USAMA NAWAZ

Email: usamanawaz789@gmail.com | **LinkedIn:** linkedin.com/in/usama-nawaz789 | **Phone:** +60137526242

Medium: medium.com/@usamanawaz789 | **Substack:** substack.com/@usamanawaz1

Website: usamanawaz.com

Summary

AI Engineer with 5+ years of experience building production-grade AI/ML systems, from multi-agent architectures and RAG pipelines to document intelligence and data platforms. Proven track record delivering end-to-end solutions: AI SaaS products, invoice extraction pipelines with 99%+ uptime, HS code classification at 95% accuracy, and ETL systems processing 25+ entity types. Proficient in Python, FastAPI, LangChain, LangGraph, CrewAI, OpenAI/Claude APIs, and cloud deployment on AWS and GCP.

Key Projects

- **AI Chatbot SaaS Platform** — Built a multi-tenant SaaS platform enabling businesses to embed customizable AI assistants. Features RAG pipeline with pgvector, real-time SSE streaming, embeddable React/TypeScript widget, intelligent lead capture, admin dashboard with analytics, WordPress plugin, and security hardening (SSRF protection, prompt injection sanitization, rate limiting).
  Tech: Python, FastAPI, React, TypeScript, PostgreSQL, pgvector, Docker, Claude/OpenAI APIs, Sentence-Transformers
- **AI-Enhanced Invoicing SaaS** — Designed a full-stack multi-tenant invoicing and ERP platform with AI agent system (orchestrator/planner/executor pattern) for natural language business queries. Supports invoices, quotations, credit notes, purchase orders, recurring invoices, financial reporting (P&L, aging, balance sheets), multi-currency, and trilingual support (English, Arabic, Urdu).
  Tech: FastAPI, Flutter, PostgreSQL, SQLAlchemy, Alembic, Redis, JWT, BLoC, ReportLab
- **Invoice Extraction & HS Code Classification** — Built a 9-phase AI extraction pipeline with cascading model fallback (Claude, GPT-5, GPT-4o) ensuring 99%+ uptime. Dual-approach HS classification combining FAISS + sentence-transformers with LLM-based classification. Supports PDF, images, DOC/DOCX, MSG/EML emails. Concurrency controls with semaphore-based thread limiting and token usage tracking.
  Tech: Python, Flask, OpenAI GPT-5, Claude API, FAISS, Sentence-Transformers, SpaCy, MySQL, AWS S3
- **Multi-Agent Customer Support** — Distributed multi-agent system using CrewAI automating customer service workflows with parallel ticket processing and intelligent task delegation. Specialized agents for classification, knowledge retrieval, and response generation, achieving 40% reduction in average response time.
  Tech: CrewAI, Python, LangChain, LLMs
- **Document Processing Automation** — End-to-end pipeline using computer vision and LLMs to process warehouse documents. Extracts and transforms unstructured data from PDFs and Word Documents into structured JSON with 90% accuracy, resulting in 60% efficiency improvement.
  Tech: Python, OpenCV, LangChain, OpenAI, PDF Processing
- **Customs Consolidation Validator** — AI-powered RESTful API for validating customs consolidation data. Multi-phase validation engine correcting product tags and flagging HS code misclassifications with domain-specific prompt engineering and concurrent async processing.
  Tech: Python, FastAPI, OpenAI API, Pydantic, AsyncIO
- **Financial Data Extraction Service** — Backend microservice automating financial data extraction, investment report parsing (Morgan Stanley, Robinhood, MidFirst Trust), transaction categorization via Plaid API, and travel document processing with two-stage AI classification pipeline.
  Tech: Python, Flask, OpenAI GPT-4o, Plaid API, Dropbox API, Docker
- **Shipping Document Extraction** — Production-grade system processing shipping labels, invoices, and logistics documents with multi-stage pipeline (OCR, vision analysis, JSON extraction). Barcode/QR detection across 8+ carriers (UPS, USPS, FedEx, DHL, Amazon) with fuzzy string matching for merchant validation.
  Tech: Python, Flask, OpenAI GPT-5, LangChain, FAISS, OpenCV, AWS S3, GitLab CI/CD
- **McMaster-Carr MCP Server** — MCP server enabling AI assistants to search, browse, and extract product data from McMaster-Carr's 700K+ industrial supply catalog. 7 MCP tools with browser automation, dynamic filter discovery, and dual-format output (Markdown & JSON).

Tech: TypeScript, Node.js, Puppeteer, Express.js, Zod, MCP SDK

- **Music Data Analytics Platform** — End-to-end system: multi-stage ETL pipeline transferring Chartmetric data from AWS S3 to BigQuery (25+ entity types) with incremental processing and automated VM lifecycle management, plus FastAPI/Flask API layer exposing 50+ endpoints with Redis caching and JWT authentication.
  Tech: Python, FastAPI, Flask, BigQuery, Boto3, GCP Cloud Storage, Cloud Run, Redis, Docker
- **AI Transaction Categorizer** — GPT-powered REST API categorizing bank transactions into custom accounting categories. Concurrent batch processing (20 workers), multi-strategy matching (exact, fuzzy, fallback), and real-time cost tracking.
  Tech: Python, FastAPI, OpenAI GPT, Pydantic, Docker, MongoDB
- **AI-Enhanced ERP System** — AI-powered ERP leveraging Frappe ERPNext with WhatsApp integration, enabling users to perform complex ERP functions through conversational NLP interfaces.
  Tech: Frappe, ERPNext, Python, WhatsApp API, NLP
- **Multi-Agent Article Generation** — Distributed multi-agent system using CrewAI to autonomously research, synthesize, and generate articles with specialized agents for research, writing, and editing.
  Tech: CrewAI, Python, LLMs, LangChain
- **Therapeutic Meditation Assistant** — Context-aware meditation guidance system using RAG architecture with LangChain, delivering personalized mindfulness experiences.
  Tech: LangChain, RAG, Python, ChromaDB, LLMs
- **Box Office Revenue Forecasting** — Prediction model using LightGBM achieving 85% accuracy in forecasting movie revenues through feature engineering and advanced regression techniques.
  Tech: Python, LightGBM, Scikit-learn, Pandas
- **Financial Transaction Security System** — Real-time fraud detection platform using ML algorithms for automated anomaly detection and alerting on suspicious financial activities.
  Tech: Python, Scikit-learn, XGBoost, Anomaly Detection
- **Intelligent Inventory Management** — AI-driven inventory optimization with predictive analytics reducing holding costs by 20% through improved demand forecasting and automated reorder calculations.
  Tech: Python, Predictive Analytics, Scikit-learn, Time Series

---

Professional Experience

Content Moderator
Concentrix Malaysia
August 2024 – Present

- Execute comprehensive content review processes according to established company policy guidelines
- Analyze and classify user-generated content to maintain platform integrity and safety standards
- Apply technical judgment to identify policy violations within the system
- Continue building AI engineering projects independently alongside this role

AI Engineer
DevStudio
May 2020 – August 2024

- Designed and implemented data pipelines to collect and store analytics from YouTube, Spotify, and TikTok for artist valuation and success prediction
- Built an automated HS code classification system using FAISS semantic search with sentence-transformers, achieving 95% accuracy across major e-commerce platforms
- Engineered document extraction systems using LangChain and OpenAI to transform unstructured PDFs, Word Documents, and other formats into structured JSON with 90% accuracy
- Developed NLP-driven chatbots reducing support tickets and improving user satisfaction
- Built anomaly detection systems using ML techniques to identify irregularities in large-scale datasets
- Optimized recommendation systems using collaborative filtering and neural networks, increasing engagement by 30%
- Conducted trend analysis and data visualization to support business decision-making

Teacher Assistant
Forman Christian College
Feb 2020 – July 2022

- Conducted coding labs and designed materials for courses like Compiler Construction, Data Structures, and MIPS Assembly Language.

Education

Bachelor's Degree in Computer Science
Forman Christian College, Lahore, Pakistan
2016 – 2019
CGPA: 3.195

Certifications

- **Claude Code in Action**
  Credential
- **AI Engineering - Specialization**
  Credential
- **Generative AI for Data Scientists - Specialization**
  Credential
- **Introduction to Machine Learning in Production**
  Credential
- **Python for Data Science, AI & Development**
  Credential
- **Supervised Machine Learning: Regression and Classification**
  Credential

Technical Skills

- **Languages:** Python, TypeScript, Java, C++
- **ML/AI:** TensorFlow, PyTorch, Scikit-learn, Keras, Hugging Face, OpenCV, NLP, Transformers, XGBoost, Pandas, NumPy
- **GenAI & LLMs:** LangChain, LangGraph, CrewAI, RAG, OpenAI GPT-4o/GPT-5, Claude API, Llama 3.2, AWS Bedrock, Sentence-Transformers, FAISS
- **Databases:** PostgreSQL, MySQL, MongoDB, BigQuery, Redis, Pinecone, ChromaDB, Supabase
- **Cloud & DevOps:** Docker, AWS (S3, EC2, Bedrock, CodeDeploy), GCP (Cloud Run, Cloud Functions, BigQuery, Cloud Storage), Oracle Cloud, Git, GitLab CI/CD
- **APIs & Web:** FastAPI, Flask, Django, React, Flutter, Node.js, Express.js, Streamlit, Pydantic

---

Personal Attributes

- Problem-solving mindset
- Excellent teamwork and communication skills
- Passionate about lifelong learning and professional growth
- Strong attention to detail and commitment to quality