ACCUKNOX AI/ML ASSIGMENT

ACCUKNOX

Pratik Kishor Nayak kishorpratik023@gmail.com *Resume Link*

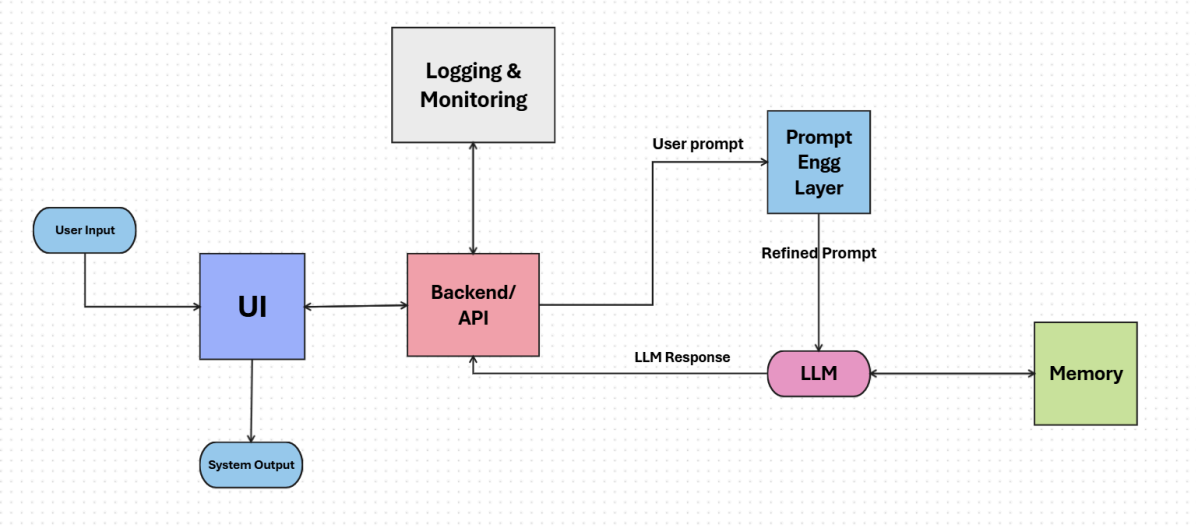
Problem statement 2

2.1 Self-Assessment of Skills (LLM, Deep Learning, AI, ML)

LLM	**A** Why ? I understand using large language models in chatbots and retrieval systems. I can integrate LLMs via APIs, design prompts, and build small apps around them.
Deep Learning	**C** Why ? I understand deep learning theory, like neural networks and backpropagation, but lack hands-on model-building experience and need guidance.
Artificial Intelligence	**B** Why ? I understand fundamental AI concepts, like problem formulation, search techniques, and machine learning for decision-making, but I need guidance with larger systems.
Machine Learning	**A** Why ? I rate myself **A** in Machine Learning as I can preprocess data, train models, evaluate performance, debug and improve models independently.

2.2 The key architectural components to create a chatbot based on LLM :

An LLM chatbot is not just a model but a system that combines UI, backend logic, prompts, memory, knowledge, and safety controls to produce helpful responses.



1. **User Interface (UI)**

- This is the **front-end** where the user inputs queries and receives replies.
- **Purpose:** To collect user input and show the chatbot's response.

2. **Backend / API Layer**

- This is the middle layer that connects everything.
- What it does:
  - Receives user messages from the UI
  - Sends requests to the LLM
  - Returns the final response to the UI
- **Purpose:** Acts as the "brain connector" between the user and the AI model.

3. **Prompt Builder (Prompt Engineering Layer)**

- Before sending a message to the LLM, the system formats the input properly.
- It may include:
  - System instructions such as “You are a helpful assistant”.
  - Chat history
  - User’s latest message
- **Purpose:** To guide the LLM to give better and more relevant answers.

4. **Large Language Model (LLM)**

- This is the core AI component in LLM based chatbot.
- What it does:
  - Understands user input
  - Generates natural language responses
- **Purpose:** To produce intelligent, human-like replies.

5. **Context & Memory Management**

- Chatbots need to remember previous messages to hold a conversation.
- This can be:
  - Short-term memory (last few messages)
  - Long-term memory (stored conversations)
- **Purpose:** To make conversations feel continuous and natural.

6. **Safety & Control Layer**

- This layer checks and validates user input and model output. It helps prevent from harmful content, policy violations and incorrect or unsafe answers.
- **Purpose:** To keep the chatbot safe and reliable.

7. **Logging & Monitoring**

- The system tracks User interactions , Errors and Response quality.
- **Purpose:** To improve the chatbot over time and fix issues.

2.3. Vector Database

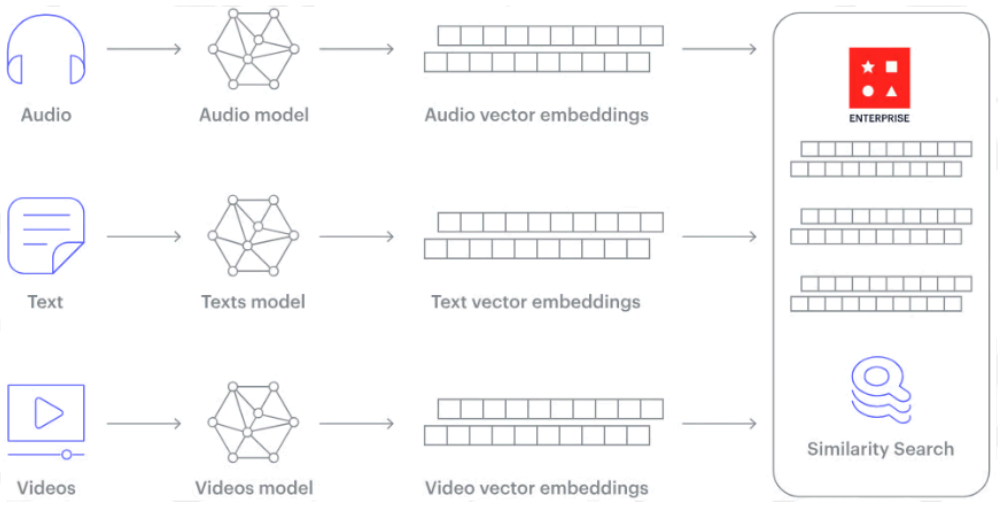
A vector database stores information as multi-dimensional vectors that capture specific characteristics.

The dimensions can vary from a few to thousands, depending on data complexity. Various types of data, such as text and images, are converted into vectors using techniques like machine learning or feature extraction.

The main benefit of a vector database is its ability to retrieve data based on vector similarity, allowing for searches based on semantic relevance rather than exact matches.

For example, it can be used to:

- Search songs by melody and rhythm.
- Find articles with specific themes.
- Identify gadgets similar in characteristics and reviews.



Problem Statement :

How Might We create an automated product matching system for multi-vendor marketplaces such as Amazon or eBay? The goal is to consolidate duplicate listings into a single product entity based on semantic and visual similarities, rather than relying solely on exact metadata.

Vector Database Choice:

Pinecone

Why ?

- **Lower Latency for real time De-duplication:**
  - Pinecone employs Approximate Nearest Neighbor (ANN) search.
  - **Efficiency:** This capability enables it to sift through millions of existing products and identify a 99% match in just milliseconds.
- **Hybrid Search :**
  - At times, the model number "WH-1000XM5" serves as the identifier that indicates a duplicate. It's important to note that pure vector search may occasionally overlook specific alphanumeric strings.
- **Metadata Filtering :**
  - Pinecone can limit its search to the "Sony" or "Electronics" segment, enhancing accuracy and preventing your Sony headphones from being mistakenly matched with a Bose headset just because they look similar.