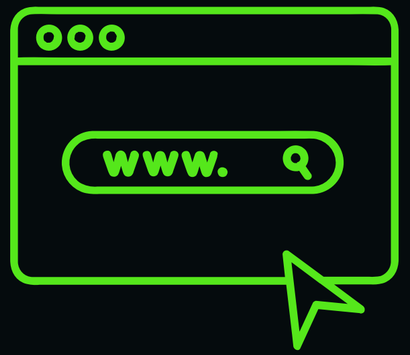
Tri-Virtual Assistant Chatbot

Using NVIDIA Workbench AI and NVIDIA NIM

Mentor :
Prof. Manjul Krishna Gupta

TEAM MEMBERS :
1) Renu Bojja:1RVU22CSE129
2) Karmishtha P:1RVU22CSE077

Project Overview



OBJECTIVE

INFORM
TROUBLESHOOT
ANSWER
ANY QUERIES ON NVIDIA WORKBENCH SOFTWARE PRODUCTS

COMPONENTS

- CUSTOMIZABLE GRADIO APPLICATION - ADDS WEBPAGES,IMAGES ,VIDEO AND PDFS
- USES LANCE DB VECTORSTORE
- USING NVIDIA LAUNCH PAD PERMISSION ,IMPLEMENTED SELF-HOSTED ENDPOINTS USING NVIDIA NIM

Model

MISTRAL-7B-INSTRUCT-V0.2 (NVIDIA NIM)

This LLM follows instructions, completes requests, and generates creative text.

ROUTER MODEL (RAG PIPELINE)

Directs user queries to the appropriate pipeline (RAG or Websearch) based on the topic.

RETREIVAL MODEL (RAG PIPELINE)

Assesses the relevance of retrieved documents from the knowledge base, filtering out non-relevant results to provide higher-quality context.

GENERATOR MODEL (RAG PIPELINE)

Generates responses based on the retrieved documents and context, answering user queries directly.

HALLUCINATOR MODEL (RAG PIPELINE)

Evaluates the response generated by the assistant to ensure it remains faithful to the information in the retrieved documents, reducing the chances of AI-generated "hallucinations."

ANSWER MODEL (RAG PIPELINE)

Grades the generated response to confirm it accurately answers the user's original query.

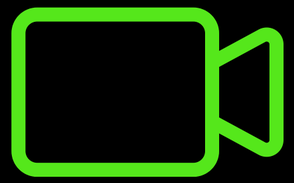
Datasets

NVIDIA AI WORKBENCH KNOWLEDGE BASE

- Purpose: Core knowledge source for NVIDIA-specific queries.
- Storage: Ingested into LanceDB vector database for efficient retrieval.

MULTIMODAL EMBEDDINGS DATASET

- Vector embeddings for text, images, audio, and video.
- Purpose: Enables semantic retrieval, matching user queries with similar content.
- Storage: Stored in LanceDB with collections for different modalities.



OCR AND AUDIO TRANSCRIPTION DATA

- Content: Text extracted from images (OCR) and transcriptions of audio/video files.
- Purpose: Supports multimodal retrieval, adding text data from visual and audio files

Tools Used

CORE FRAMEWORKS AND LIBRARIES

Python 3
Gradio

DATABASE AND STORAGE

LanceDB
FFMEG

MACHINE LEARNING AND NLP LIBRARIES

LangChain
LLamaIndex And Mistral
LangGraph

EMBEDDINGS AND MODEL INFERENCE

NVIDIA Embeddings
NVIDIA NIM
NVIDIA Workbench AI
NVIDIA Launchpad

MULTIMEDIA AND DOCUMENT PROCESSING

Pytube
MoviePy

DEVELOPMENT AND MONITORING

Docker
Tesseract OCR
Speech Recognition

Goals

Architecture Type: Client-Server Model with RAG Pipeline

Frontend (Client)

- Tool: Gradio
- Function: Provides user interface for Control-Panel and Public-Chat applications.
- Components:
  - Chat window for user queries.
  - Knowledge base management and settings panel.

Data Storage Layer

- Database: LanceDB
- Function: Stores embeddings of documents, images, videos, and other content for retrieval.
- Collections: Organized by content type (e.g., web pages, PDFs, images).

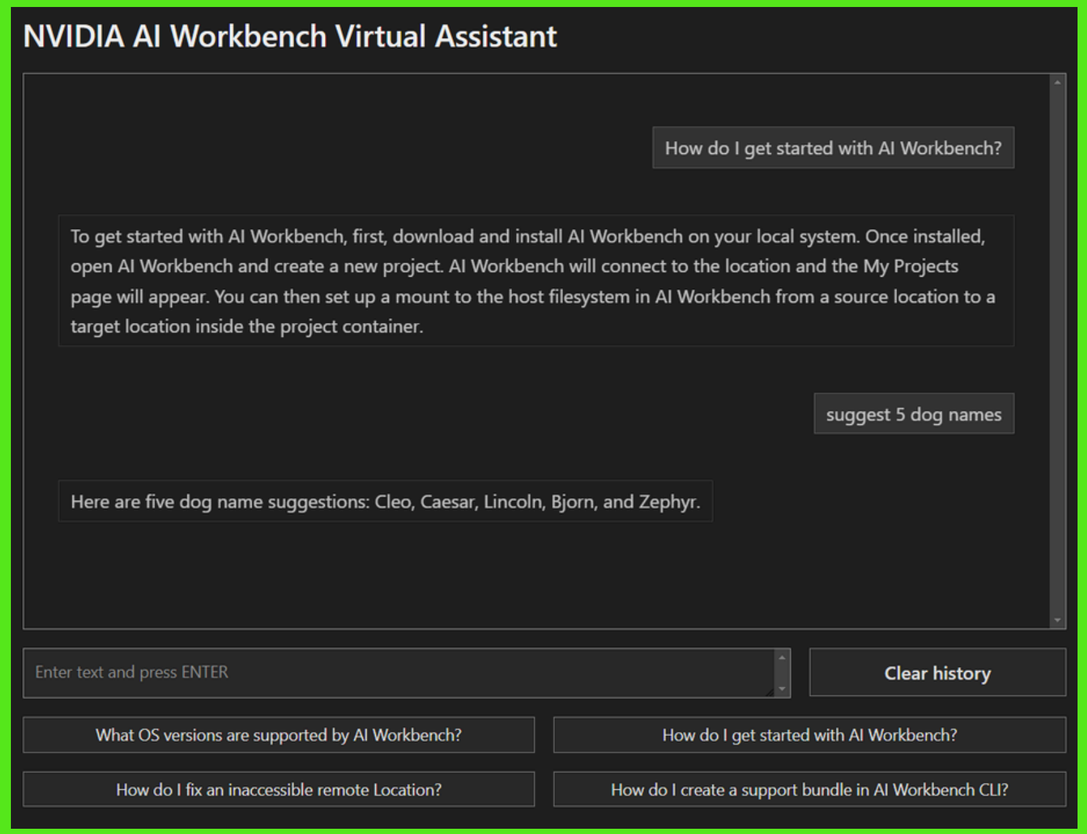
RAG Pipeline (Retrieval-Augmented Generation)

- Tool: LangGraph
- Function: Directs user queries through the following workflow:

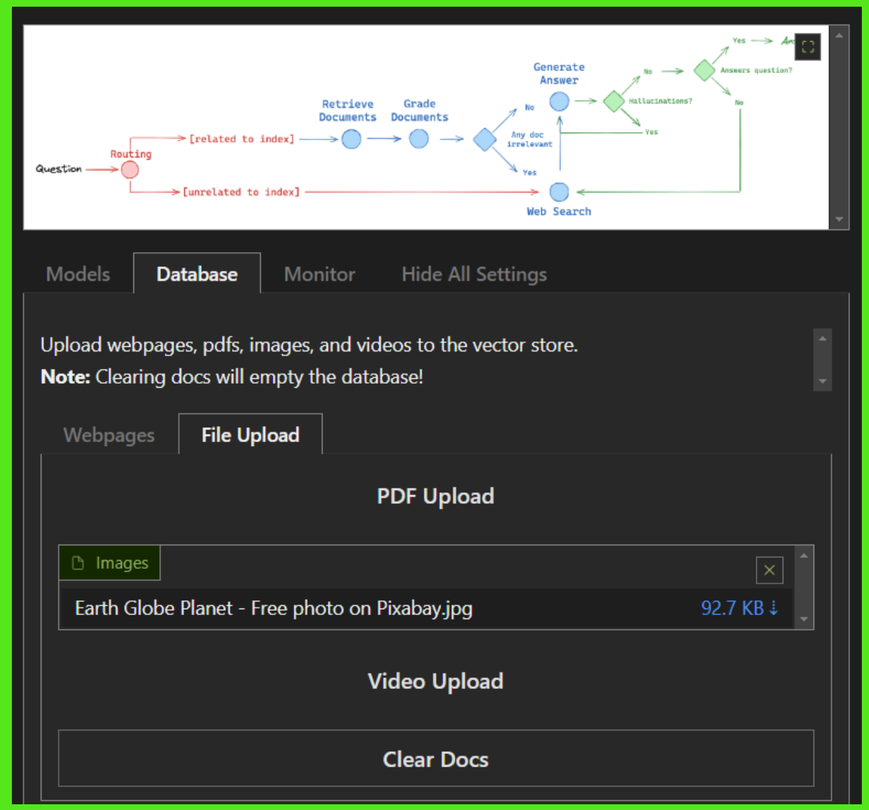
Inference & Embeddings

- Models: NVIDIA API Catalog for cloud inference or NIM for self-hosted inference.
- Purpose: Generate embeddings and process language tasks.
- Compatibility: GPU-optimized for efficient local processing.

UI/UX Overview

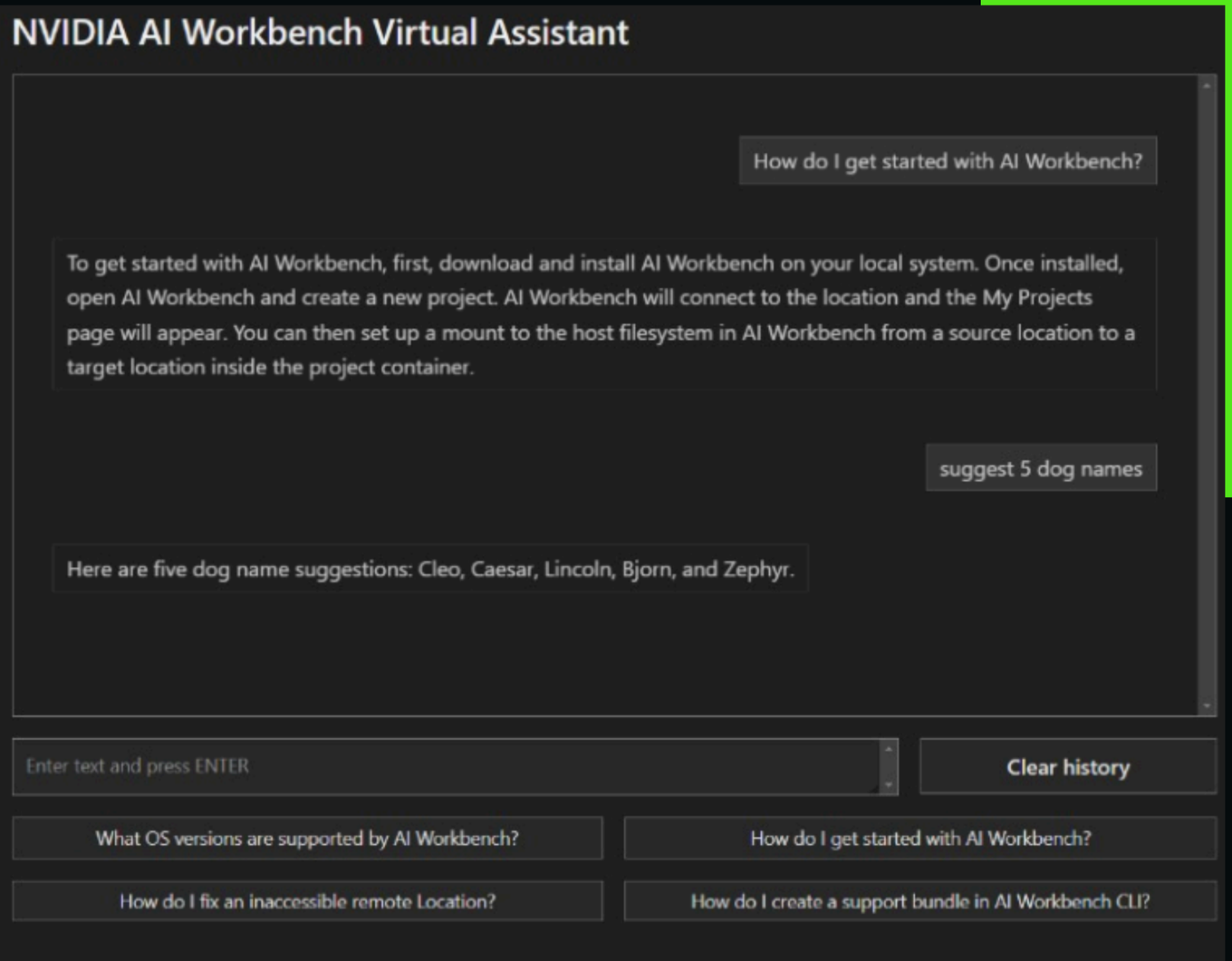


- Framework: Gradio
- Design Goals:
- Intuitive, clean layout for easy interaction.
- Minimalist style with collapsible menus for a focused user experience.



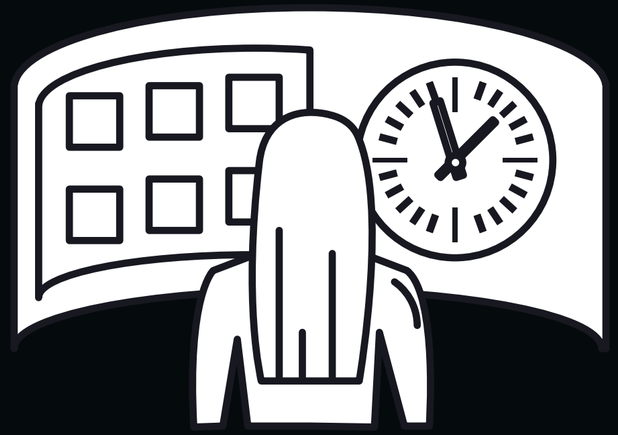
- Chat Window: Main area where users input queries and view responses.
- Knowledge Base Management: Allows users to upload and manage documents (Control-Panel app only).
- Settings Panel: Collapsible side panel with tabs for model and database settings, configuration, and monitoring.

Outputs



- Objective: Developed a multimodal virtual assistant for NVIDIA AI Workbench using RAG pipeline and web search integration.
- System Functionality:
- Ingests and processes various document types (web pages, PDFs, images, videos) into LanceDB.
- Uses multiple LLMs for query routing, answer generation, and quality assessment.
- Testing: Deployed and tested in NVIDIA AI Workbench environment, integrated with Docker for optimized performance.
- Outcome: Prototype successfully aids users in troubleshooting and navigating the AI Workbench.

Future Work



CLOUD ENDPOINTS:

- INTEGRATE CLOUD AND SELF-HOSTED INFERENCE.

MULTIMEDIA:

- IMPROVE AUDIO, VIDEO, AND OCR PROCESSING.

DYNAMIC UPDATES:

- ENABLE AUTO-UPDATES AND ADAPT TO NEW CONTENT.

THIRD-PARTY APIS:

- EXPAND INTEGRATIONS FOR BROADER KNOWLEDGE.

ADVANCED QUERIES:

- SUPPORT MULTI-TURN CONVERSATIONS AND FEEDBACK.

SCALABILITY:

- OPTIMIZE PERFORMANCE UNDER LOAD.

PERSONALIZATION:

- ADD USER PROFILES AND CUSTOMIZATION.