Suryansh Gupta

AI/ML Engineer

✉ suryanshgupta9933@gmail.com ☎ +918700644659 Delhi, India

https://linkedin.com/in/suryanshgupta9933 https://github.com/suryanshgupta9933

PROFESSIONAL EXPERIENCE

AI/ML Intern 06/2024 – present
Inspire AI Delhi, India

- Developed pipeline for **personalized AI news feed** catering domain specific news according to user preference from multiple sources and condensing using **Retrieval Augmented Generation (RAG)** utilizing **LlamaIndex** and **Langchain.**
- Designed and implemented **user onboarding chatbot** for generating hyper focused user profile for better **recommendation modelling.**
- **Fine-tuning Large Language Models (LLM)** on Linkedin post data for creating an AI ghostwriter, reducing model hallucination and AI content generation detection upto 80%.

AI/ML Intern 01/2024 – 04/2024
Acencore Delhi, India

- Developed an **AI powered resume builder and review system** for tailoring resumes to job-specific requirements utilizing **Prompt engineering for LLMs.**
- Implemented predictive skill forecasting and skill gap analysis using **clustering algorithm** and **cosine similarity** helping users to align their skills with trending and in-demand technologies helping users close career gaps.
- Implemented a **gaze detection** system using **CV** and **mediapipe** to monitor interviewee gaze during virtual interviews, resulting in a 30% improvement in process integrity.

AI/ML Intern 09/2022 – 12/2023
Prodigal AI Delhi, India

- Engineering **Anomaly Detection** system for company data by creating **data pipelines** and **data manipulation** techniques.
- Developed a **Gen-AI pipeline** for creating children e-book utilizing **Llama2** for generating captivating stories and titles and **Kandinsky** image generative model for cartoonized images for story through **FastAPI** endpoints.
- Worked on **Transformers** architecture for training in-house **LLMs** from scratch.
- Implemented an **ATS system** powered by **LLMs**, improving resume parsing accuracy by 35%.
- Developed **RAG** systems using **Langchain** enabling conversation with multiple file sources like links, pdfs, csv etc. in natural language and storing **chat history** and **conversation memory.**
- Worked on open source **Speech-to-Text(STT)** models like **Whisper (OpenAI)** and **Talking AI avatars.**

SKILLS

Python | PyTorch | Large Language Models (LLMs) | LlamaIndex | Langchain | FastAPI | Google Cloud Platform (GCP) | MLOps | Retrieval Augmented Generation (RAG) | CI/CD | GIT

EDUCATION

B.Tech - Artificial Intelligence and Machine Learning 2021 – 2025
Guru Gobind Singh Indraprastha University Delhi, India

High School 2018 – 2020
Ryan International School Ghaziabad, India

PROJECTS

Scene Sense 🔗

AI-powered Image Search Engine

- Scene Sense uses the **CLIP** model to understand natural language queries and image content, providing accurate and contextually relevant results.
- App handles multiple users, a robust user authentication system is implemented using **FastAPI.** User credentials are stored in **MongoDB** and it uses password hashing techniques and **JSON Web Token (JWT).**
- The app leverages **Google Cloud Platform** for storage and deployment and **Pinecone** for embedding management, enabling efficient and fast image retrieval.

CAMEL RAG 🔗

Roleplaying agent pipeline using RAG and custom google search agent

- This pipeline replicates roleplaying agents from **CAMEL research paper** to form context for the query.
- A custom google search agent is deployed for retrieving latest information given the task.
- All the unstructured data is fed into **FAISS vector database** for further retrieval.
- Followed by a **LLM** which generates a more informed and factually correct responses.

Dense Object Detection 🔗

Object Detection Model for Dense Environment

- The **YOLOv5x** model was trained from scratch on the **SKU110K dataset** with a maximum of 50 epochs, achieving a **mAP50 of almost 0.6** on the training set after 30 epochs.
- Quantizing the model to **INT8 precision** reduced the model size to **83 MB** and resulted in a minimal loss of **accuracy of 0.007**, with a **precision of 0.917**, **recall of 0.865**, and **mAP50 of 0.915.**
- Inference results showed that the **INT8 model performed the best in terms of accuracy**, outperforming the other models, including the PyTorch FP16 and Tflite FP32 models.
- It is suitable for **deployment on edge devices** with limited memory resources.

Customer Churn Prediction 🔗

MLOps pipeline with ZenML and MLFlow

- Comprehensive and modular pipeline with **data ingestion, data preprocessing, feature engineering, training using Logistic Regression, evaluation and fine-tuning using Grid Search.**
- This project uses **ZenML** for maintaining the pipeline with **MLFlow** for experiment tracking and deployment.

ACCOMPLISHMENTS

Co-Author in Research Paper 🔗

- "Enhancing Cloud Analytics with ZSearch: A Generative AI Chatbot for Real-time Insights".
- This paper was presented at the prestigious **14th International Conference on Recent Engineering and Technology (ICRET)** in 2023.

Lead at IEEE USAR

- Served as the Artificial Intelligence and Machine Learning Chapter Lead at the **IEEE GGSIPU EDC.**

LANGUAGES

English

Hindi