YASH KUMAR SINGH

AI/ML Engineer

+91 -6261312563 • Gorakhpur, Uttar Pradesh • Github • ys98931@gmail.com

ABOUT ME

Aspiring AI Engineer with a strong foundation in machine learning, deep learning, and data analysis. I ampassionate about building intelligent systems that solve real-world problems and am currently honing mytechnical skills through hands-on projects and continuous learning. Seeking an entry-level position to apply my expertise in a collaborative environment and contribute to cutting-edge AI innovations.

WORK EXPERIENCE

INFINITY AI Solutions - AI Intern **JULY 2025 - PRESENT**

Worked on end-to-end projects within the machine learning lifecycle, specializing in Computer Vision and Large Language Models. My experience includes developing and optimizing neural network architectures, managing large-scale datasets to enhance model performance, and contributing to the deployment of production-level AI solutions

- Supported the core team in implementing fundamental Generative AI practices, specifically assisting with the setup for fine-tuning Large Language Models (LLMs).
- Assisted in the preparation and preprocessing of large-scale datasets, ensuring data readiness to support the team's model training workflows.
- Collaborated with senior developers to test and deploy basic model integrations via REST APIs, helping to validate early-stage AI solutions .

PROJECTS

SnapDiet -AI powered Food Classification platform

Tech Stack : TensorFlow, Keras, OpenCV, FastAPI, React, Spoonacular API, REST APIs

- Developed a Convolutional Neural Network (CNN) with TensorFlow and Keras for multi-class classification of Indian cuisine, achieving precise food recognition from images.
- Engineered a responsive frontend using React, enabling users to upload food images in real-time for immediate analysis and interaction.
- Deployed a high-performance FastAPI backend to serve the CNN model, establishing an efficient API endpoint for prediction requests.
- Implemented RESTful endpoints to handle image processing and integrated predictions.
- Integrated the complete stack to seamlessly display the predicted dish and its nutritional breakdown enhancing the overall user engagement.

MediLLaMA - Fine-Tuned LLM for Medical Q&A

Tech Stack : LLaMA-2, Hugging Face (Transformers/PEFT), LoRA, PyTorch, Quantization

- Fine-tuned LLaMA-2-7B using LoRA (Low-Rank Adaptation) and 4-bit quantization for efficient training on limited GPU resources.
- Specialized the model for healthcare Q&A using the MedDialog (English) dataset, enhancing its ability to understand patient symptoms and generate medically relevant responses.
- Applied parameter-efficient fine-tuning (PEFT) with Hugging Face's transformers, datasets, and peft libraries for rapid iteration and low memory consumption.
- Designed a custom prompt-response format to mimic realistic patient-doctor interactions.
- Deployed and evaluated the model using Hugging Face pipelines to simulate clinical consultation scenarios.

CERTIFICATIONS

- **Career Essentials in Generative AI** – Microsoft & LinkedIn (**2024**)
- **Deep Learning A-Z** – Udemy (**2025**)
- **Generative AI for Beginners** – Udemy (**2025**)

EDUCATION

SHRI RAMSWAROOP MEMORIAL UNIVERSITY **2021-2025**

B.Tech in Computer Science

- Relevant Coursework - Specialization - AI/ML/DL
- CGPA - 7.26

SKILLS

- Languages – Python, C++, JavaScript
- AI/ML – TensorFlow, PyTorch, OpenCV, Hugging Face
- GenAI – LLMs, Transformers, RAG, LoRA, Fine-Tuning
- Tools – Git, Docker