Project Title: AI-Assisted Learning for NVIDIA SDKs and Toolkits

Objective:

Develop a sophisticated LLM model utilising deep learning techniques to provide a seamless and intelligent learning experience to users for NVIDIA SDKs and Toolkits.

Key Features:

Natural Language Processing (NLP):

- Implement advanced open source **Llama-2-7b-chat** model to comprehend and generate human-like responses.
- Fine-tune models on NVIDIA-specific datasets for improved contextual understanding using **Hugging Face transformers** and **LangChain** framework with **FAISS library** vector store .

Context Management:

- Develop a context-aware system to maintain continuity in conversations.
- Implement memory networks or attention mechanisms to understand and remember user inputs over time.

Technology Stack:

Google Colab:

- **Description:** A cloud-based platform that provides free access to GPUs and TPUs for running Python code. Ideal for data analysis, machine learning, and collaborative projects.
- **Purpose:** Running the project in a cloud environment with GPU support for efficient model training and experimentation.

Python:

- **Description:** A high-level programming language commonly used for machine learning and data science tasks.
- **Purpose:** Writing and executing the project code, including model training, data preprocessing, and integration with various libraries.

Hugging Face Transformers:

- **Description:** An open-source library that provides pre-trained models and tools for working with transformer-based models, including Llama-2-7b-chat.
- **Purpose:** Utilizing pre-trained models, fine-tuning if necessary, and accessing utilities for working with transformer architectures.

LangChain Framework:

- **Description:** A framework for natural language processing (NLP) tasks, focusing on chat applications and conversational AI.
- **Purpose:** Integrating the Llama-2-7b-chat model into the LangChain framework to leverage its specific functionalities and features.

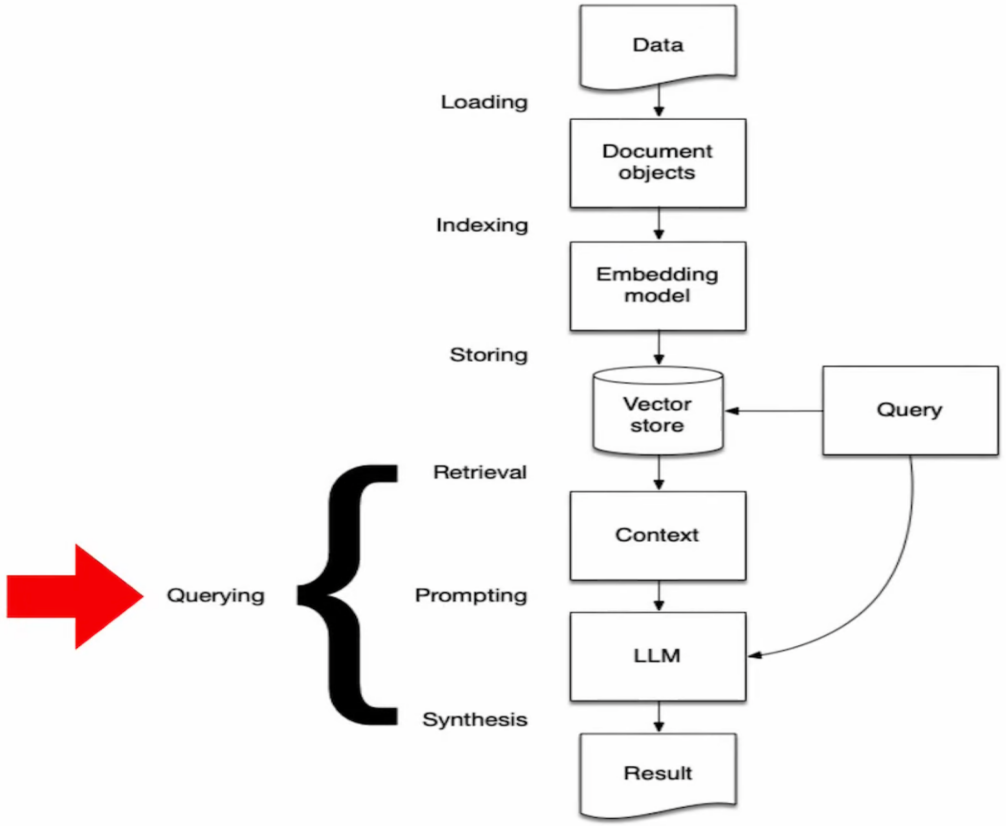
FAISS Library:

- **Description:** A library for efficient similarity search and clustering of dense vectors. It is often used for fast and scalable similarity search in large datasets.
- **Purpose:** Integrating FAISS for efficient similarity search operations, potentially in the context of the chat model or related tasks.

Dependencies and Libraries:

- **Description:** Additional Python libraries and dependencies required for tasks such as data manipulation, visualization, and other functionalities specific to the project.
- **Purpose:** Supporting various aspects of the project, such as data preprocessing, visualization, and other necessary tasks.

Development Phases in diagram :



Conclusion:

The model aims to offer practical guidance and address specific user concerns on NVIDIA SDKs and toolkits. The project prioritizes user experience, security, and scalability, ensuring a robust and intelligent platform for seamless communication.