MUHAMMED AFSAL P M

AI Engineer | Agentic AI | LLM Engineering | RAG Systems

Kochi, Kerala, India | +91 98460 93491
muhammedafsalpmtech@gmail.com
LinkedIn | GitHub | Hugging Face | Portfolio | Kaggle
Notice Period: 30 Days
Preferred Locations: Kochi | Thiruvananthapuram | Bengaluru | Hyderabad | Chennai | Remote

Professional Summary

AI Engineer with 3+ years of experience building and deploying production-grade Generative AI and agentic applications across oil & gas, BFSI, healthcare and e-commerce. Experienced in multi-agent orchestration (LangGraph, CrewAI), RAG, NL-to-SQL, MCP integrations, LLM evaluation and vision-language models, using both OpenAI and self-hosted LLMs (Qwen, Llama via Ollama). Delivers end to end with Python, FastAPI, React/Next.js, Docker, Kubernetes, AWS and Azure. Fine-tunes and quantizes open-weight LLMs, with published Hugging Face models exceeding 1,000 downloads.

Technical Skills

Agentic AI & LLMs: Generative AI, Agentic AI, Multi-Agent Systems, LangGraph, LangChain, CrewAI, RAG, RAGFlow, NL-to-SQL, MCP, Tool Calling, Human-in-the-Loop, Prompt Engineering, Embeddings
Models & Providers: OpenAI GPT-4, Azure OpenAI, Gemini, Groq, Sarvam, Qwen, Llama, Qwen3-VL, LLaVA, Ollama, Hugging Face
Fine-tuning, Evaluation & LLMOps: LoRA / PEFT, SFT (TRL), llama.cpp, GGUF Quantization (imatrix), LLM-as-a-Judge, LLM Evaluation, Langfuse, Perplexity / KL Divergence
Backend & Frontend: Python, SQL, FastAPI, REST APIs, Microservices, Async Programming, Celery, RabbitMQ; Next.js, React.js, Angular, TypeScript, Tailwind CSS
Databases & Vector Search: PostgreSQL, MySQL, MongoDB, Redis, pgvector, Qdrant, ChromaDB, FAISS, Elasticsearch
Cloud & DevOps: AWS, Microsoft Azure, Docker, Docker Compose, Kubernetes, CI/CD (GitHub Actions, Drone CI), Git, SonarQube
Voice, Vision & Analytics: Speech-to-Text, Text-to-Speech (Piper), LiveKit, Twilio, YOLO, DeepSORT, scikit-learn, Power BI, Tableau

Professional Experience

NGXP Technologies LLP — Kochi, Kerala, India
Agentic AI Engineer — *06/2026 – Present*

- Built an NL-to-SQL database assistant for an oil & gas client using LangGraph multi-agent orchestration with self-hosted Qwen (Ollama) and OpenAI models, translating business questions into SQL over enterprise databases; added voice (STT, Piper TTS) and LLM-as-a-Judge response evaluation.
- Delivered crew management, well-slippage tracking, automated daily reporting, anomaly detection and a database-migration validator for the same client, with Next.js/React frontends on FastAPI; owned deployment, testing and production debugging.
- Architected a supervisor-based multi-agent platform (LangGraph) coordinating web-search, communication and verifier agents with parallel execution, failure replanning, human-in-the-loop approvals and code-enforced policy guardrails; containerised with Docker Compose.
- Implemented an enterprise document-intelligence (RAG) platform on RAGFlow with vector databases, self-hosted Qwen and OpenAI models; added Langfuse tracing and evaluation, persistent agent memory (Supermemory) and Drone CI/CD.
- Developed real-time driver-safety monitoring combining on-device detection with VLMs (Qwen3-VL 8B, LLaVA 7B) via Ollama and FastAPI to flag phone use, smoking, eating, drinking and seatbelt violations, with structured JSON output and confidence-tiered alerts.

Parity Cube Private Limited — Mumbai, Maharashtra, India
Integration & Automation Engineer — *04/2026 – 05/2026*

- Built a Model Context Protocol (MCP) server for Cuelinks, the company's affiliate-marketing platform, exposing its REST API as LLM tools with structured schemas and endpoint-specific prompts, so AI clients (ChatGPT, Claude) select the right API operation and payload.
- Developed an agentic competitor-intelligence platform (LangGraph, FastAPI, OpenAI, Docker) integrating affiliate-network APIs to compare product availability, offers and pricing.

Techvantage Systems Pvt. Ltd. Thiruvananthapuram, Kerala, India
Associate Software Engineer (AI) *12/2024 – 04/2026*

- Designed and deployed multi-agent systems (CrewAI; GPT-4, Groq, Llama via Ollama), including an intelligent loan-processing platform automating document verification, eligibility checks, risk assessment and decision-making for banking, insurance and audit clients.
- **Zentis AI** (zentis.ai; Techvantage-incubated agentic AI platform for regulated BFSI; core team): helped architect the multi-agent orchestration platform from inception, using microservices, autonomous agents, RAG, vector search, Azure OpenAI, Groq and LLM-as-a-Judge for workflow automation, document intelligence and audit.
- **Go-Do** (go-do.ai; multilingual voice and WhatsApp AI agents; core team): built conversational AI agents (FastAPI, LiveKit, Twilio WhatsApp API, Redis, MongoDB, FAISS-based RAG; Gemini, OpenAI, Sarvam) with STT/TTS, tool calling and persistent sessions, serving customers in multiple Indian languages.
- Built an HR-policy chatbot assistant (CrewAI, OpenAI) with agent-based intent classification, plus AI products for recruitment (resume shortlisting, adaptive tests, LLM-generated questions) and contract intelligence (clause detection, risk and compliance checks, contract Q&A).
- Shipped production services with FastAPI, MongoDB and RabbitMQ on Kubernetes (Azure, AWS): OTP authentication, multi-language RAG, fraud detection and automated reports; set up CI/CD, SonarQube and monitoring; also built computer vision (YOLO, DeepSORT) and ML pipelines (FTI, scikit-learn).

Bharat Petroleum Corporation Ltd. (Kochi Refinery) Kochi, Kerala, India
AI Engineer *12/2023 – 12/2024*

- Developed an AI chatbot with a Retrieval-Augmented Generation (RAG) pipeline that gives context-aware answers from internal refinery documents.
- Designed agent-based AI data-auditing and report-generation workflows using RAG to automate data validation, summarisation and compliance reporting.
- Extracted, validated and managed data from multiple sources with SQL/MySQL to ensure data quality and integrity for petroleum operations, adding agent-based anomaly detection.
- Developed and optimised machine-learning models and applied statistical and stochastic analysis to support optimisation of extraction, refining and distribution processes.
- Built interactive Power BI, Tableau and Excel dashboards with AI-generated insights and predictive analytics to improve operational efficiency and decision-making.

Open-Source Projects & Models

Qwen3.8-27B GGUF Quantization | huggingface.co/muhammed-afsal-p-m/Qwen3.8-27B-GGUF

- Reproducible llama.cpp pipeline converting Qwen/Qwen3.8-27B BF16 weights into 7 imatrix GGUF quantizations (Q8_0 to IQ3_M); 1,000+ downloads on Hugging Face.
- Q4_K_M cuts weight files from 54 GB to 17 GB with mean KL divergence 0.012 and 95% top-token agreement vs BF16 on held-out text; tool calling and vision tested via llama-server.

NL2SQL Agent | github.com/muhammedafsalpm/nl2sql-agent

- Multi-agent Text-to-SQL assistant for SQL Server (LangGraph) with planner, SQL author, verifier and synthesizer agents, a read-only SQL validator, charts and Excel/Word export; runs on OpenAI or local models (Ollama).

Agentic Document Intelligence | github.com/muhammedafsalpm/agentic-document-intelligence

- Agentic RAG platform with an intent router, ReAct analysis agent and verifier agent; hybrid BM25 + bge-m3 search in Elasticsearch with reranking; FastAPI, React/TypeScript, Docker.

Llama 3.1 8B Invoice Extraction (fine-tuned) | hf.co/muhammed-afsal-p-m/Llama-base-3.1-8B-invoice-gguf-sft

- LoRA (r=128) SFT with TRL/PEFT for structured JSON extraction from invoices and purchase orders; released adapter, merged weights and GGUF for local inference with Ollama.

Education

Bachelor of Technology, Computer Science and Engineering Kerala, India
APJ Abdul Kalam Technological University | Minor: Electronics & Communication *2019 – 2023*
CGPA 8.46 / 10 (First Class with Distinction)

Certified Specialist in Data Science and Analytics Kerala, India
ICT Academy of Kerala *05/2024 – 10/2024*

Certifications, Publication & Languages

- Microsoft Certified: Azure Data Scientist Associate (DP-100) (2024); Azure Data Fundamentals (DP-900) (2024)
- Databricks: Generative AI Fundamentals (2025); IBM Data Science Professional Certificate (2024)
- DeepLearning.AI: Introducing Multimodal Llama 3.2 (2025); Alteryx: Machine Learning Fundamentals (2024)
- Publication: “Smart Application for Visually Impaired People”, IJARIIE, Vol. 9, Issue 3, 2023
- Languages: English (professional), Malayalam (native), Tamil (conversational), Hindi (conversational)