

Muhammad Mubashir

AI/ML Engineer | Data Scientist | Full-Stack Developer | RAG & LLM Systems Specialist

mubashir.data555@gmail.com | [LinkedIn](#) | [Upwork](#) | [Portfolio](#) | +92-337-9854767 | Peshawar, Pakistan

PROFESSIONAL SUMMARY

AI/ML Engineer and Full-Stack Developer with 3+ years of experience designing and shipping production-grade AI systems - Retrieval-Augmented Generation (RAG) architectures, multi-agent and Model Context Protocol (MCP) systems, and full-stack SaaS platforms. Deep experience across legal-tech, government, healthcare, fintech, and consumer AI: built enterprise RAG systems processing 50,000+ documents, a political AI assistant sized for 10,000+ queries/day, PII-anonymization engines for legal case data, conversational AWS infrastructure agents on Amazon Bedrock, and AI-powered SaaS products integrating Claude, GPT-4o, Stripe, and Twilio. Comfortable across the full LLM stack - orchestration (LangChain, LangGraph, CrewAI, MCP), vector search (Pinecone, Elasticsearch), and business-automation tooling (GoHighLevel, n8n, Make.com). Top Rated on Upwork with a 100% Job Success score across 17 completed engagements (5.0-star average) and 6 more in progress. Comfortable owning a project end-to-end - architecture, backend implementation, testing, and deployment.

CORE COMPETENCIES

AI/LLM Engineering: Claude (Anthropic) API, Claude Code, OpenAI GPT-4o API, LangChain, LangGraph, CrewAI, Model Context Protocol (MCP), Retrieval-Augmented Generation (RAG), Prompt Engineering & Caching, AI Chatbot Development

Agentic AI & Orchestration: Multi-Agent Systems, Human-in-the-Loop (HITL) Workflows, Tool-Calling/Function-Calling, LLM Orchestration, State Checkpointing, Agent Memory

Machine Learning & Data Science: PyTorch, TensorFlow, Scikit-learn, Computer Vision, Natural Language Processing, Statistical Modeling, Pandas, NumPy

Backend & Full-Stack Development: Python, Node.js, JavaScript, FastAPI, Flask, Express.js, REST API Design, Git/GitHub

Data & Vector Infrastructure: Pinecone, Elasticsearch, MongoDB, SQL, Vector Search & Embeddings, FAISS

Cloud & DevOps: AWS (Bedrock, boto3), Docker, Render, CI/CD Fundamentals

Business Automation & Integrations: GoHighLevel (GHL), n8n, Make.com, Zapier, Stripe (Subscriptions & Webhooks), Twilio/WhatsApp Business API, OpenAI Whisper

PROFESSIONAL EXPERIENCE

Upwork - Freelance AI/ML Engineer & Full-Stack Developer

Jan 2023 – Present

Remote | Top Rated, 100% Job Success, 17 completed contracts (5.0 ★avg), \$10K+ earned

- Architected a production RAG platform (Datagreffe Justice) processing 50,000+ legal documents with Elasticsearch and Pinecone, cutting document-generation time by 80% and sustaining sub-2-second query responses at 95%+ confidence.
- Built a multi-jurisdiction PII-anonymization engine with a triple-search architecture for legal case decisions, reaching 98% detection accuracy, including added safeguards for protecting minors and other vulnerable parties.
- Designed a dual-endpoint (public/field) political RAG assistant with temporal query routing and namespace-based retrieval, integrating Claude, OpenAI, and Pinecone, sized for 10,000+ queries/day.
- Built the backend for an AI-powered recovery-support SaaS product (AI Sponsor): Claude-based chat with prompt caching (~90% input-token cost reduction), Stripe subscription billing (trials, webhooks), and a WhatsApp voice pipeline (Whisper transcription + synthesized voice replies).
- Built a conversational AWS infrastructure agent on Amazon Bedrock (Claude 3.5 Sonnet) with LangChain memory, letting users create, update, and manage AWS resources through natural language instead of the CLI.
- Engineered a Model Context Protocol (MCP) multi-tool agent platform on LangGraph with async state checkpointing, human-in-the-loop interrupts, and live tool-calling (web search, financial data).
- Built a CrewAI multi-agent pipeline that extracts, summarizes, and structures financial earnings-call transcripts into client-ready reports.
- Delivered computer-vision models for medical imaging (brain-tumor classification, diabetic retinopathy) in TensorFlow, and a 5-stage OCR-to-RAG pipeline converting scanned documents into vector-ready JSON.
- Integrated Amazon Bedrock session management for client applications and mentored clients on agentic AI frameworks (CrewAI, LangChain, LangGraph).
- Built a human-in-the-loop LinkedIn content and job-application AI agent (Content Copilot) using a three-agent architecture with web-search-grounded topic scoring.

ItecExperts Pvt Ltd - AI/ML Engineer

Oct 2023 – Feb 2024

Peshawar, Pakistan · On-site

- Contributed to data science and deep learning projects as part of the core AI engineering team, building and evaluating ML models.

Giant Tech Solutions LLC - Artificial Intelligence Engineer

Jul 2023 – Oct 2023

Peshawar, Pakistan · On-site

- Delivered AI/ML applications with a focus on statistical data analysis and deep learning.
“Collaborating with Mubashir, an exceptional AI engineer, has been truly inspiring. He consistently delivers top-notch results without compromising on deadlines or quality. His exceptional communication skills make complex AI concepts easy to understand.” - CEO, Giant Tech Solutions

Machine Learning Intern - CodSoft

Aug 2023 (1 month)

Remote (India)

- Completed a short-term internship applying core machine learning concepts on guided project work.

SELECTED PROJECTS

Datagreffe Justice - Enterprise Legal AI Platform

Flask · Pinecone · Elasticsearch · Claude · OpenAI · Python · Backend

Enterprise RAG system processing 50,000+ legal documents for a legal-tech client. AI document-generation module cutting creation time by 80%, with sub-2-second query response at 95%+ confidence. Includes a multi-jurisdiction PII-anonymization module with a triple-search architecture (98% detection accuracy) and dedicated safeguards for protecting minors and other vulnerable parties named in legal decisions.

RHDP Political RAG Assistant

Flask · Pinecone · Claude · OpenAI

Dual-endpoint (public/field) AI assistant built for a West African political organization, sized for 10,000+ queries/day. Implements temporal query routing (past vs. future program data) and namespace-based retrieval so the same system serves accurate, context-appropriate answers to two very different audiences from one knowledge base.

Conversational AWS Infrastructure Agent

Amazon Bedrock (Claude 3.5 Sonnet) · LangChain · Flask · boto3

Chat-driven AWS assistant that lets non-technical users create, describe, update, and delete AWS resources in natural language instead of the CLI - with mandatory field confirmation and explicit yes/no approval before any resource-changing action executes, to prevent accidental infrastructure changes.

MCP Multi-Tool Agent Platform

LangGraph · Model Context Protocol (MCP) · Async SQLite · Streamlit

Agent platform built on LangGraph's StateGraph with a Model Context Protocol client for external tool access, async SQLite-backed conversation checkpointing for persistent multi-session memory, and human-in-the-loop interrupt/approval steps before sensitive tool calls (e.g., live financial-data lookups) execute.

AI Sponsor - AI-Powered Recovery Support SaaS

Node.js · Express · Claude API · Stripe · Twilio

Backend for a 24/7 AI sponsor chatbot for 12-step recovery programs. Implemented prompt caching for ~90% input-token cost savings, Stripe subscription billing (trials, webhooks), and a WhatsApp voice pipeline combining Whisper transcription with synthesized voice replies. Built and passed a 12-category automated test suite covering chat, sessions, webhooks, and security.

Multi-Agent Financial Document Analysis

CrewAI · LangChain · Streamlit · PDF Extraction

Multi-agent pipeline (built with CrewAI) that ingests earnings-call transcripts and financial PDFs, delegates summarization and structuring across specialized agents, and outputs client-ready report documents.

Namhee - AI Therapy Chatbot (MVP)

FastAPI · Pinecone · LangChain · GPT-4o

Mental-health support chatbot with mood assessment and tracking, CBT/mindfulness-based guidance, session memory, and automatic risk detection that surfaces crisis helplines. Shipped production-ready with a 100% pass rate across 11 core automated tests and documented performance targets (sub-5-second RAG responses).

Content Copilot - Human-in-the-Loop LinkedIn & Job-Application Agent

Python · React · OpenAI · Web Search

Local-first, three-agent tool that drafts LinkedIn posts/comments and tailored job-application cover letters, with web-search-grounded topic scoring. Designed so every publish/apply action stays manual, avoiding platform ToS violations.

Computer Vision for Medical Imaging

TensorFlow · CNNs · Deep Learning

Built and trained CNN-based classification models for brain-tumor detection and diabetic retinopathy screening from medical scans.

OCR-to-RAG Pipeline

Python · OCR · Vector Embeddings

Five-stage pipeline (PDF → image → preprocessing → OCR → structuring) converting scanned documents into clean, vector-store-ready JSON for downstream RAG ingestion.

Real-Time NHL Match Outcome Predictor

Python · Statistical Modeling

Predictive model using a Poisson-distribution approach for match outcomes, achieving 95%+ backtested accuracy.

FREELANCE TRACK RECORD (UPWORK)

Top Rated badge · 100% Job Success Score · 23 total contracts (17 completed, 6 in progress) · 5.0★average rating on every rated job · \$10K+ total earnings · 324+ hours logged

“Muhammad was great, he went above and beyond the scope of work and truly cares about giving his clients the best experience - he is also an AI expert.” - Client, AI Specialist / Custom GPT Builder

“It has been a real pleasure working with Mubashir on this highly technical and time-sensitive legal-tech project. From the outset, he demonstrated strong professionalism and deep technical knowledge.” - Client, Backend Enhancement Project

“He is extremely knowledgeable and professional. I recommend him 100%.” - Client, 30-Minute Consultation

“Muhammad consistently demonstrated exceptional availability and support throughout the design and study phases of my project, which involved creating a chatbot. His expertise in this area is evident and invaluable.” - Client, Deep Learning Lab

CERTIFICATIONS

- IBM Data Science Professional Certificate - Coursera, Sep 2022 (Credential ID: VQXW63KF6KG4)
- Upwork Skill Certification - Generative AI - Sep 2024

EDUCATION

University of Engineering & Technology, Peshawar
Bachelor of Science, Computer Science (Data Science Focus)

Sep 2019 – Sep 2023

LANGUAGES

English (Fluent) · Urdu (Native) · Hindi (Conversational)