

HARDIK

AI/ML Engineer

New Delhi, Delhi | +91 9310898123 | hardik21673@gmail.com
LinkedIn: linkedin.com/in/hardik0o0 | GitHub: github.com/HARD1Kk

PROFESSIONAL SUMMARY

AI/ML Engineer specializing in the **architecture, optimization, and deployment of production-grade AI Agents, Retrieval-Augmented Generation (RAG) pipelines, and multimodal chatbots**. Proven expertise **bridging advanced machine learning** (PyTorch, TensorFlow, Hugging Face) with **robust backend engineering** (Python, JavaScript, C++) to deliver secure, scalable AI solutions. Deeply experienced in **designing agentic frameworks, optimizing vector/relational retrieval pipelines, and eliminating hallucinations** to build **high-accuracy, enterprise-ready AI systems**.

KEY SKILLS

- **Languages & Core:** Python, C++, JavaScript/TypeScript, HTML/CSS
- **AI & Machine Learning:** Generative AI, Large Language Models (LLMs), RAG Pipelines, Prompt Engineering, MCP, A2A, PyTorch, TensorFlow, Hugging Face
- **Data & Databases:** PostgreSQL, MSSQL, ChromaDB (Vector DB), REST APIs, GraphQL
- **Ingestion & OCR:** PyMuPDF4LLM, Docling, PaddleOCR, Scrapy, BeautifulSoup
- **Cloud & DevOps:** Microsoft Azure, Docker, Git/GitHub, CI/CD pipelines, System Monitoring & Alerts

PROFESSIONAL EXPERIENCE

Software Development Engineer-1, **Mobcoder** — Noida

Nov 2025 – Present

- Developed and optimized large-scale backend data pipelines to support production AI applications.
- Established Azure cloud infrastructure to streamline data management and storage processes.
- Containerized microservices using Docker and scheduled daily ETL runs on Microsoft Azure to ensure zero-touch, automated pipelines.
- Collaborated with cross-functional teams to resolve product issues, improve system reliability, and deliver customer-facing features.

AI/ML Engineering Intern, **GetMyQuotation** — Delhi

Aug 2025 – Oct 2025

- Deployed a production-ready multimodal RAG chatbot integrating voice and visual inputs with grounded retrieval, enhancing customer support accessibility and user satisfaction.
- Designed and implemented a Retrieval-Augmented Generation (RAG) system (LangChain, MSSQL) that restricts outputs to cited source data, eliminating LLM hallucinations and ensuring accuracy.
- Contributed AI solutions to streamline vendor comparison workflows, helping customers find clearer, faster options for home renovation and furnishing.
- Assisted in developing software features for the customer quotation platform and collaborated with engineers to troubleshoot and resolve technical issues.

Artificial Intelligence and Data Science Intern, **CSRBOX** — Delhi

June 2024 – July 2024

- Developed and implemented a conversational AI chatbot utilizing large language models (LLMs) and NLP techniques to enhance user engagement.
- Optimized chatbot performance and user interactions through data analysis, improving response accuracy and user satisfaction.
- Collaborated successfully with a team of interns to accomplish project goals and resolve issues related to AI products.

KEY PROJECTS

DocuFlow — Enterprise Document Ingestion & RAG Orchestrator

Python, ChromaDB, Docling, PaddleOCR

- **Layout-Aware Ingestion & OCR:** Architected an ingestion engine processing PDFs, Word docs, spreadsheets, and images while preserving document structures and tables.
- **Structure-Aware Chunking:** Built chunking engine with regex detectors preserving Markdown headings and tables; indexed vectors into ChromaDB.
- **Advanced RAG Pipeline:** Engineered RAG orchestrator featuring Cross-Encoder Reranking and dual-layer LLM validation (Gatekeeper & Auditor) to eliminate hallucinations.

A2A Conversational Analytics ChatBot — Conversational Data Analysis Platform

Python, Next.js, Azure, LangChain, MSSQL

- **Conversational Analytics:** Architected a full-stack Next.js application enabling semantic search and conversational analysis over complex enterprise datasets.
- **Secure RAG Orchestration:** Engineered LangChain pipelines that strictly ground responses in verified MSSQL data to eliminate hallucinations.
- **Multi-Agent Framework:** Designed an Agent-to-Agent (A2A) system with adaptive routing, supporting both high-speed responses ('Fast Mode') and comprehensive information extraction ('Deep Research Mode').

Judicial Document Orchestration Pipeline — Legal Document Scraping & Storage Pipeline

Python, Scrapy, Azure, SQL, Docker

- **Mass Ingestion:** Built Scrapy scraper to crawl High Court websites; processed and OCR'd 1.4M+ documents to extract metadata. Optimizing data extraction, cleansing, and storage within a relational SQL database for rapid querying and downstream analysis.
- **Reliability & Orchestration:** Scheduled daily ETL pipeline containerized with Docker, logging, monitoring, and automated alerts.

AI Voice Assistant — Customer Support Automation

Python, LiveKit, Azure OpenAI, PyTorch

- **Real-Time Voice Pipeline Architecture:** Engineered a low-latency, conversational voice agent using LiveKit WebSockets, integrating **AssemblyAI** for real-time speech-to-text (STT), **Cartesia Sonic** for ultra-fast text-to-speech (TTS), and **Silero VAD** for accurate turn detection and voice activity monitoring.
- **Dual-Mode Conversational AI:** Built a unified customer support engine for the *GetMyQuotation* platform, combining an **OpenAI GPT-4o-mini** voice assistant with an **Azure OpenAI** text chatbot, implementing background noise cancellation (BVC) to ensure seamless, high-accuracy interactions.

Vision Based Agent — Vision-Based Computer-Use Automation

Python 3.13, LangGraph, Google GenAI, PyAutoGUI, Pillow

- **Autonomous Agentic Loop:** Engineered a continuous, closed-loop agent that interprets screenshots and leverages multimodal LLMs to execute precise mouse and keyboard actions for task automation.
- **Computer-Use Orchestration:** Designed a framework for executing complex, multi-step tasks described in plain English, enabling seamless cross-application interaction.
- **Systems Architecture:** Developed a specialized architecture for Python developers, demonstrating advanced integration of vision models with low-level OS automation tools.

EDUCATION

B.Tech in Computer Science & Engineering, Maharaja Agrasen Institute of Technology — Delhi	2022 – 2026
CGPA: 8.3 / 10.0	
Class 12th (CBSE), SRDAV Public School, Dayanand Vihar — Delhi	2020 – 2022
CGPA: 8.3 / 10.0	
Class 10th (CBSE), SRDAV Public School, Dayanand Vihar — Delhi	2018 – 2020
CGPA: 8.2 / 10.0	

CERTIFICATIONS

- **DSA from Coding Ninjas** - Completed the Data Structures and Algorithms certification from Coding Ninjas, gaining proficiency in core concepts such as arrays, linked lists, trees, graphs, dynamic programming, and problem-solving techniques.
- **LLM Engineering: Master AI, Large Language Models, and Agents** - Completed an advanced course covering large language model engineering with practical experience in Retrieval-Augmented Generation (RAG), LangChain frameworks, prompt engineering, fine-tuning, and deploying AI agents.
- **Full-Stack Web Development** - The Complete Web Developer: Zero to Mastery (Udemy) — Attained comprehensive full-stack expertise, acquiring proficiency in modern frontend and backend technologies including React, Node.js, SQL, and JavaScript.

EXTRACURRICULAR ACTIVITIES

- **Event Head at TnM (Techsurge & Mridang)**: Led planning and promotion, coordinating student volunteers, securing sponsorships, and implementing outreach to improve attendee engagement.
- **Hackathon Organizer**: Organized multiple hackathons, tech events, and seminars at MAIT to promote innovation and knowledge-sharing within the student community.