# Avdhesh Singh Chouhan

London, UK

(+44) 7429146921

ai.avdhesh@gmail.com
https://www.linkedin.com/in/avdheshchouhan

## Professional Summary

**AI** and **Generative AI Engineer** with over **8 years of experience** designing, developing, and deploying **intelligent systems** across **cloud, edge, and embedded environments.** I specialize in building **end-to-end solutions** involving **large and small language models (LLMs & SLMs)**, including **RAG**, **memory-enhanced architectures**, and **agentic AI frameworks** for **autonomous decision making**. My work spans the **full AI stack — language and vision-based**, covering everything from **data preparation** and **model training** to **deployment** and **optimization** across various **compute platforms**.

A **self-driven** and **dependable professional**, I thrive in solving **complex challenges**, whether working independently or **leading cross-functional teams** to deliver **high-impact AI solutions**.

## Skills

AI, Generative AI, Agentic AI, Deep Learning, NLP, Computer Vision, Multimodal AI, LLMs, SLMs, RAG, Vision Language Models (VLMs), Transformers, Reinforcement Learning, PyTorch, TensorFlow, Hugging Face, Fine-Tuning, LoRA, QLoRA, PEFT, Prompt Engineering, LLM Evaluation, AI Evaluation Frameworks, Agentic Benchmarking, LangChain, LangGraph, CrewAI, AutoGen, Vector Databases, Knowledge Graphs, Pinecone, Neo4j, Model Serving, FastAPI, Docker, Kubernetes, MLOps, CI/CD, Python, C++, SQL, AWS, Google Cloud Platform (GCP), Azure, NVIDIA CUDA, TensorRT, Qualcomm Snapdragon, Edge AI, Microservices, Solution Architecture, Stakeholder Management, AI Strategy

## EXPERIENCE

### Turing Inc, London, UK - **Senior AI Consultant**

OCT 2025 - PRESENT

- **Designed and implemented a scalable agentic evaluation orchestration layer for a Google AI program**, enabling high-throughput benchmarking of multi-turn LLM agents across Gemini, OpenAI, and Anthropic models.
- **Built a containerized, parallel execution framework** using Docker and a threaded worker–queue architecture to reliably evaluate complex agent workflows at scale.
- **Developed automated auto-rating systems for agent reliability and instruction adherence**, leveraging execution traces, environment state deltas, and failure analysis to generate objective, reproducible evaluation metrics.
- **Engineered enterprise-grade observability and telemetry pipelines** using Google Cloud BigQuery and Protobuf to capture fine-grained agent reasoning steps, tool usage, and state transitions for data-driven model optimization.
- **Delivered end-to-end dataset validation and delivery pipelines for agent training and evaluation**, integrating GCP and Workspace APIs to ensure data integrity across the full agent development lifecycle.

### Capgemini Engineering, Bangalore, India - *Lead AI Engineer*

JUN 2022 - OCT 2025

- **Led design and deployment of scalable AI & GenAI systems across cloud and edge platforms** (AWS, GCP, Azure, NVIDIA Jetson, Qualcomm Snapdragon).
- **Defined AI solution strategy and roadmaps** in collaboration with C-suite, product leaders, and business stakeholders.
- **Architected an in-house GenAI accelerator platform**, automating the full AI lifecycle from data ingestion to CI/CD and production deployment using Docker and Kubernetes.
- Built and deployed production GenAI applications including **RAG pipelines, multimodal workflows, vector databases**, and enterprise LLM integrations.
- **Led and mentored a 10-member AI team**, driving hiring, upskilling, and R&D across LLMs, SLMs, and MLOps.
- Showcased agentic and domain-specific GenAI solutions at **CES 2025**, MWC, NRF, and CG Powerplay; **recipient of Capgemini's 2024 Annual Engineering Excellence Award**.

### ConnectWise Inc, Pune, India - *AI/ML Engineer*

SEP 2019 - JUN 2022

- **Managed cloud infrastructure** for enterprise data backup and recovery platforms, contributing to microservices-based architectures **supporting scalable AI/ML workloads.**
- Translated business and operational requirements into **ML-driven monitoring solutions**, delivering working prototypes to validate impact prior to production rollout.
- Led proof-of-concepts for **anomaly detection and fault prediction**, analyzing system telemetry and presenting actionable insights to technical and business stakeholders.
- **Integrated real-time ML pipelines into existing enterprise products**, partnering with DevOps and support teams to ensure low-latency inference, high availability, and smooth production adoption.

### Calsoft Inc, Pune, India - *Python Developer*

JUL 2017 - SEP 2019

- **Designed and developed data collection engines for converged infrastructure** (compute, storage, networking, virtualization), implementing backend logic in **Python** with **RESTful integrations**.
- **Contributed to the architecture and implementation of IaaS platforms**, leveraging **microservices architectures** and **Docker-based containerization** for scalable, modular deployments.
- **Built internal tooling and automation frameworks** to streamline **infrastructure monitoring, diagnostics, and lifecycle management**.
- **Supported AI/ML model validation and benchmarking**, developing **data pipelines and simulation environments** to evaluate performance and improve overall **system reliability**.

## EDUCATION

### *Master of Science, Machine Learning and Artificial Intelligence*

*Liverpool John Moores University, UK*

JUL 2020 - MAY 2022

### *Post Graduate Diploma, Machine Learning and Artificial Intelligence*

*International Institute of Information Technology, Bangalore*

JUL 2020 - SEP 2021

### *Bachelor of Engineering, Information Technology, HONS.*

*Rajiv Gandhi Technical University*

JUL 2013 - MAY 2017