Shiva Ram Godar

Gongabu, 44600, Kathmandu, Nepal
shivagodar44@gmail.com linkedin.com/in/kamalgodar
github.com/kamalgodar (+977) 9841959244

Education

Pulchowk Campus, Institute of Engineering, Tribhuvan University (TU) Nov 2017 - Apr 2022
Bachelor of Electronics and Communication Engineering Lalitpur, Nepal
Major Courses: *Object Oriented Programming, Artificial Intelligence, Data Mining, Big Data, Engineering Mathematics, Computer Networks, Computer Graphics, Embedded System, Microprocessor, Computer Architecture*

Work Experience

Fusemachines Jun 2022 - Present
Machine Learning Engineer Kathmandu, Nepal

- Squadery: worked on Document AI developing a Talent hiring platform, where I performed the tasks of parsing and segmentating the content of resume, developed a Named Recognition Model (NER) to extract the essential information from resume and finally score the candidate's resume based on the job description
- ResumeWriter: secured a second position in AI Hackathon where we leveraged ChatGPT to build a platform that assists people on writing the Resume providing the necessary suggestions and recommendations based on Job Description, ATS score was seen to have been reached to 85%
- Chatbots: developed various Retrieval Augmented Generation (RAG) based chatbots as a Proof of Concept (PoC) on various domains including food, finance and airlines using various open source and closed source Generative AI models
- AI Studio (AnswerGen): Developed one of the engines called AnswerGen using the open source LLM (Mistral-Instruct) and closed source LLM (ChatGPT) where the system is able to answer the queries related to uploaded document

Project Experience

AI Studio (AnswerGen) March 2024 - Jul 2022

- Built one of the engines in AI Studio that can answer the questions based on the files being uploaded to the project; the system supported unstructured documents with pdf, docx, txt extensions
- The uploaded documents were stored in the vector database, ChromaDB, which facilitated context retrieval during inference based on user queries. This retrieved context was then utilized by the open-source model, Mistral-Instruct, to generate appropriate responses to the queries.
- Deployed the Mistral-7B-Instruct using NVIDIA Triton Inference Server and vLLM

Tools: *Langchain, FastAPI, ChromaDB, Mistral, RabbitMQ, MongoDB*

Chatbot May 2023 - Feb 2024

- Built a customized RAG system leveraging Large Language Model (LLM) that automates the drive through service helping customers to make the order of the food items based off the restaurant's food menu
  - Used Pinecone as a vector database to store all the details of the food items of the restaurant and made the system to note down the order details (change, add, cancel) based on the conversation with the customer
- Built a generative AI chatbot for the debt settlement company that answers the queries based off the knowledge base and also gather users information to test the qualification of the candidate for the debt settlement program
  - Used Milvus to store the knowledge base of the company that included the information on various debt programs with the eligibility criteria for each
  - Experimented with open source models Falcon 40B and Llama2 and used the Mistral-Instruct model to respond to the users queries
  - Fine tuned Mistral-Instruct-7B on the QA and conversational datasets using PEFT (LoRA) for better conversation flow to make it able to ask the required questions to know the eligibility of the candidate
- Built a Chatbot to answer the questions based on the structured and unstructured document types that contained tables and text
  - Leveraged the capability of the agents to answer the questions based on the data stored in the relational database (SQLite)
  - Used FAISS as a vector database to store the textual and tabular content of the unstructured file based off which the right context is retrieved for the response generation by LLM

Tools: *Langchain, Pinecone, Milvus, FastAPI, chatGPT, Mistral, Falcon, Llama, EC2*

Squadery Sep 2022 - May 2023

- A hiring platform that helps recruiter to manage the hiring process efficiently with the use of AI powered functionalities filter out the applicants applying for the job
- Worked on the resume parsing and segmentation of the resume information where various text parsing tools were experimented like tika parser, pdfminer, pdftotext and pymupdf was chosen based on its relative performance and the segmentation of the resume was done based on the identification of heading from the resume
- Built and fine tuned a Roberta model using spacy3 to perform Named Entity Recognition on text data with an accuracy of 70% for extracting the necessary information from resume that helps the recruiter to know more about the candidacy
- Built an NER model for extracting the necessary information from resume
- Built an scoring system that scores and evaluates the resume based on the Job Description

Tools: *PyTorch, AWS Lambda, PyMuPDF, Spacy, MongoDB*

Skills

Programming Languages Python, C/C++

Skills	Machine Learning (ML), Deep Learning (DL), Natural Language Processing (NLP), Recommendation Systems
Tools	Pandas, NumPy, PyTorch, TensorFlow, SpaCy, NLTK, Langchain, RASA, OpenCV, Beautiful Soup, Scrapy Vector Databases (Pinecone, Milvus, ChromaDB)
Others	FastAPI, Docker, Kubernetes, MongoDB, PostgreSQL, Amazon EC2, Amazon SQS, AWS Lambda, GIT, LaTeX

Personal Project

Optimized Itinerary Recommender using AI Mar 2021 - Apr 2022

- developed a web-based application that recommends destinations and provides optimized itinerary connecting the best destinations based on user's travel preferences and overall statistics of the destinations, ensuring a tailored travel experience
- implemented advanced filtering techniques utilizing content-based, collaborative and hybrid filtering methods, to recommend top destinations with detailed information of each, aligning with user interests
- applied the Ant-colony optimization (ACO) algorithm as an approach to solve the Travelling Salesman Problem (TSP) which helped to find the shortest possible route connecting all the recommended and preferred destinations, optimizing travel efficiency of an user

Tools: *Pandas, matplotlib, scikit-learn, ReactJS, Django, PostgreSQL*

Margadarshan: A Smart Traffic Management System Jun 2020 - Jan 2021

- developed a web-based application to manage traffic by alerting users about different unforeseen conditions such as heavy traffic, constructions, accidents and diverging them to less crowded routes to optimize their travel time
- implemented a vehicle counting system using Mixture of Gaussian (MOG) background subtraction, object detection by contour, and a object counting algorithm to assess traffic conditions in major city areas through video analysis
- integrated a feature allowing users to post real-time road conditions (construction, accidents, traffic jams, etc.) with a credit point system to verify the authenticity of posts, aiding other commuters

Tools: *Django, OpenCV, PostgreSQL*

Others

Tic-Tac-Toe	C++
Four Seasons	A Graphics Project to design and model four seasons using Blender

Trainings and Certificates

Generative AI with Large Language Models Coursera
AI Fellowship, 2022 Fusemachines
Introduction to Data Science in Python Coursera
SQL for Data Science Coursera
Applied Machine Learning in Python Coursera
Neural Network and Deep Learning Coursera
RASA Certificate RASA
HTML, CSS and JS for Web Developers Coursera

Extra Activities

Award and Recognition
AI Hackathon, Fusemachines, 2023 (Second) Certificate
Golden Jubliee Scholarship, 2018

Hobbies
Football, Cricket, Movies, Hiking