Saad Ashraf

Lahore, Pakistan | saadashraf3519@gmail.com | (+92) 310 8900129 | LinkedIn Profile

"Machine Learning and GenerativeAI Engineer | 1 Year of Experience"

SKILLSET

Toolset: Python | Numpy | Pandas | YOLO | Resnet | CNN | Transformers | Natural Language Processing | | Tensorflow | Pytorch | Scikit-learn | C++ | Fine-tuning | ARIMA | NLTK | Hugging face | Mediapipe | |GANs | Keras | Matplotlib | Langchain | RAG | LLMs |

Code Freak | Teamwork | English, Urdu – All professional proficiency or above

PROFESSIONAL EXPERIENCE

Simplelogix

AI/ML Engineer Oct 2025 – Present

1- Developed a machine learning model to predict patient bill payment behavior using classification algorithms; optimized data pipelines, feature engineering, and model evaluation for improved financial forecasting.
2- Built an end-to-end job search web application in Flask that unifies role search, company lookup, and NAICS-based industry exploration within a single responsive interface. Integrated SerpAPI for real-time job scraping and used the OpenAI GPT API to analyze each company—generating insights on hiring challenges, industry classification, and role-specific skill requirements.
3- Developed a security analysis tool that ingests CVE codes and automatically identifies associated exploits, impact severity, and potential risks. Integrated the OpenAI GPT API to generate clear, actionable remediation recommendations for each vulnerability, enabling rapid threat assessment and decision-making. Designed the system to streamline vulnerability management and enhance overall security posture.
4- Built an AI-driven voice chatbot that converts spoken queries into database insights using for speech-to-text and text- to-speech. Implemented an LLM-based natural-language-to-SQL engine that transforms user questions into executable SQL queries, retrieves results, and explains them back in clear, conversational language. Designed the system to enable hands-free, voice-controlled data interrogation with accurate, real-time responses.
5- Developed an end-to-end system that integrates Microsoft Graph API to access and transcribe Microsoft Teams meetings, processes the transcripts using OpenAI LLMs to extract action items, and automatically assigns tasks and tickets to relevant team members in ClickUp via its API. This solution streamlines post-meeting workflows by converting discussions directly into actionable tasks with minimal human intervention.
6- Built an automation system using Microsoft Graph API to access and process Microsoft Outlook emails, leveraging OpenAI LLMs to extract action items and automatically create and assign tasks or tickets to relevant team members in ClickUp via its API. The solution converts email discussions directly into structured, actionable tasks, improving productivity and reducing manual follow-ups.
7- Built a real-time AI Voice Assistant that auto-joins Microsoft Teams, Zoom and Google Meet meetings. The system uses python WebSocket server with OpenAI for conversational responses, React/Typescript frontend with Web Speech API for speech recognition (STT) and speech generation (TTS), and integrated with ClickUp API to provide context-aware answers about project tasks, status, progress report etc. Implemented features like barge-in interruption, conversation and history management.

Visionerds

AI Developer Aug 2025 – October 2025

1- Developed a a multi-agent conversational AI system with parent–sub agent routing using LangChain and LangGraph. Built Python backend for RAG pipelines, embedding logic (PDF, CSV, text), and real-time text/voice interactions with customizable LLM and TTS models.

AI Intern

Jun 2024 – Sep 2024

2- Perform camera calibration using checkerboard images, detects corners, calculates calibration values, saves calibrated images with markings, and demonstrates image undistortion, providing a comprehensive tool for camera calibration and distortion correction tasks.
3- Developed and evaluated a YOLO-based eye disease classifier with data preprocessing, training setup, and performance

analysis using precision, recall, mAP, and confusion matrix visualizations.

4- Built a web app that converts voice inputs into AI-generated images using Whisper for transcription, GPT-4 for prompt refinement, and DALL·E 3 for real-time image creation.

5- Developed a real-time driver drowsiness detection system using YOLOv5, integrated with OpenCV and Dlib for facial landmark detection. The system monitors Eye Aspect Ratio (EAR) and Mouth Aspect Ratio (MAR) to identify signs of fatigue, such as prolonged eye closure and yawning. Achieved high performance metrics with precision at 94.3%, recall at 94.8%, and an F1-score of 94.0%.

FAST-NUCES

Research Assistant Jul 2024 – Aug 2024

1- Developed a sentiment analysis model to assess Reddit users' opinions on gaming topics.
2- Utilized BERT embeddings for advanced text feature extraction.
3- Implemented a Random Forest classifier to categorize sentiments effectively.
4- Conducted data preprocessing to clean and prepare Reddit comments for analysis.
5- Visualized sentiment distribution and model performance using confusion matrices.
6- Leveraged Python and Jupyter Notebook for data science and machine learning tasks.
7- Enhanced understanding of public sentiment towards gaming discussions on Reddit.

SHOWCASE PROJECTS

Voice-EnabledRAGLegal Assistant

Enabled voice-based legal queries, simplifying research for professionals and students.

Tech Stack: Deepgram STT/TTS, LangChain, OpenAI API, FAISS

Voice-enabled RAG system using GPT and FAISS to answer Pakistan Law queries through natural interactions.

RAG ForDiagnostic For Clinical Notes DiReCTs

Enabled clinical query answering with contextual diagnostics using RAG and the MIMIC-IV-Ext dataset.

Tech Stack: Python, LangChain, OpenAI GPT, FAISS,

RAG-based clinical assistant using LLMs and vector retrieval to answer diagnostic queries from the MIMIC -IV-Ext dataset.

Emoji Math

Solved emoji-based math puzzles by interpreting emoji inputs using a lightweight fine-tuned Llama model.

Tech Stack: Python, DeepSeek-R1-Distill-Llama-8B, LoRA, TRL's SFTTrainer

Developed a lightweight model fine-tuned with LoRA adapters and 4-bit quantization to solve emoji-based math puzzles using DeepSeek-R1-Distill-Llama-8B.

Math Meme Repair

Corrected viral math meme errors using a fine-tuned Llama model with witty explanations.

Tech Stack: Python, DeepSeek-R1-Distill-Llama-8B, LoRA, TRL's SFTTrainer

Developed a fine-tuned Llama model to humorously correct viral math meme errors, providing witty and clear explanations.

Facial Feature Detection

Detected facial features like gender, glasses, and clothing color from images.

Tech Stack: Python, OpenCV, Pytorch, TensorFlow, efficientnetb0

Developed a facial feature detection system to identify gender, glasses, and shirt color from images using efficientnetb0 and KNN.

EDUCATION

National University of Computer and Emerging Sciences (FAST - NUCES)

Bachelor of Computer Science (BS CS) 2021 - 2025