

TAHA AHMAD

taha.ahmad2215@gmail.com | +92-321-4870680 | [Portfolio](#) | [GitHub](#) | [LinkedIn](#) | Lahore, Pakistan

PROFILE

Full-Stack AI Engineer with **3+ years of experience** designing and delivering **production-grade AI systems, enterprise web applications, and scalable backend platforms**. Skilled in **LLMs, RAG, AI Agents, Computer Vision, Voice AI, FastAPI, React, Laravel, and cloud technologies**, with a proven track record of building **end-to-end products** that **automate business workflows, improve operational efficiency, and reduce costs**.

SKILLS & TOOLS

Frontend: HTML | CSS | JavaScript | React | Next.js

Backend: Python | FastAPI | Pydantic | REST APIs | AsyncIO | WebSockets | SQL

AI & ML: LLMs | VLMs | RAG | AI Agents | LangChain | LangGraph | MCP | Hugging Face | Fine-tuning | Embeddings | Vector Search | PyTorch | TensorFlow | OpenCV | Computer Vision | Voice AI (STT/TTS)

Data: PostgreSQL | Redis | ChromaDB | Pinecone | pg-vector

Cloud & DevOps: Docker | Google Cloud (GCP) | Vercel | Hostinger | Git | Linux | Stripe

EXPERIENCE

Symufolk — AI/ML Engineer

Lahore, Pakistan | Oct 2025 – Present

Tech-Stack: Python, PyTorch, FastAPI, React, SQL, Redis, Docker, GCP, YOLO, VLMs, LLMs, Computer Vision, OpenCV, Label Studio, Git, Linux.

- Built an **on-demand GPU inference platform** with **Docker** and **Redis**, enabling **asynchronous YOLO annotation** and **VJEP2 action recognition**. Automated **GPU scaling** based on queue demand, reducing **infrastructure costs by 60%** while delivering **scalable, low-latency inference**.
- **Fine-tuned** YOLO, VLMs, and LLMs for object detection, segmentation, and action recognition. Built augmentation pipelines, optimized GCP GPU training, and benchmarked pre/post-fine-tuning to improve accuracy, latency, and deployment readiness.
- Enhanced **Label Studio Community Edition** into an **enterprise-grade annotation platform** with **RBAC, organization management, annotation review workflows, GCP integration, YOLO-assisted auto-annotation, and annotation propagation**. Reduced manual labeling effort and eliminated **~\$70K/year** in licensing and annotation costs.
- **Built** an automated **benchmarking framework** for YOLO, SAM, Grounding DINO, and VJEP2, enabling **one-click evaluation** against ground-truth datasets. Generated accuracy, latency, and performance reports to accelerate model selection for production deployments.
- Led the **end-to-end AI MVP delivery**, driving **client discovery, solution architecture, project planning, and technical delivery**. Built the **React frontend, FastAPI + PostgreSQL backend**, architected **secure cloud audio/video storage**, and led **two junior engineers** through development and deployment.

Aspire Analytica — Software Engineer

Lahore, Pakistan | Jul 2023 – Jun 2025

Tech-stack: Python, FastAPI, Odoo, SQL, LLMs, RAG, AI Agents, LangChain, Hugging Face, ElevenLabs

- Built an **agentic AI financial reconciliation platform** using **RAG, tool-augmented agents, and React**.

Automated **order-to-payment reconciliation**, **vendor communication**, **sales performance analysis**, and **case escalation**, reducing reconciliation time by **80%**, maintaining **100% accuracy**, eliminating **~\$50K/year** in audit costs, and replacing spreadsheet-based workflows with **real-time operational visibility**.

- Built a **production-grade Voice AI platform** for the automotive industry using **FastAPI**, **RAG**, and a **self-hosted 7B LLM**. Automated **appointment scheduling** and **customer support** across multiple U.S. locations, increasing **appointment bookings by 15%**, saving **~\$10K/month** in support costs, and improving customer experience.
- Built a **Chrome-based web scraping platform** in **3 weeks**, processing **10K+ products** using **residential proxies**, **session management**, and **account rotation**. Automated **market-aligned pricing**, reducing manual effort and accelerating competitive price updates.
- Developed **backend modules** and **database schemas** for **5 ERP implementations** across **POS, HR & Payroll, Inventory, and Accounting**. Collaborated with clients on **requirements gathering**, **ERP workflow design**, **user training**, and **milestone delivery**, helping digitize operations and improve business efficiency.

PROJECTS

Twicsy UK

[Link](#)

Tech Stack: Laravel | PHP | JavaScript | MySQL | REST APIs | Stripe | Apify | PHPUnit | Git | Linux | SEO

- Architected and delivered a **Laravel 13 e-commerce platform** for digital services. Built **155+ dynamic pages**, a **custom CMS**, **SEO infrastructure**, **Stripe payments**, **RBAC**, and **API integrations**, while leading **deployment**, **optimization**, and **stakeholder collaboration**.

DigitsCode

[Link](#)

Tech Stack: React | Node.js | JavaScript | REST APIs | Hostinger | Git

- Architected and developed the **official DigitsCode corporate website** using **React** and **Node.js**. Built responsive pages showcasing company services, contact, and policy information, integrated dynamic content and contact forms, and led deployment, optimization, and ongoing maintenance on **Hostinger**.

FinEval

[Link](#)

Tech Stack: Python | Flask | React | Gemini | LLM | LLM Evaluation | Prompt Engineering | Git

- Built a **1,000+ question LLM benchmark** for **numerical reasoning**, **multi-step problem solving**, and **factual accuracy**. Implemented **LLM-as-a-Judge** evaluation pipelines and developed a **React analytics dashboard** for **performance comparison** and **error analysis**.

EDUCATION

BS Computer Science — GCU, Lahore

Develop Generative AI Applications - IBM

[Link](#)

Programming in Python - Meta

[Link](#)

Foundations of AI and Machine Learning - Microsoft

[Link](#)

AI and Machine Learning Algorithms and Techniques - Microsoft

[Link](#)

Building Intelligent Troubleshooting Agents - Microsoft

[Link](#)