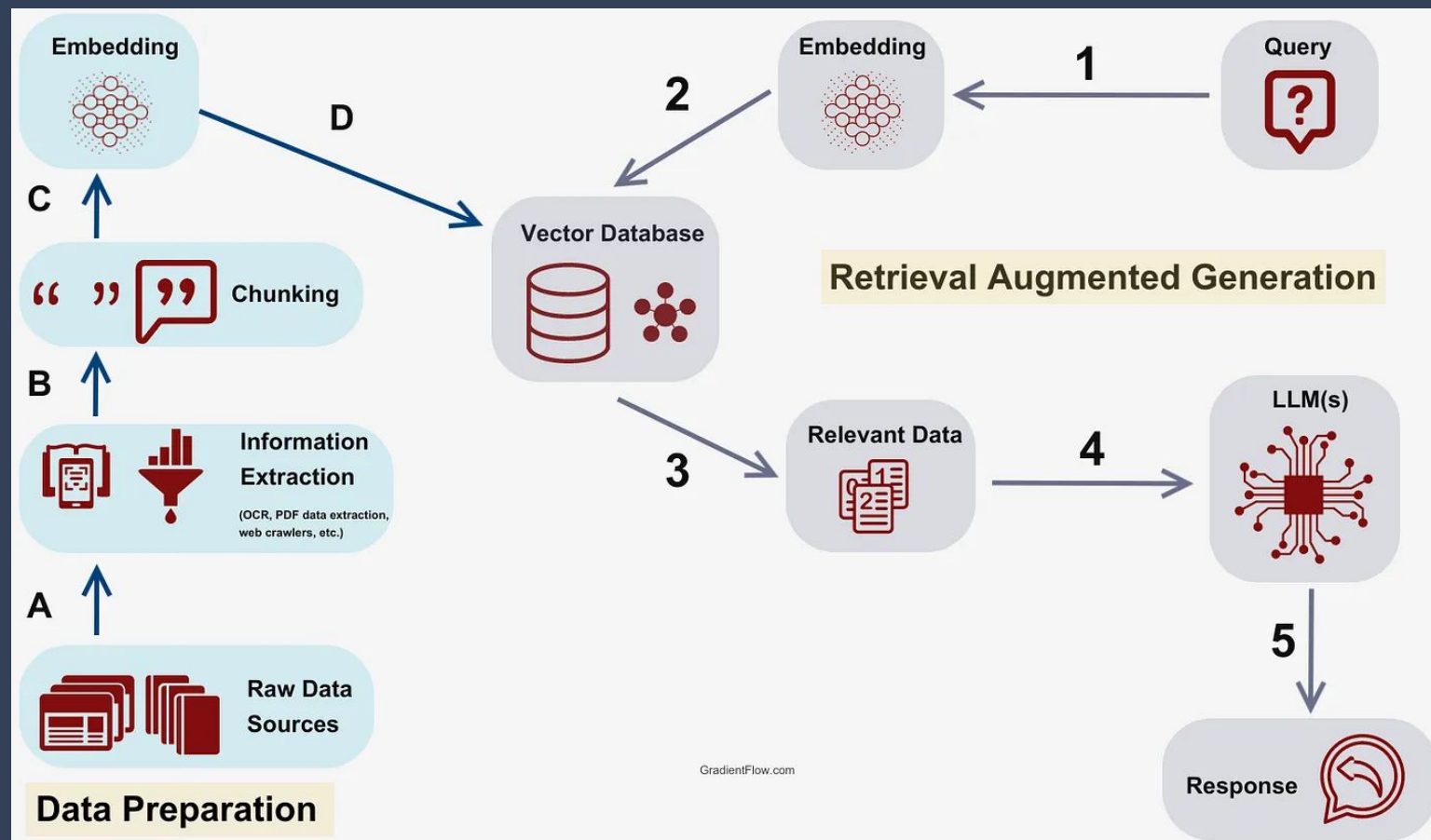


RAG comes to rescue | Retrieval Augmented Generation

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?



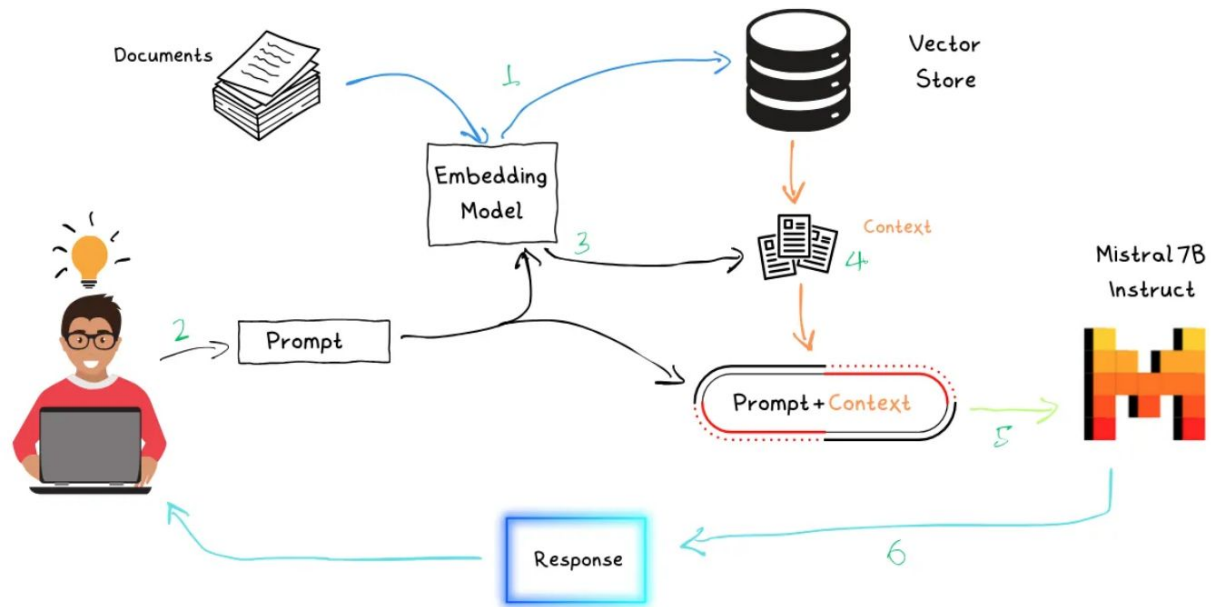
RAG Architecture in Practice

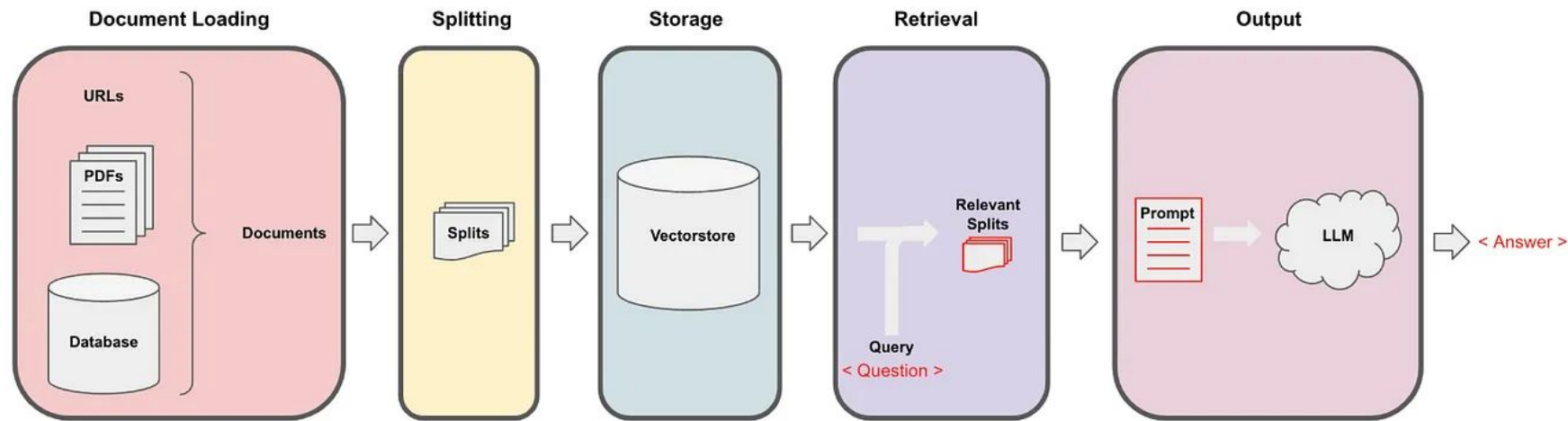


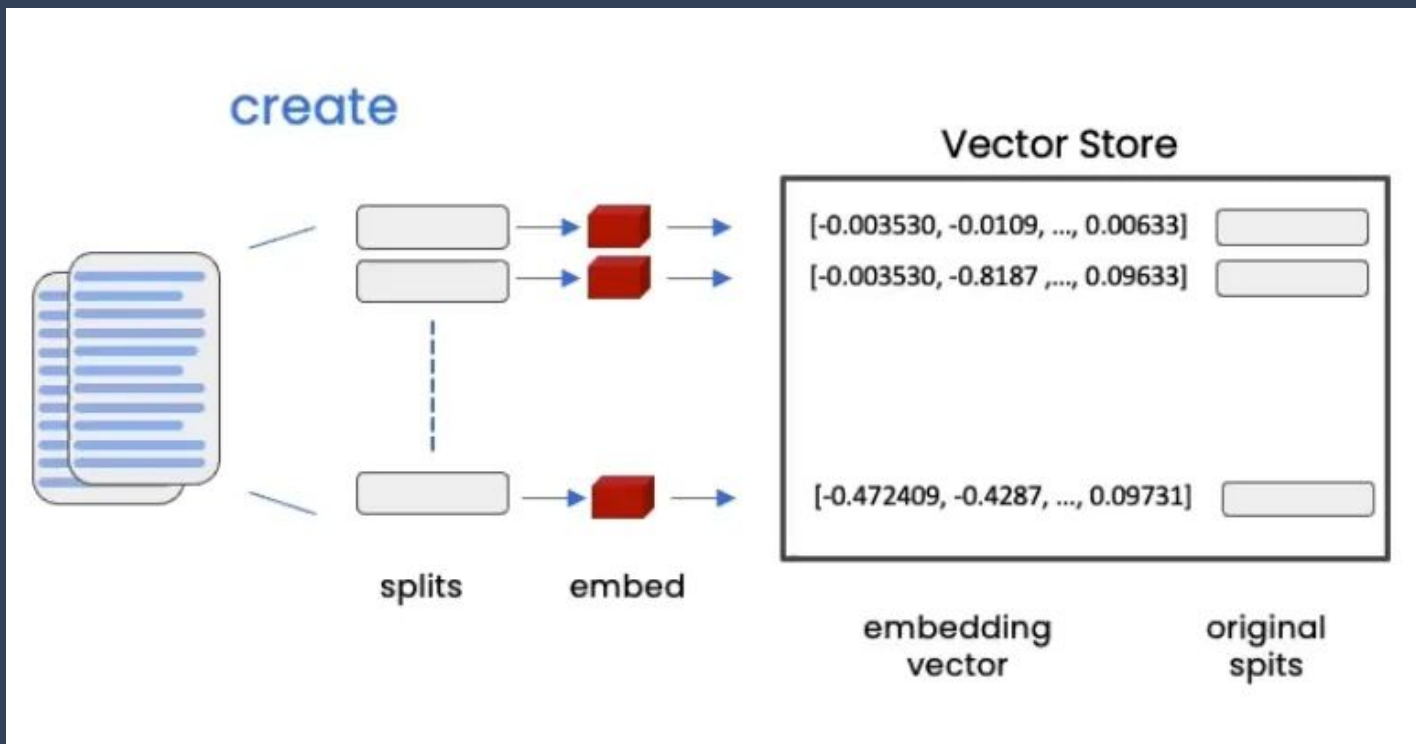
LectureBot Demo | Tools/APIs Used

- 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

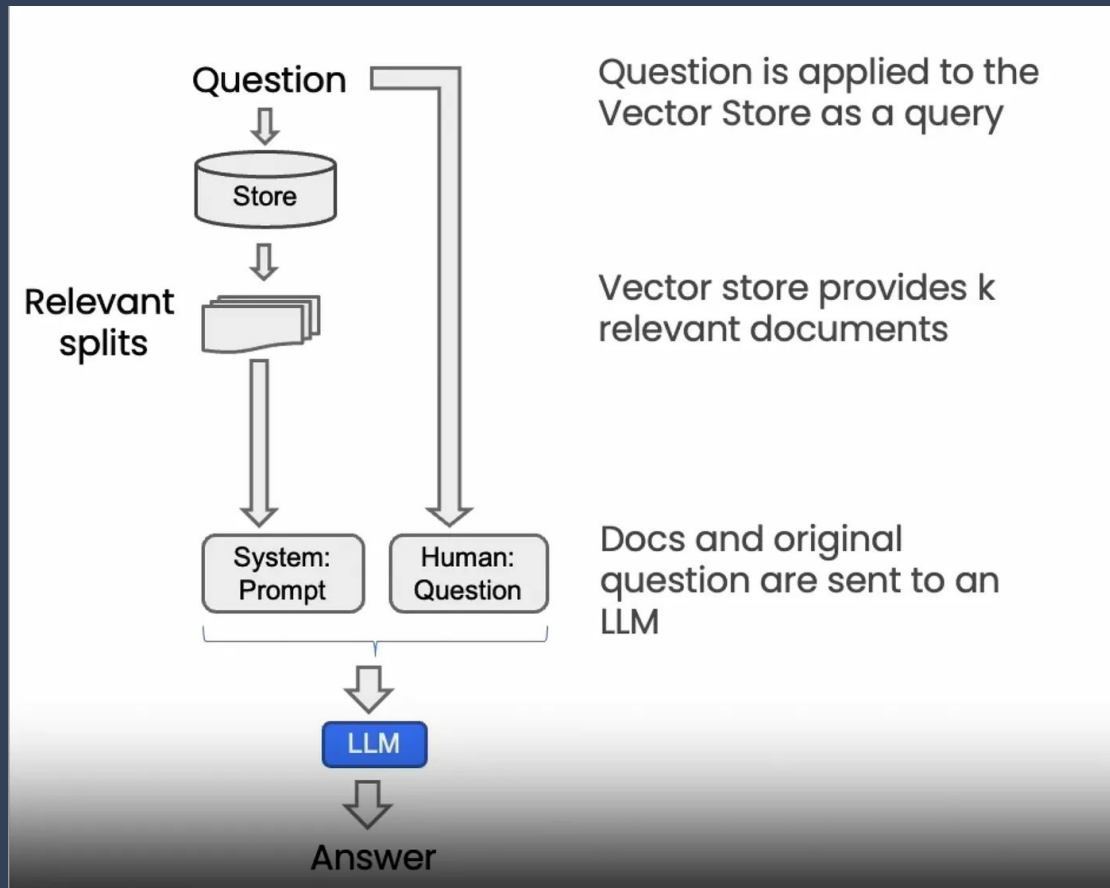
Demo LectureBot



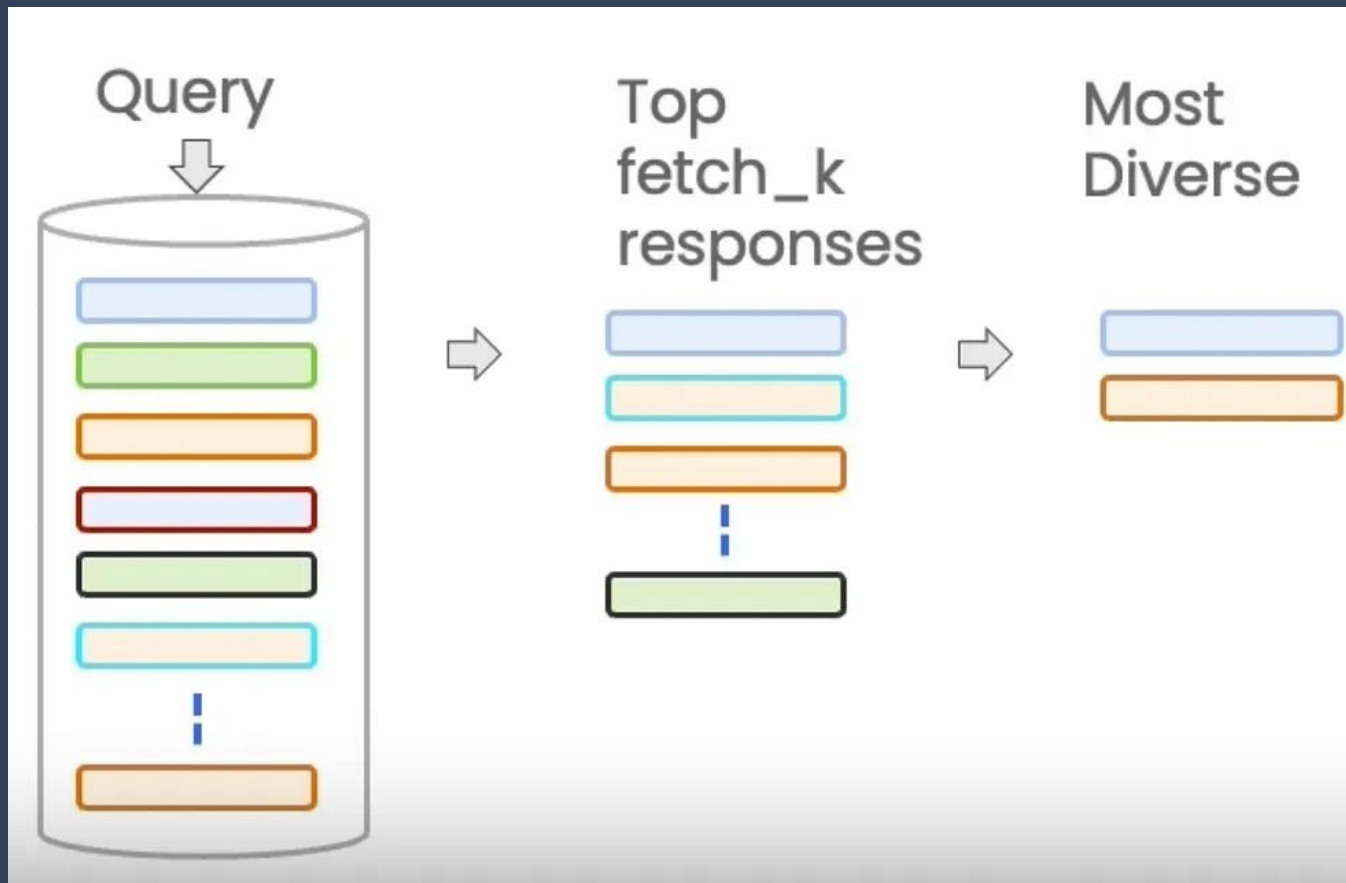




Document processing | Retrieval

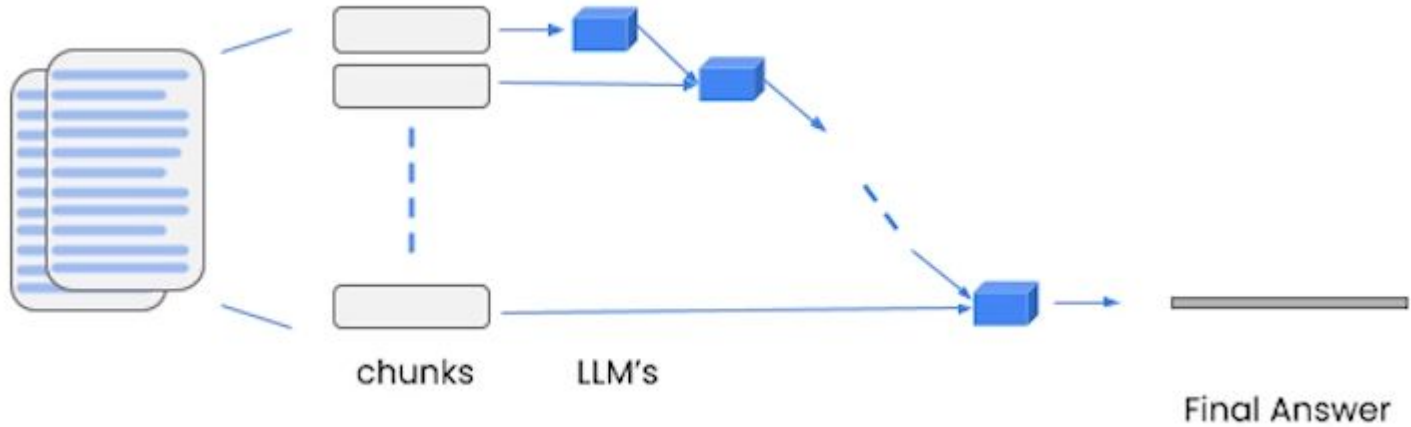


Retrieval Query| MMR Maximum Marginal Relevance



Prompting Context | Stuff Documents vs. Refine Docs

Refine



Is GenAI relevant in my job domain?

Software Engineers | Why should they upskill in Generative AI

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 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

Apply

Save

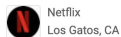
...

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.

Product Managers | Why should they upskill in Generative AI

Product Manager, Consumer Intelligence Algorithms



\$ 110K–190K a year  Full-time



AI Product Manager

Microsoft · Mountain View, CA

Base pay range

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

Senior Product Manager, Generative AI



Google
Portland, OR



1000+ Openings
for AI PM jobs

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

Product Manager - Generative AI



Meta
Menlo Park, CA

Qualifications

- Employer will accept a Bachelor's degree in Computer Science or related field, followed by 12 years of experience in job offered or in an IT Project management-related occupation. CSM, PMI-ACP, PRINCE2, SAFe, or IC Agile certification is required.
- Working with technical teams developing software solutions, including acting as a Scrum Master, Kanban, or Agile Coach.
- **Artificial intelligence, machine learning, database, and big data stacks.**
- Project Management Methodology.
- Experience in Product Management to help Product Managers from ideate to launch a new product

What You'll Bring

- Strong project management skills, including planning, scheduling, and budgeting.
- Knowledge of project management methodologies and tools
- Excellent communication and interpersonal skills to collaborate with technical teams and stakeholders.
- Ability to manage and motivate cross-functional teams to achieve project objectives.
- Strong analytical and problem-solving skills to identify and resolve project issues.
- **Technical expertise in AI and GenAI**
- Ability to prioritize and manage multiple projects simultaneously.
- Strong leadership skills to guide project teams towards successful project completion.

Preferred Qualifications

- Master's Degree AND 1+ year(s) experience in engineering, product/technical program management, data analysis, or product development or equivalent experience.
- 4+ years of experience managing cross-functional and/or cross-team projects.
- **1+ year of experience in AI/ML**
- 1+ year(s) experience reading and/or writing code (e.g., sample documentation, product demos).
- Has a basic understanding of the Hardware (Silicon+System) milestones, deliverables and interdependencies for establishing an accurate and effective business and technical requirements.
- Background in Electronic Design Automation (EDA) methodologies and Computer Aided Design (CAD)
- Proficiency in project management tools and methodologies like Jira, Azure DevOps, or equivalent platforms.

Technical Program Manager | Why should they upskill in Generative AI

Sr. Technical Program Manager - Money En...
Databricks · San Francisco, CA (On-site)

What We Look For

- 7+ years of technical program management experience.
- Bachelor's degree in a related field (EE, Computer Science, Computer Engineering, Software Engineering).
- **Technical knowledge and experience of Cloud infrastructure, Big Data and AI**
- Experience in program management, process definition and improvements and influencing adoption of defined processes across multiple teams or organizations.
- Ability to facilitate conversations to prioritize, manage tradeoffs, identify gaps and risks, drive accountability, and measure successes.
- Comfortable handling conflicts and escalations
- Experience in project/sprint planning, execution driving, risk management and effective communication to the business stakeholders
- Experience operating autonomously across multiple teams and organizations
- Familiar with agile methodology in software development and program management and collaboration tools such as Jira, spreadsheets, etc.
- Experience with at least one cloud provider: AWS, Azure, or GCP

Technical Program Manager, CX Applications
Coinbase · Seattle, WA

What We Look For In You (ie. Job Requirements)

- BA/BS degree in Information Management Systems or equivalent experience.
- 3+ years experience, preferably in business applications development management, product management for internal products.
- Experience in large tooling delivery.
- Knowledge and experience working with Interactive Voice Response (IVR), Content Management System (CMS), Knowledge Management System (KMS), Machine Learning (ML), chatbots, and internal home-grown tooling (Amazon Connect, Contentful, Qualtrics, Salesforce preferred).
- Strong verbal and written communication skills.
- Self-starter.
- Experience tackling complex, ambiguous technical challenges.
- Background working within an Agile environment and using Jira.
- Salesforce Administrator Certification (Advanced Admin, App Builder, Service Cloud Consultant preferred).

Nice To Haves

Technical Program Manager – LLM, Gen AI
Cognite · Austin, TX (Hybrid)

Who You Are

- **3+ years of AI/ML/LLM experience**
- 7+ years of working directly with engineering teams experience
- 5+ years of technical product or program management experience
- 3+ years of software development experience
- 5+ years of technical program management working directly with software engineering teams experience
- Experience managing programs across cross functional teams, building processes and coordinating release schedules
- 5+ years of project management disciplines including scope, schedule, budget, quality, along with risk and critical path management experience strongly preferred
- Experience managing projects across cross functional teams, building sustainable processes and coordinating release schedules strongly preferred
- Experience defining KPI's/SLA's used to drive multi-million dollar businesses

Technical Program Manager, ML / AI
Meta · Austin, TX

- Drive technical excellence within the team, coordinating and contributing to engineering deliverables including architectural diagrams, specifications, launch criteria, and test plans.
- Work cross-functionally to develop best practices and development processes in a quickly-changing and dynamic environment, drive impact through deployment of key initiatives and garner adoption of those processes.

Minimum Qualifications:

- B.S. in Computer Science or a related technical discipline, or equivalent experience.
- 7+ years of technical program management, software engineering, or systems engineering experience.
- **Experience building Machine Learning technologies, and shipping them into products.**
- Demonstrated experience leading execution across highly ambiguous, complex products and programs, with experience making technical and product tradeoffs balancing business needs and technical constraints. Knowledge of user needs, gathering requirements, and defining scope.

Engineering Managers (Tech) | Why should they upskill in Generative AI

- 8 years of experience with software development in one or more programming languages (e.g., Python, C, C++, Java, JavaScript).
- 3 years of experience in a technical leadership role; overseeing strategic projects, with 2 years of experience in a people management, supervision/team leadership role.
- **Experience in Generative AI (Large Language Models, Multi-Modal, Large Vision Models).**
- Experience with machine learning algorithms and tools (e.g., TensorFlow), artificial intelligence, deep learning, natural language processing or other ML discipline.

Preferred qualifications:

- Master's degree or PhD in Engineering, Computer Science, or a related technical field.
- 3 years of experience working in a complex, matrixed organization.

Engineering Manager, Lens Studio

Snap Inc. · Los Angeles, CA

Apply

Save

...

EXPERIENCE

- Build and grow a team of exceptional software engineers and technical leaders
- Create growth opportunities, give regular feedback, promote talent, manage performance

Knowledge, Skills & Abilities:

- Knowledge of game engine design patterns
- Strong computer science fundamentals
- Expertise in modern C++
- Solid understanding of generative ML workflows, like ComfyUI, and other ML technologies which are democratizing complex creation processes in this and similar fields
- Strong product sense
- Ability to collaborate with internal and external stakeholders at all levels of a company
- Excellent written and verbal communication skills
- Ability to influence and convey messages to a wide range of stakeholders

Engineering Manager - Growth Messaging

Netflix · United States (Remote)

Apply

Save

...

Nice To Have Skills

- Experience working on Messaging.
- Experience working with ML Researchers, Algorithm Engineers or Data Scientists.
- **Knowledge and experience working with machine learning systems.**

Netflix has a unique culture that values employee freedom and responsibility. We seek to grow inclusive and diverse teams that will enhance our perspectives, skill sets, and behaviors. We highly encourage you to apply if your background will complement us, even if your experience doesn't precisely match the job description. Your skills and passion will stand out—and set you apart—especially if your career has taken some

Senior Manager, AI Engineering - Marketplace Monetization AI

LinkedIn · Sunnyvale, CA (Hybrid)

Apply

Save

...

BASIC QUALIFICATIONS:

- 7+ years of relevant professional experience
- 3+ years of management experience
- BA/BS in Computer Science or other technical discipline, or related practical technical experience
- Hands on experience in data modeling and machine learning engineering

PREFERRED QUALIFICATIONS:

- 10+ years of relevant professional work experience
- 5+ years of experience leading engineering teams.
- At least one year of experience managing other managers and technical leads.
- Domain experience in Ads AI or Marketplace AI
- MS or PhD in Computer Science, Machine Learning, Statistics or related fields

SUGGESTED SKILLS:

- **Machine Learning & AI**
- Engineering Leadership

Basic Qualifications

- B.S. in Computer Science, Electrical, Computer Engineering, Data Science, or equivalent.
- 5+ years experience as a software engineer, and programming skills in Java, Go, or Python
- 5+ years full-time engineering management work experience.
- Experience leading both engineers and applied/data scientists.
- Ability to problems solve and make complex decisions with incomplete information in highly ambiguous situations and environments.
- Great interpersonal skills, deep technical ability, and a track record of successful execution in cloud security engineering or product development at cloud-scale.

PREFERRED QUALIFICATIONS

- Experience building privacy preserving technologies.
- Prior experience in leading teams over multiple locations, and working across multiple in-house engineering organizations
- Experience in algorithm development and prototyping.
- Experience with productionizing algorithms for real-time systems.
- **Good understanding of LLM fine-tuning, RAG, and guardrails**

Engineering Manager, Payment Intelligence...

Stripe · United States (Remote)

Apply

Save

...

Who you are

We're looking for someone who meets the minimum requirements to be considered for the role. If you meet these requirements, you are encouraged to apply. The preferred qualifications are a bonus, not a requirement.

Minimum Requirements

- 3+ years of direct engineering management experience
- 1+ year of experience working within a team responsible for developing, managing, and optimizing ML models or ML infrastructure

Preferred Qualifications

- Proven track record of building and deploying machine learning models or systems that have effectively solved critical business problems
- Experience managing teams that leverage real-time, distributed data processing
- Experience managing teams that leverage batch processing pipelines
- Experience building sustainable operations for managing many ML models, including CI/CD, auto-training, auto-deployment, and continuous model refreshes
- Experience managing teams that owned many diverse ML models
- Experience in adversarial domains like Fraud, Trust, or Safety
- Past experience operating under team goal-setting frameworks such as OKRs

Applied Gen AI | Curriculum Outline

Python Crash Course

Python fundamentals
Python Libraries for Machine Learning

Getting Started Generative AI

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

Building Gen AI Applications

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

Domain	Product Managers	Backend, Frontend, Fullstack, Test Engineers	Default	Engineering Manager	Technical Program Managers
--------	------------------	--	---------	---------------------	----------------------------

W11-12	Product Management(Tech) Specialization AI Product Management Gen AI Product Strategies, Roadmap & Execution	Software Engineering Specialization AI Engineering Tech Stack Full Stack AI Engineering	Capstone Project Industry Project BYOP	Engineering Manager Specialization Strategic Integration of LLMs in Engineering Projects System Design Innovations with LLMs	Technical Program Manager Specialization Mastering Generative AI Project Management Strategic AI System Design for Program Managers
--------	---	--	--	---	--

W13-14	Capstone Project AI Product Strategies, Roadmaps & Execution BYOP	Capstone Project Full Stack LLM Application Development BYOP		Capstone Project Gen AI Innovation & Integration for Engineering Leaders BYOP	Capstone Project Generative AI Program Management BYOP
--------	---	--	--	---	--

Q&A

