Mahammad Shaik

AI/ML Engineer

(747) 202-5243• mahammadabzal74@gmail.com • New York, NY • https://www.linkedin.com/in/mahammad-shaik-b48410218/

SUMMARY

AI/ML Engineer with 5+ years of experience building production-grade Agentic AI, Generative AI, machine learning, and cloudnative solutions. Specialized in Python, LLMs, multi-agent systems, RAG, MCP, vector search, and scalable AI orchestration, with strong expertise across IBM Cloud, AWS, and Google Cloud. Proven track record of improving AI accuracy, reducing inference latency and cost, and deploying reliable ML/LLM services using Docker, Kubernetes, and modern MLOps practices.

PROFESSIONAL EXPERIENCE

AI/ML Engineer | IBM | New York, NY **May 2026 – Present**

- Engineered Python-based multi-agent orchestration services integrating LLMs, RAG, tool calling, and workflow routing, improving task execution and response reliability across enterprise knowledge and automation use cases.
- Applied LLM fine-tuning techniques (LoRA, QLoRA, PEFT) using Hugging Face Transformers, Trainer and Datasets; optimized inference with caching, batching and GPU profiling, reducing response latency by 31%.
- Implemented agent evaluation with Ragas, DeepEval, golden test sets and regression benchmarks for function-calling accuracy, retrieval precision, task success, groundedness and accuracy-versus-latency trade-offs.
- Built agent workflows using BeeAI, Python, MCP tools, and vector retrieval, supporting multi-step planning, contextual memory, tool execution, and coordinated task completion across enterprise AI applications.
- Developed RAG pipelines using Python, embedding models, document chunking, metadata enrichment, vector search, and reranking, enabling context-aware responses over enterprise knowledge sources and documents effectively.
- Integrated LLM services, REST APIs, FastAPI, Pydantic, and asynchronous Python components, enabling agents to securely execute business tools and return reliable structured outputs consistently across workflows.
- Designed containerized microservices on IBM Cloud with Docker and Kubernetes, separating API, agent orchestration, retrieval, evaluation, and observability workloads for scalable enterprise deployment and independent production releases.
- Deployed cloud-native AI services on IBM Cloud using Kubernetes, object storage, secrets management, and IAM controls, supporting secure and resilient workloads across development, staging, and production environments.
- Implemented CI/CD pipelines using Git, IBM Cloud container tooling, automated testing, Docker builds, security scanning, and Kubernetes deployments, accelerating releases across staging and production.
- Established AI service observability using OpenTelemetry, Prometheus, Grafana, structured logging, and agent traces, enabling production monitoring of response latency, token usage, failures, and workflow execution behavior.

AI/ML Engineer | Globant | San Francisco, CA **Sep 2024 – Apr 2026**

- Built Python-based multimodal pipelines that improved metadata processing accuracy by 27% across 1,500+ daily users, enabling scalable video ingestion, transcription, and semantic retrieval workflows in production.
- Reduced end-to-end search latency by 38% by optimizing vector retrieval, API orchestration, caching, and asynchronous processing, improving responsiveness for high-volume archive queries in production environments.
- Reduced inference cost by 33% through model routing, batching and caching, assessing GPU memory, throughput and latency trade-offs while preserving retrieval accuracy and production response quality.
- Developed Gemini and Vertex AI workflows for multimodal understanding, combining transcripts, frames, metadata, and embeddings to support natural-language search, contextual recommendations, and content discovery at scale.
- Implemented Python-based embedding and retrieval pipelines with vector indexing, semantic ranking, and metadata filtering, improving search relevance across large media archives and diverse content discovery queries.
- Built machine learning pipelines for transcript processing, content classification, entity extraction, and automated metadata enrichment using Python and cloud-native services across distributed workloads and media archives.
- Developed cloud-native media processing services across Google Cloud and AWS using containerized workloads, object storage, asynchronous pipelines, and scalable APIs for reliable video and metadata processing.
- Designed microservices for media ingestion, search, metadata processing, and content delivery, enabling independent scaling and reliable integration with downstream content management and streaming platform workflows.

- Implemented CI/CD workflows for Python and AI services using Docker, AWS, and Google Cloud infrastructure, enabling automated testing, deployment, rollback, and environment consistency across environments.
- Integrated secure APIs with authentication, authorization, monitoring, and fault handling, enabling downstream CMS and streaming platforms to reliably consume enriched metadata, semantic search, and content discovery services.

Machine Learning Engineer | Infosys |India **Jul 2020 – Dec 2023** • Developed Python-based ML pipelines using OCR, feature engineering, anomaly detection, and rule validation, contributing to 18% improved compliance exception detection accuracy across procurement transactions.

- Optimized Python inference workflows and model thresholds, contributing to 32% lower inference latency while maintaining 94% detection accuracy across compliance-risk screening and exception identification processes.
- Implemented Python-based OCR processing with SQL, Pandas, and scikit-learn, extracting compliance evidence from documents and supporting anomaly identification for downstream risk-scoring workflows.
- Developed feature-engineering workflows using Python, NumPy, Pandas, and SQL, transforming vendor, transaction, approval, and documentation data into structured features for ML models.
- Supported AWS data pipelines using Python, SQL, and object storage, processing procurement records, preparing training datasets, generating features, and enabling scalable batch inference workflows.
- Contributed to AWS-based ML service deployments using Docker, Kubernetes, and REST APIs, supporting reliable model inference, compliance-risk scoring, and integration with downstream enterprise applications.
- Worked on microservices for OCR, rule evaluation, risk scoring, and alerts, supporting modular deployments, fault isolation, scalable processing, and integration across compliance monitoring workflows.

TECHNICAL SKILLS

Programming & Software Engineering: Python, SQL, FastAPI, Pydantic, REST APIs, AsyncIO, Object-Oriented Programming, Data Structures & Algorithms, Microservices, Software Engineering, Software Development, Debugging, C++, Linux

Generative AI & Agentic AI: LLMs, Agentic AI, AI Agents, Multi-Agent Systems, BeeAI, LangGraph, MCP, Tool Calling, Function Calling, RAG, Prompt Engineering, Agent Memory, Agent Evaluation, Artificial Intelligence, Automation

Machine Learning & NLP: Scikit-learn, Pandas, NumPy, Feature Engineering, Anomaly Detection, Classification, Entity Extraction, OCR, Model Evaluation, Model Optimization, Machine Learning, Deep Learning, PyTorch, TensorFlow, Neural Networks, Natural Language Processing (NLP), Computer Vision, OpenCV, Predictive Modeling, Model Training, Reinforcement Learning, Forecasting, Clustering, Decision Trees, Statistics, A/B Testing

LLM & Retrieval Systems: Embeddings, Vector Databases, Vector Search, Semantic Search, Hybrid Search, Reranking, Document Chunking, Metadata Enrichment, Ragas, DeepEval

Cloud & AI Platforms: IBM Cloud, AWS, Google Cloud Platform (GCP), Vertex AI, Gemini, Object Storage, IAM, Secrets Management, Cloud-Native Services, GPU Computing

MLOps, DevOps & Infrastructure: Docker, Kubernetes, CI/CD, Git, Model Deployment, Model Serving, Automated Testing, Security Scanning, Containerized Workloads, DevOps, Kubeflow, Apache Airflow

Data Engineering & Distributed Systems: SQL, Data Pipelines, Batch Processing, Asynchronous Processing, Event-Driven Architecture, Distributed Systems, Caching, Data Processing Pipelines, Scalability, Data Analysis, Apache Spark, Kafka, BigQuery

Monitoring & Observability: OpenTelemetry, OpenInference, Prometheus, Grafana, Structured Logging, Distributed Tracing, Agent Tracing, LLM Observability, Performance Monitoring, Performance Metrics

Soft Skills: Technical Leadership, Problem Solving, System Design, Technical Communication, Agile Development, Code Review, Mentoring, Production Troubleshooting, Planning, Decision Making, Prioritization, Continuous Improvement

CERTIFICATIONS

AWS Certified Machine Learning Engineer – Associate
IBM Generative AI Engineering Professional Certificate
Certified Kubernetes Application Developer (CKAD)

EDUCATION

Master of Science in Computer Science and Information Systems
California State University Dominguez Hills