MOHCINE MADKOUR

Senior AI Software Engineer · Agentic AI Systems · LLM Platform Engineering

Plano, TX
281-652-7118
mohcine.madkour@gmail.com
LinkedIn · GitHub

I build AI systems that run themselves — multi-agent orchestration pipelines, self-monitoring feedback loops, and LLM platforms that ship to production and stay there. At Intuitive Surgical, I delivered a supervisor-orchestrated multi-agent system processing real-time robotics telemetry at 99.9% uptime with zero babysitting. At Cummins, I built prognostic systems that autonomously triggered repair workflows at fleet scale. I've genuinely rethought how I work because of agentic AI — I use it every day to eliminate repetitive work, accelerate iteration, and focus on the hard problems. I'm not just an AI expert — I'm an AI-native engineer who builds things other engineers look at and ask 'how did you do that?'

TECHNICAL SKILLS

Agentic AI	Multi-agent orchestration (LangGraph, AutoGen), ReAct patterns, tool-calling, supervisor architectures, human-in-the-loop checkpoints
Self-Improving Systems	Feedback loops, continuous evaluation pipelines, data flywheels, production drift detection, auto-retraining triggers, LangSmith observability
LLM Fine-Tuning	Domain-specific dataset prep, fine-tuning runs (Hugging Face), rigorous pre/post evaluation, prompt optimization, model tradeoff analysis
RAG & Retrieval	Hybrid search, semantic caching, reranking, vector DBs (Pinecone, pgvector, Weaviate, Qdrant, ChromaDB), enterprise Q&A pipelines
MLOps & Infra	MLflow, CI/CD for ML, Docker, Kubernetes, AWS SageMaker, Azure ML, GCP Vertex AI, Databricks, OpenTelemetry
Stack Overlap	Node.js / TypeScript (familiar), Python (10+ yrs), SQL, REST APIs, Elasticsearch, cloud-native SaaS architecture

PROFESSIONAL EXPERIENCE

Senior AI Architect & ML Engineer | Intuitive Surgical *Jul 2021 – Jan 2026 · Remote*

- **Agentic AI (Production):** Designed and shipped a production multi-agent LLM platform for surgical robotics telemetry — supervisor-orchestrated architecture with four specialized agents (Data Analysis, RAG Chat, Visualization, Formatting) running at 99.9% uptime with no manual intervention; achieved 25% lift in proactive service rates.
- **Self-Improving Systems:** Built self-improving GenAI pipelines with automated evaluation, A/B testing, and LangSmith observability — systems that measured their own quality over time and auto-adjusted; delivered 97% maintenance effectiveness gain and 40% data reliability improvement.
- **Feedback Loops & Data Flywheels:** Designed continuous feedback loops where production signals fed back into prompt optimization and fine-tuning — closing the loop between user interactions and model improvement without manual re-training cycles.
- **RAG & Retrieval Systems:** Built enterprise RAG pipelines with hybrid search, semantic caching, reranking, and pgvector/Pinecone integration — enabling context-aware Q&A for field engineers over domain-specific knowledge bases.
- **Responsible AI:** Architected production-grade responsible AI guardrails: prompt injection defense, PII-aware pipelines, bias detection, and full model auditability — meeting enterprise security and compliance standards.
- **Technical Leadership:** Led cross-functional AI engineering squads; established ML architecture standards, conducted code reviews, maintained Agile backlogs across concurrent AI initiatives.

Senior Data Scientist & ML Engineer | Cummins Inc. *Oct 2018 – Jul 2021 · Columbus, IN*

- **Autonomous Production Systems:** Architected PreventTech — a distributed fault isolation intelligence system over a 500k+ vehicle fleet using PySpark, causal inference, and unsupervised clustering; built a production API-integrated prognostic platform that autonomously triggered repair workflows at org scale, delivering $700K in annual savings.
- **Self-Monitoring Fleet Intelligence:** Designed Connected Diagnostics: multivariate anomaly detection with time-series forecasting (ARIMA, Prophet) and hierarchical failure stratification — deployed, monitored, and continuously improved end-to-end; improved engine reliability 17% and saved $110K.
- **Data Engineering & MLOps:** Built scalable ETL pipelines and feature engineering frameworks on Databricks and AWS SageMaker — owned full ML lifecycle from data collection through validation, deployment, and ongoing performance monitoring.

Assistant Research Professor — Data Science & AI | UF Shands Hospital / UF Health *Jun 2017 – Sep 2018 · Gainesville, FL*

- **Clinical AI Platform:** Conceived and clinically deployed MySurgeryRisk — a real-time surgical AI risk platform via FHIR/EHR integration; probabilistic scoring for 8 postoperative complications achieving AUC 0.82–0.94; adopted by emergency physicians after rigorous validation.

SELECTED AI SYSTEMS BUILT

- **Industrial IoT:** Predictive Maintenance RAG System — full-stack: XGBoost anomaly detection, ChromaDB vector store, Claude API orchestration, Streamlit dashboard, automated evaluation pipeline (AI4I 2020 dataset)
- **Medical Robotics:** Da Vinci Surgical Robotics RAG Agent — LangGraph + ChromaDB multi-agent system with FRD/TDD documentation; domain knowledge retrieval for surgical robotics engineering
- **Computer Vision:** Warehouse Computer Vision System — YOLOv8 + ByteTrack + QR code tracking (pyzbar) with RTSP/webcam support; real-time worker and asset tracking pipeline
- **LLM Application:** Voice AI Agent (Gen AI) — hybrid SQL + ChromaDB architecture over IMDB top 1000 dataset; real-time query routing and NL-to-SQL for a Gen AI conversational agent

EDUCATION & CERTIFICATIONS

PhD, Computer Science
University of Mohamed 5 Agdal, Morocco

Postdoctoral Fellow, Biomedical Informatics
UT Health Science Center, Houston

Azure Data Scientist Associate
Microsoft Certified

Certified Professional Mirth Connect Developer
NextGen Healthcare