Sambhav Khanna

sambhavKhanna | sambhavKhanna | sambhav-khanna | s4khanna@uwaterloo.ca | 519-404-4057

Education

University of Waterloo Sep 2022 – Apr 2027
Bachelor of Computer Science, AI Specialization (Cumulative Average 89.45%) *Waterloo, Canada*

- Relevant Coursework: Data Structures and Algorithms, Operating System, Numerical Computation, Cryptography, Object-Oriented Programming, Sequential Programs, Computer Architecture, Statistics

Experience

Software Engineer Jun 2024 – Sep 2024
Tallgeese AI | Flask, ChromaDB, Langchain, React.js, Tailwind *San Jose, United States*

- Developing RAG agent and full stack application to help universities, assess and teach their students effectively.

Software Engineer Apr 2024 – Jun 2024
Qatar Airways | Milvus, LlamaIndex, OpenAI API, Flask, React.js *Doha, Qatar*

- Developed a RAG system for storing **100+** documents, each with **200+** pages containing airplane technical data.
- Indexed the embeddings and stored them in a cloud vector databse, to decrease the time to first token by **80%**.
- Implemented data cleaning pipeline, to clean tabular data, thereby increasing the accuracy of responses by **100%**.
- Managed an intern and assited them in prompt engineering, RAG API development and frontend development.

Full Stack Engineer Sep 2023 – Dec 2023
Toronto Transit Commission | Python, TypeScript, Electron.js, Docker *Toronto, Canada*

- Developed an encrypted password management application to store and manage **1000+** logins for **40+** databases.
- Reduced the application render time by **80%**, by migrating the codebase from Python **Tkinter** to **React.js**.
- Containerized the application, and used caching to reduce the build time from **5 minutes to 5 seconds**.
- Implemented SQL triggers and transactions to reduce data processing time, by decreasing unnecessary query runs.

Projects

llama3 RAG (LLM) | *Python, Langchain, Ollama*

- Developed a **RAG** pipeline for Meta's llama3:8b model using Langchain and Ollama to pull the model locally.
- Used **ChromaDB** vector database to store the docs embeddings and Ollama embeddings to embed the docs.
- Tested the RAG model with PDF files containing tables and achieved successful response on **80%** of the prompts, asking about data entries of specific cells.

ChatrBox | *MongoDB, React, Express, Node JS, socket.io, nginx, Docker, AWS*

- Developed an end-to-end chat app using the **MERN** stack (**MongoDB** as database and **React** front-end).
- Implemented a real-time chat server using the **websocket** implementation provided by **socket.io** API.
- Implemented custom hooks and custom contexts using **React Context API** to avoid prop-drilling.
- Developed a client-side routing system using **React-Router-DOM** for smooth transition.
- Containerized the application using **Docker** and deployed the image on an **AWS EC2** instance.

Multi Layered Perceptron (MLP) | *Python*

- Developed a MLP from scratch without using PyTorch. Defined class to keep track of the gradient when arithmetic operations are applied.
- Created a functions to perform forward pass and back propagation through the neural network.
- Created training and testing datasets and trained the model in multiple batches, performing gradient descent on the loss function at each iteration.

Technical Skills

Languages: Python, TypeScript, JavaScript, C/C++, HTML/CSS
Frameworks: PyTorch, Numpy, Pandas, Langchain, LlamaIndex, Express.js, Next.js, Node.js, React.js, Flask, Django
Developer Tools: Ollama, MongoDB, PostgreSQL, MySQL, AWS, GCP, Docker, Git, Linux, Azure