Paawan Purdhani

+91 9958847613 paawanpurdhani@gmail.com LinkedIn

EDUCATION

Delhi University Delhi, India
Bachelor of Science (Honors) Electronics & Computer Science Nov 2020 – June 2023

- Key Courses – Python Programming, Embedded Systems, Database Management
- Authored the Research Paper **An Adaptive System For Predicting Student Attentiveness In Online Classrooms** published in Indonesian Journal of Electrical Engineering and Computer Science.

Delhi Public School, Dwarka Delhi, India
High School Diploma, PCM Apr 2017 – Apr 2020

- Scholar Badge Recipient (Award given for academic excellence Class V - XI)
- Sports Captain, Batch of 2020

PROFESSIONAL EXPERIENCE

HCLTech Noida, UP, India
Senior AI Engineer - Full Time April 2025 – Present

- Engineered and deployed custom AI/ML solutions aligned with specific client use cases across government, healthcare, finance, and retail, achieving up to **35% improvement in processing efficiency** and scalability across cloud-native environments.
- Designed and deployed AI Models and Agents On Premises NVIDIA GPUs and on Leading Cloud Platforms including Microsoft Azure, Google Cloud Platform (GCP), and Amazon Web Services (AWS).
- Presented AI solutions and proof-of-concepts to clients, effectively communicating complex technical concepts in a business-friendly manner.
- Skills: Python, Generative AI, Llama Index, FineTuning, Vector Databases, Model Deployment, FineTuning

CCS Computers Pvt. Ltd. New Delhi, India
AI Developer - Full Time June 2023 – March 2025

- Spearheaded the creation and deployment of various cutting edge Generative AI solutions for Clients leveraging technologies like RAG, Fine Tuning (Supervised FT, Parameter Efficient FT and Low Rank Adaption) and Prompt Tuning Large Language Models.
- Consulted and Tutored over 10 Clients on how to Develop AI Solutions and Integrate AI in your daily workload.
- Conducted AI Benchmark testing on various types of GPUs ensuring efficient operation and optimizing performance.
- Skills: Python,Generative AI, Retrieval Augmented Generation , Docker, Kubernetes, NVIDIA AI Stack

Team Computers New Delhi, India
Cloud Engineer - Internship July 2022 – October 2022

- Assisting in the design and deployment of 5+ scalable cloud solutions using Amazon Web Services.
- Managed and monitored AWS accounts for multiple organizations using Zabbix for supervision and performance tracking resulting in 50% reduction in infrastructure downtime while optimizing resource allocation.
- Skills: AWS, Linux

PROJECTS & PUBLICATIONS

RAG (Retrieval Augmented Generation) System for NIC Delhi January 2024

- Spearheaded Development of RAG Project in NIC Delhi for Internal Documents decreasing data analysis time by 70% for the organization.
- Integration the RAG Pipeline with the ongoing systems of the organization to use RAG on 50,000 documents.
- Technologies Used: Pytorch, Llama Index, Milvus DB, Llama 3.1 - 8B

Make Your Own Bot (Demo Video - Link) (Independent Project) December 2024

- Spearheaded the Creation of an AI Product where anyone can Build AI Bots for their Website without any Coding
- User has to just upload their Website Link / Document or PPT and an AI Chatbot will be created on that along with the Frontend Code to be inserted to User's Website.
- Used by over 20 Companies
- Technologies Used - Python, Llama Index ,NVIDIA NIM, FastAPI

Fraud Complaint Classification for Ministry of Home Affairs (MHA) July 2024

- Developed a Generative AI Model to Categorize Complaints on the Cyber Crime Portal (https://cybercrime.gov.in/) for 54 Categories
- Fintuned and Prompt Tuned Models (Llama 3.1 8B and Mistral 7B) .
- Quantized the AI Model for memory and GPU Optimization.
- Decrease Complaint Registration and Time to Action time by 40%
- Technologies Used: PyTorch, HuggingFace, CUDA, NVIDIA Triton Server

SKILLS

Python, Azure (AI Fundamentals Certified), Generative AI, AI Agents, AI Solutions (Fine Tuning and Prompt Tuning), Embedding Models, Model Deployment, Llama Index, Langchain, RAG, Vector Database, MilvusDB, Multi-Node Training, Pytorch, NVIDIA Stack (CUDA, Tensor-RT, Triton Server, NEMO, NIM for Inference), Linux, Docker, Kubernetes