Khushpreet Singh

khushpreets016@gmail.com (+91)6239550058 in/khushpreets016 github.com/Kalrakhush

SUMMARY

Results-driven AI Engineer with expertise in **machine learning (ML), artificial intelligence (AI), deep learning, and data analytics.** Proficient in building scalable AI solutions, deploying production ML models, and designing **Retrieval-Augmented Generation (RAG)** systems. Skilled in Python, PyTorch, TensorFlow, computer vision, and SQL. Experience with cloud technologies, model deployment, data preprocessing, and model optimization to enhance business outcomes. Strong ability to transform complex business requirements into actionable AI insights. Proven track record of improving model performance and deployment efficiency

TECHNICAL SKILLS

- **Programming Languages:** Python, SQL, Java
- **AI/ML Frameworks:** PyTorch, TensorFlow, scikit-learn, **NLP** techniques, Keras, OpenCV, Hugging Face Transformers
- **Machine Learning:** Deep Learning, Neural Networks**,** Natural Language Processing (NLP), Predictive Modeling, Feature Engineering
- **Agentic Ai:** CrewAI, LangGraph, LangChain, Google A2A, Multi-Agent Systems
- **MLOps & Deployment:** Model Deployment, Production ML, Model Monitoring, Experiment Tracking, Model Versioning, MLflow
- **Data Science Tools:** Power BI, Tableau, Excel, Pandas, NumPy, Matplotlib, Seaborn
- **Cloud Platforms:** Microsoft Azure, AWS, Google Cloud Platform (GCP)
- **DevOps:** Git, GitHub, CI/CD Pipelines, Docker, Data Pipeline, Automated Testing
- **Specializations:** Large Language Models (LLMs), RAG Systems, Fine-tuning, Prompt Engineering, Hyperparameter Optimization, A/B Testing
- **APIs:** REST API design, development, and integration

EXPERIENCE

AIML Engineer

FinAGG Technologies, Noida ***May 2025 – Present***

- Lead development of voice bot platform with continuous call experience, supporting real-time interruption and multi-turn dialog via WebSockets.
- Architected full calling system integrating telephony APIs (e.g., Ozonetel), audio recording, and PDF summary generation for post-call reports.
- Designed and implemented Invoice AI module: PDF parsing, data extraction, validation, and automated invoice reconciliation.
- Built AI video generation pipeline: frame-by-frame snapshot extraction, audio mapping, and stitched output using Python and FFMPEG.
- Optimized system for production deployment: containerization, CI/CD pipelines, monitoring, and logging, reducing release cycle by 30%.

Associate Ai Engineer

DeepForrest AI, Hyderabad **June 2024 – May 2025**

- Built and optimized ETL pipelines to process large-scale datasets (10TB+) for analytics and machine learning.
- Designed and deployed ML models using TensorFlow and PyTorch, reducing inference time by 25%.
- Engineered RAG systems by integrating vector databases, enhancing LLM contextual accuracy.
- Developed and fine-tuned LLM-based chatbots for enterprise applications and customer support, serving 1000+ daily queries
- Designed REST APIs for model deployment and consumption, integrating them seamlessly into enterprise applications with automated testing.
- Conducted hyperparameter optimization, improving model accuracy by 15%

PROJECTS

AI-Powered Document Processing Chatbot

December 2024 - January 2025

- **Developed a chatbot** enabling users to **query budget-related and legal documents**, assisting **500+ users** in financial and legislative research.
- Implemented PDF parsing using **PDFplumber** to extract key financial/legal terms from Republic Acts, budgets, and regulations.
- Integrated LlamaIndex for efficient indexing and retrieval of large-scale legal and financial datasets, reducing query response time by 60%.
- Built **NLP query handling** with OpenAI GPT models for structured document retrieval.
- Deployed production-ready application using FastAPI and Streamlit with real-time document processing capabilities
- **Stack**: Python, LlamaIndex, Langchain, OpenAI GPT, Pinecone, FastAPI, Streamlit, React, Elasticsearch, Docker.

Fine-Tuned Sentiment Analysis Model
January 2025 – February 2025

- Fine-tuned BERT-based sentiment analysis model using QLoRA (Quantized Low-Rank Adaptation) to improve classification accuracy for neutral reviews by 12%
- Curated and preprocessed custom dataset with 10,000+ balanced samples for positive, negative, and neutral sentiments, addressing class imbalance through advanced upsampling and synthetic data generation
- Integrated and optimized LoRA adapters using PEFT library (LoraConfig from Hugging Face), reducing model size by 50% while maintaining performance
- Achieved significant improvement in F1-score for neutral reviews and reduced misclassification between neutral vs positive/negative by 18%
- Deployed fine-tuned model for real-time review tagging in production sentiment analysis dashboard using Gradio
- **Stack:** Python, Hugging Face Transformers, QLoRA, PEFT, Datasets, PyTorch, Gradio, Weights & Biases, CUDA

AgriGuide – AI-Powered Crop Recommendation System

Github Link Website Link • November 2022 - April 2023

- Developed **ML models** analysing soil nutrient data to suggest optimal crops, enhancing agricultural yield.
- Integrated **meteorological data analysis** to predict rice disease outbreaks, reducing crop loss.
- Built a web app using **ReactJS and Flask** for real-time agricultural insights.
- Implemented ensemble methods combining regression models and decision trees for crop yield prediction with 88% accuracy
- Used **regression models and decision trees** for crop yield prediction.
- **Stack**: Python, Scikit-learn, Flask, ReactJS, Pandas, NumPy.

EDUCATION

B. Tech in Artificial intelligence and Machine learning, Computer Science

CGC Jhanjeri • June 2024 • 8.34 CGPA

12th

Non-Medical • Panacea Senior Secondary Public school • Jalalabad west • March 2019 • 78 %

10th

Panacea Senior Secondary Public school • Jalalabad west • March 2017 • 10 CGPA

CERTIFICATIONS

Microsoft Certified: Azure Data Scientist Associate - Microsoft
Google Data Analytics Professional Certificate - Google
Machine Learning Specialization - Coursera
Deep Learning Specialization - Coursera
Java Programming Certification - HackerRank
Python Programming Certification - HackerRank

PUBLICATION

The Agriguide : A Crop Recommendation System

Contributions:

- Conducted extensive data analysis on crop patterns and soil nutrients.
- Designed and implemented the AgriGuide system with integrated API endpoints for scalability.

INTERPERSONAL SKILLS

- Problem Solving Approach
- Team Collaboration
- Data Storytelling
- Communication and Presentation Skills