Hameem M Mahdi, B.S.C.S., M.S.E., Ph.D.

SENIOR PRINCIPAL APPLIED SCIENTIST

+1 (779) 200-5027 | mahdi.hameem@gmail.com | github.com/HAMEEMM | linkedin.com/in/HAMEEMM

PROFESSIONAL SUMMARY

Agentic AI & Production ML Systems Architect — 15+ years building and shipping multimodal LLM, RAG, and computer-vision systems at enterprise scale. Led deployments generating $2.3M+ annual savings (45% compute reduction), scaled services to 500K+ daily users (99.5% uptime), and mentored 20+ researchers. Core expertise: LLMs & GenAI, RAG, MLOps (K8s, vLLM, Triton), model optimization/quantization, healthcare cloud integration (Oracle OIC), and privacy-preserving ML for regulated environments.

TECHNICAL SKILLS

Programming Languages

Python (Primary), Jupyter Notebook, MATLAB, NumPy, pandas, R, TypeScript, JavaScript • ***Systems***: Bash, C/C++, Go, Java, Rust • ***Mobile***: Kotlin, Swift • ***Web***: ASP.NET Core, C#, CSS3/SASS, HTML5, MVC, Perl, PHP, XML • ***Database***: SQL, T-SQL, PL/SQL, PL/PgSQL, NoSQL, OQL, XQuery • ***Scientific***: Dask, Julia, Scala

AI/ML Frameworks & Tools

Core Frameworks: PyTorch, TensorFlow, JAX, scikit-learn, XGBoost, Hugging Face (Transformers, Datasets, Accelerate, Hub) • ***Distributed Training & Acceleration***: PyTorch Lightning, DeepSpeed, Horovod, Distributed Training, Multi-GPU Training, NCCL • ***NLP & Language Models***: BERT, SentenceTransformers, spaCy, NLTK, Gensim, LangChain, LlamaIndex • ***Large Language Models (LLMs)***: GPT-4, Claude, Llama, Mistral, Gemini, Mixtral • ***Fine-Tuning & Alignment***: PEFT, SFT, LoRA, Quantization, RLHF, DPO, PPO, GRPO, Reward Models, Prompt Tuning, Instruction Tuning • ***Computer Vision & Multimodal***: Vision Transformers (ViTs), CNNs, ResNet, YOLO, Diffusion Models, Stable Diffusion, Flux.jl, DiT, LLaVA, LayoutLM, TrOCR, SimCLR, OpenCV • ***Model Optimization***: Pruning, Distillation, Model Compression, Transfer Learning, ONNX, ML.NET • ***Deployment & Serving***: vLLM, TensorRT, Triton Inference Server, BentoML, FastAPI, Captum • ***Explainability***: SHAP, LIME, Intrinsic & Post-hoc Methods, Optuna

AI Systems Types

Agentic, Analytical, Autonomous (Non-Agentic), Bayesian/Probabilistic, Cognitive/Neuro-Symbolic, Conversational, Evolutionary/Genetic, Explainable (XAI), Federated/Privacy-Preserving, Generative, Multimodal Perception, Optimization/Operations Research, Physical/Embodied, Predictive/Discriminative, Reactive, Recommendation/Retrieval, Reinforcement Learning, Scientific/Simulation, Symbolic/Rule-Based

ML & Data Engineering

Apache Airflow, Apache Atlas (Data Lineage), Apache Beam, Apache Spark, Batch/Stream Processing, Data Lineage, Data Preprocessing, Data Versioning (DVC), dbt, Delta Lake, Distributed Training, ETL Pipelines, Feature Engineering, Feature Stores (Feast), Hadoop, Kafka, Large-Scale Data Analysis, multicore SMP, Parquet, PySpark, Signal/Image Processing, Trino/Presto, Vector Databases (FAISS, Milvus, Weaviate, Pinecone)

MLOps & Deployment

Container & Orchestration: Docker, Kubernetes, Helm, Kustomize • ***Workflow & Training***: Argo Workflows, Kubeflow, Distributed Training, NCCL • ***Model Serving***: Model Serving (vLLM, Triton Inference Server, BentoML, Seldon Core, KServe), Ray (Ray Tune/Ray Serve) • ***Model Management***: MLflow, TFX, Model Registry, Model Optimization, RAGAS (RAG Evaluation) • ***Architecture & APIs***: Microservices Architecture, API Integration, gRPC, REST, Celery/RabbitMQ • ***Security & Scanning***: Container image scanning (Trivy, Clair), Container Registries (DockerHub, ECR, GCR) • ***Enterprise***: Enterprise AI Deployment

Cloud Platforms & Services

Cloud Platforms: AWS (CloudWatch, EC2, EKS, Lambda, S3, SageMaker), GCP (BigQuery, GKE, Vertex AI), Azure ML • ***ML-Specific Services***: Vertex AI, SageMaker, Azure ML, Azure OpenAI APIs, OpenAI APIs • ***Infrastructure & Compute***: Google Cloud TPU, Cloud-Native ML Infrastructure, Cloud Security (IAM, VPC) • ***Data Warehousing***: Snowflake, BigQuery

Data Storage & Databases

Vector Databases: FAISS, Milvus, Pinecone, Weaviate • ***SQL Databases***: MySQL, Oracle, PostgreSQL • ***NoSQL***: Cassandra, MongoDB, Redis • Graph Databases: Neo4j • ***Search & Analytics***: Elasticsearch/OpenSearch • ***Data Formats & Architecture***: Apache Iceberg, Delta Lake, HDF5, Parquet, Data Lake Architecture

Hardware & GPU Acceleration

Automatic Mixed Precision (AMP)/FP16, bfloat16 (BF16), cuDNN, CUDA, CUDA Toolkit/nvcc, Edge TPU (Coral), FlashAttention, GPU Computing, Hardware Acceleration, Multi-GPU Training, NCCL, NVIDIA DGX, NVIDIA Jetson, ONNX Runtime, OpenVINO, ROCm, TensorRT, Triton Inference Server, vLLM

Edge Computing & Mobile ML

CoreML, Edge Computing, Embedded AI Systems, Model Quantization for Edge, Mobile Deployment Optimization, On-Device Model Performance Optimization, ONNX Mobile, TensorFlow Lite

Monitoring & Evaluation

Experiment & Model Tracking: Weights & Biases (Experiment Tracking), Model Performance Monitoring, A/B Testing • ***Infrastructure & System Monitoring***: Prometheus, Grafana, OpenTelemetry, Sentry • ***Data Quality & Evaluation***: Great Expectations, Evidently AI, WhyLabs, RAGAS (RAG Evaluation) • ***Analytics & Visualization***: Tableau

Dev Tools & Versioning

API Integration, CI/CD Pipelines (GitHub Actions, GitLab CI, Jenkins), Docker Compose, DVC (Data/Model Versioning), Git/GitHub/GitLab, Helm, Infrastructure as Code, Terraform, Version Control Best Practices, pytest

Visualization & Analysis

Analytics Workflows, Dashboard Development, Data Storytelling, EDA, Matplotlib, Plotly, Seaborn, Tableau, V-Ray

Security, Privacy & Governance

Explainability & Interpretability: Explainable AI (XAI), Model Interpretability (SHAP, LIME), Intrinsic & Post-hoc Explainability Methods, Model Cards/Datasheets for Datasets • ***Privacy & Security***: Differential Privacy, Privacy-Preserving ML, Federated Learning, Adversarial Robustness/Robust ML, Secure ML Pipelines • ***Governance & Compliance***: Compliance & Governance (GDPR), Data Lineage/Provenance

Research & Quantitative Skills

Research Areas: Bioinformatics, Healthcare AI, Quantum Machine Learning, Neuromorphic Computing, Signal/Image Processing, Edge Computing, Causal Inference, Optimization (Convex/Nonconvex) • ***Academic Publications***: NeurIPS, CVPR, ICLR, ICCV, ICML, ACL, ECCV, KDD, ICASSP, InterSpeech Publications, arXiv Preprints • ***Impact & Contributions***: Patent Development, Open-Source Contributions

PROFESSIONAL EXPERIENCE

The Technology Innovation Institute (TII) November 2024 – **Present**
Sr. Principal Applied Scientist Contract, Remote

- Led production deployment of agentic AI and RAG systems (vision + NLP) for 3 enterprise clients; architected model-serving pipelines (Kubernetes, vLLM/Triton) and inference optimizations that reduced computational overhead 45% (≈ $2.3M annual savings) while maintaining SLA targets. Mentored 20+ researchers through architecture reviews and design patterns for safe tool use and alignment.

Mayo Clinic November 2015 – **Present**
Sr. Cloud Integration Engineer Remote

- Architected and governed Oracle Fusion Cloud integrations (OIC) with legacy EHR systems—designed 300+ real-time and batch integration flows using FBDI/HDL automation—reducing manual processing by 71% and integration latency by 50%. Implemented HIPAA-compliant data pipelines and federated training patterns for privacy-preserving analytics.

Note*: Contract roles with TII, G42, and IT-Serve were performed concurrently with Mayo Clinic responsibilities (20 hrs/week max), with explicit written approval and no conflicts of interest.*

G42 October 2020 – November 2024
Principal Applied Scientist Contract, Remote

- Led design and optimization of 6+ deep learning/Generative AI models for finance, energy, and healthcare. Achieved 92% accuracy on multimodal tasks and 48% inference latency reduction, driving the research-to-product pipeline for 4 major enterprise deployments.

IT-Serve October 2015 – October 2020
Sr. Applied Scientist Contract, Remote

- Spearheaded deployment of 8+ ML and GenAI models for computer vision and NLP. Served 500K+ daily users with 99.5% uptime and achieved a 40% reduction in latency, contributing 2 patents and 4 peer-reviewed publications.

Mayo Clinic November 2013 – November 2015
Data Analyst Contract, Rochester, MN

- Developed predictive ML models for high-risk patient cohorts, achieving 89% accuracy and a 23% reduction in readmissions. Analyzed 5M+ EHR records to build analytics dashboards across 7 hospital systems.

Samsung Ads September 2010 – November 2013
Applied Behavioral Scientist Chicago, IL

- Integrated behavioral science frameworks into health apps, driving a 42% increase in conversion rates. Designed 25+ A/B tests and rapid experimentation protocols using the COM-B model.

Cleo Communications August 2008 – September 2010
Sr. Software Engineer II Rockford, IL

- Engineered scalable SaaS components and an EDI translation engine, achieving 52% performance optimization and 67% error reduction via TypeScript/Node.js solutions.

EDUCATION

- **National University** San Diego, CA
  Ph.D. in Data Science, GPA: 3.88/4.0 (Omega Nu Lambda Honors) **December 2025**
- **Pennsylvania State University** University Park, Pennsylvania
  M.S. in Software Engineering, GPA: 3.39/4.0 (Magna Cum Laude Honors) **May 2020**
- **Rochester Community and Technical College** Rochester, MN
  Diploma in Health Informatics, GPA: 4.00/4.0 (High Honors) **August 2015**
- **University of Baghdad** Iraq, Baghdad
  B.S. in Computer Science, GPA: 4.00/4.0 (High Honors) **June 2007**

SELECTED PUBLICATIONS

- Text and Audio Classification Enabled Diagnosis for Treatment Applications by Natural Language Processing (NLP) and Deep Learning (DL), 2025 (DOI 10.5281/zenodo.19394304)
- Perceive, Plan, Act, Self-Correct: An Architectural Framework for Goal-Directed Agentic AI Systems (DOI: 10.31224/6738)

FORTHCOMING PUBLICATIONS

- A Unified Taxonomy of 19 AI System Types: Architecture, Capability, and Deployment Analysis for the Modern AI Landscape
- Causal Inference at Scale: How Analytical AI Transforms Pattern Mining Into Actionable Business Intelligence
- Autonomy Without Agency: A Formal Distinction Between Autonomous and Agentic AI Systems in Safety-Critical Domains
- From Bayes' Theorem to Clinical Deployment: Probabilistic AI for High-Stakes Decision-Making Under Uncertainty
- Beyond Hallucination: How Neuro-Symbolic Fusion Enables Verifiable Reasoning in Large Language Models
- Where Humans Meet Machines: A Survey of Conversational AI Pipelines From NLU to Real-Time Response Generation
- Population-Based Search for Neural Architecture Design: Evolutionary AI in AutoML, Drug Discovery, and Robot Morphology
- SHAP, LIME, and Beyond: Post-Hoc Explanation Methods for Black-Box Models in Healthcare, Finance, and Criminal Justice
- Trustless Collaboration: Harnessing Cryptographic Computation and Federated AI for Secure Multi-Institutional Synthesis
- Architectural Evolution: From Statistical Autoregression to Grounded and Verifiable Reasoning in Generative Systems
- Seeing, Reading, and Hearing Simultaneously: How Multimodal Perception AI Achieves Understanding Beyond Any Single Modality
- Exact Solvers, Heuristics, and Hybrid AI: A Survey of Optimization Architectures for Supply Chain, Routing, and Resource Allocation
- AI in the Physical World: Foundation Models for Robotics, Sim-to-Real Transfer, and Real-Time Embodied Intelligence
- Supervised Learning in Production: How Predictive AI Drives Billions of Daily Decisions in Medical Diagnosis
- Stateless by Design: Why Reactive AI Systems Without Memory, Learning, or Planning Remain Essential in Safety-Critical Domains
- From Collaborative Filtering to Neural Retrieval: The Evolution of Recommendation AI in Search and Social Media Platforms
- Trial, Error, and Reward: How Reinforcement Learning Discovers Strategies No Human Taught From Game Play to Protein Folding
- AlphaFold, GNoME, and GraphCast: How Scientific AI Is Transforming Chemistry, Materials Science, and Weather Forecasting
- Explicit Knowledge, Traceable Reasoning: A Survey of Symbolic and Rule-Based AI From GOFAI to Modern Hybrid Systems

AWARDS & HONORS

- ACM SIGSOFT Distinguished Paper for research advancing software engineering practices and innovation (2023)
- Mayo Clinic Mae Berry Service Excellence for exceptional contributions to organizational impact and team excellence (2016)

CERTIFICATIONS AND LICENSES

Oracle Fusion AI Agent Studio Certified Foundations Associate - Rel 1 (2026) • Google Cloud Professional Data Engineer (2026) • Scaled Agile Framework (SAFe) (2025) • AI For Medical Diagnosis (2024) • AWS Certified Solutions Architect - Associate (2022) • Microsoft Azure Fundamentals (AZ900) (2023) • Certified Scrum Master (CSM) (2020)

LANGUAGES

• Arabic (Fluent) • English (Native) • Spanish (Conversational)