

Huzaiifa Ahmad Gill

(+92) 301 0507689 | huzaiifaahmad2210@gmail.com | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)

TECHNICAL SKILLS

AI/ML & MLOps: PyTorch, TensorFlow, LangChain, LangGraph, LLaMA 3.x, OpenAI API, HuggingFace, RAG, Multi-Agent Systems, ChromaDB, MLflow, DVC, SageMaker, CNN, LSTM
Languages: Python, Go, Java, TypeScript, JavaScript, C#, C++
Backend & APIs: FastAPI, Django, Spring Boot, ASP.NET Core, Node.js, NestJS, gRPC, GraphQL, REST, WebSockets, Apache Kafka, RabbitMQ
Frontend: React, Next.js, Vue.js, Tailwind CSS, HTML/CSS
Databases: PostgreSQL, MySQL, MongoDB, Redis, DynamoDB, Neo4j, SQLite
Cloud & DevOps: AWS (EKS, Lambda, RDS, MSK, Bedrock, S3, DynamoDB, Kinesis), Terraform, Docker, Kubernetes, Helm, ArgoCD, Karpenter, KEDA, Istio, Ansible
DevSecOps & Observability: Prometheus, Grafana, Loki, OpenTelemetry, GitHub Actions, Jenkins, Checkov, Trivy, OPA/Rego, Infracost

EXPERIENCE

Machine Learning Intern

Apr 2026 – May 2026

Red Buffer — Onsite

[GitHub](#)

- Built **FlowForge**, an AI-powered planning platform that converts client proposals into validated timelines, Gantt charts, and architecture diagrams via a LangGraph multi-agent pipeline.
- Architected a 4-agent LangGraph workflow with self-healing retry loops that automatically recover from invalid or malformed agent outputs.
- Integrated HuggingFace Inference API and Together AI with LangChain prompt optimisation; shipped a Dockerised FastAPI REST service backed by PostgreSQL (SQLAlchemy + Alembic) for output versioning.

PROJECTS

Multi-Tenant SaaS Platform | *Go, NestJS, Spring Boot, gRPC, Kafka, EKS, ArgoCD*

Ongoing | [GitHub](#)

- Engineered a production-style polyglot microservices platform (5 services, each in its own repo) with independent CI/CD, versioning, and deployment lifecycle per service.
- Built a Go/go-kit **API Gateway** with Keycloak JWT validation, per-endpoint Redis token-bucket rate limiting, gobreaker circuit breaking (50% error threshold), and OpenTelemetry tracing over GraphQL and REST.
- Provisioned AWS infrastructure via 12 Terraform modules (EKS, RDS, MSK, ElastiCache, ECR) delivered through **ArgoCD GitOps** with Argo Rollouts Istio canary releases and KEDA autoscaling.

Kinetics MLOps Platform | *PyTorch, SageMaker HyperPod, EKS, MLflow, ArgoCD*

Ongoing | [GitHub](#)

- Delivered an end-to-end MLOps platform training a **CNN-LSTM** video action-recognition model on Kinetics-400, orchestrated by SageMaker HyperPod on EKS.
- Loaded a **435+ GB Kinetics-400** dataset from S3 into an **FSx for Lustre** POSIX file system (auto-import, lazy-loaded), pre-sharded by an offline PyAV→WebDataset ETL for optimal sequential-read throughput.
- Engineered **scale-to-zero GPU infrastructure** with Karpenter Spot provisioning for significant cost savings, and distributed DDP training with AMP bf16 + S3 checkpointing for auto-resume on Spot interruption.
- Wired a full lifecycle: MLflow tracking, approval-gated Model Registry, dual serving (SageMaker Endpoint + FastAPI), with Checkov, Trivy, OPA/Rego, and Infracost for DevSecOps and FinOps.

Multi-Agent RAG System | *LangGraph, Claude Sonnet, LLaMA 3.3 70B, ChromaDB, FastAPI, Vue 3*

[GitHub](#)

- Architected a multi-agent RAG system where a LangGraph Planner routes queries across three specialised agents — RAG (LLaMA 3.3 70B + ChromaDB), Web Search (Tavily), and Memory (SQLite).
- Ran each agent as a standalone MCP microservice with a ReAct-style Replan loop and a Claude Sonnet Aggregator; **cut end-to-end latency ~60%** via async parallel web search and fire-and-forget memory writes.

Voice AI Web Application | *Django, Next.js, Whisper AI, LLaMA 3.3, XTTS v2, Kubernetes, Helm*

[GitHub](#)

- Built a real-time voice AI assistant handling the full loop — Whisper + Silero VAD transcription, LLaMA 3.3 reasoning, XTTS v2 synthesis — decoupled via a RabbitMQ async queue for low latency.
- Engineered for scale with async Django (gRPC + WebSockets), PostgreSQL via PgBouncer, Redis caching, and Kubernetes deployment via Helm and GitHub Actions CI/CD.

Online Banking System | *Spring Boot, Next.js, Kafka, Keycloak, Kubernetes, Grafana, Prometheus*

[GitHub](#)

- Built a Spring Boot banking platform with **hexagonal (ports & adapters) architecture** across Account, Transaction, and Payment microservices, plus Kafka event-driven messaging and Keycloak SSO.
- Instrumented observability (Grafana, Prometheus, Loki, Tempo) and deployed to a local Kubernetes (KIND) cluster via Jenkins CI/CD with Helm and Testcontainers.

EDUCATION

University of Hertfordshire

BSc (Hons) Computer Science — Upper Second Class Honours (2:1)

Hatfield, UK

Oct 2022 – Aug 2025