Bashir Mohammed, PhD

+1-510-356-7983 | bashirmx6@gmail.com | Personal Website | LinkedIn |Twitter| GitHub | Google Scholar
San Francisco, CA 94105, USA

SUMMARY

I'm a hands-on Senior Lead GenAI Solutions Architect on the Frontier AI team at AWS, where I partner with the world's most innovative founders and frontier AI startups to turn breakthrough ideas into production-grade, cost-efficient GenAI and agentic applications on AWS, from prototype to production. I specialize in architecting reliable, high-performance AI infrastructure with a deep focus on model behavior, evaluation rigor, and responsible deployment.
Previously, I was a Principal Staff AI Architect at Intel in the Office of the CTO, where I led anti-hallucination and guardrail initiatives for LLMs, vision-language models, and multi-agent systems, designing evaluation frameworks and prompt-based safety control layers that measurably improved reliability across enterprise deployments.
I hold a Ph.D. in Computer Science and bring a strong research foundation from Lawrence Berkeley National Laboratory, where my work focused on intelligent networks, quantum communication systems, and high-performance computing and distributed systems.
At the intersection of deep science and real-world deployment, I build AI solutions that are safe, responsible, and ready to scale globally.

EXPERIENCE

Amazon Web Services (AWS) *March 2025 – Present*
Senior Lead GenAI Solutions Architect — Frontier AI Team *San Francisco, CA*

- **Frontier AI Partnership:** Architect and scale production-grade GenAI and agentic systems for Silicon Valley's leading frontier AI startups, partnering directly with founding engineers and CTOs from prototype through enterprise deployment on Amazon Bedrock, SageMaker, and Trainium/Inferentia.
- **Joint Reference Architecture (Oumi + Bedrock):** Co-authored the AWS Machine Learning Blog "Accelerate custom LLM deployment: Fine-tune with Oumi and deploy to Amazon Bedrock" (March 2026), establishing a codified joint solution with the Oumi team for the open-source training-to-production gap. Reference architecture covers FSDP/DeepSpeed/DDP distributed training on EC2, S3 artifact versioning, and managed inference via Bedrock Custom Model Import.
- **Developer Enablement & Partner Events:** Co-led the Stanford University StartX & Friends Agentic AI Deep Dive workshop in Palo Alto for the Stanford-affiliated founder community — delivering hands-on sessions on Amazon Bedrock, AgentCore (Runtime, Identity, Gateway, Memory), Kiro, Amazon Q Developer CLI, and multi-agent collaboration architectures.
- **Agentic AI Systems Architecture:** Design and deploy multi-agent workflows for frontier startups, addressing long-horizon agent state management, tool-use orchestration, eval rigor, and the demo-to-production gap. Hands-on expertise spans LangGraph, CrewAI, Strands, MCP, A2A, and Bedrock AgentCore.
- **LLM Customization & Fine-Tuning:** Lead end-to-end fine-tuning engagements covering full fine-tuning, LoRA/QLoRA, RLHF and post-training optimization. Build reproducible recipe-driven training pipelines with integrated evaluation and synthetic data generation workflows.
- **Evaluation & Benchmarking:** Architect production-grade LLM evaluation and observability frameworks for partner deployments, including hallucination detection, instruction-following metrics, RAG eval pipelines, and agentic trajectory analysis.
- **Mentorship & Community:** Mentor early-career engineers across AWS and the partner ecosystem; contribute to internal technical enablement programs and external developer communities.

Intel Corporation *February 2023 – February 2025*
Principal Staff AI Solutions Architect, Office of the Chief Technology Officer (OCTO) *Santa Clara, CA*

- **AI Safety & Guardrails Leadership:** Led empirical research on hallucination, over-confidence, and instruction-following failures in LLMs and multimodal models. Designed evaluation frameworks to systematically measure and characterize unsafe or misleading model behaviors in real-world deployments.
- **Production Safety Control Layers:** Designed and deployed prompt-engine and guardrail systems including dynamic instruction steering, uncertainty-aware responses, and retrieval-augmented workflows — measurably reducing hallucination and improving reliability across enterprise production deployments.
- **SEAL (SmartEdge Agent):** Lead architect and inventor of an LLM-Powered Conversational Control plane for Advanced Edge Manageability — a novel solution revolutionizing edge management through conversational command and control.
- **Visual-RAG Theft Detection (Major Retail Client):** Led delivery of a real-time multi-modal video understanding solution combining LVMs, LLMs, and multi-agent frameworks on low-cost Intel hardware, addressing extensive shoplifting challenges while ensuring strict safety guardrail compliance and minimizing hallucinations.
- **GenAI Stack Expertise:** Specialized in LLMs, LVMs, and SLMs, building Retrieval-Augmented Generation pipelines with LangChain, LlamaIndex, and Haystack. Hands-on with NVIDIA NeMo, NIM, RAPIDS, and CUDA.

Lawrence Berkeley National Laboratory *June 2022 – January 2023*
Computational Research Engineer / Scientist *Berkeley, CA*

- **CRD-NERSC Supporting Workflows:** Advanced intelligent scientific workflow data management at the National Energy Research Scientific Computing Center, with emphasis on real-time stream processing and data provenance for optimized scientific computing.
- **QUANT-NET (Quantum Application Network Testbed):** Developed a proof-of-concept quantum network linking Berkeley Lab and UC Berkeley featuring entanglement swapping over optical fiber. Collaborated with leading experts from Berkeley Lab, UC Berkeley, and Caltech to demonstrate entanglement between small-scale quantum computers.[Link]
- **Cyber-Physical Systems Security:** Built self-driving synthetic biology labs using ML processes and Bayesian ensemble modeling through the Automated Recommendation Tool (ART) to secure and optimize cyber-physical processes.[Link]

Lawrence Berkeley National Laboratory *April 2019 – May 2022*
Postdoctoral Research Fellow *Berkeley, CA*

- **"Large-scale Deep Learning for Intelligent Networks" (DOE-funded):** Led development of AI/ML algorithms to optimize control of distributed network resources, enhance high-speed data transfers, and minimize network downtime for exascale scientific workflows. *Won Best Paper Award at the Machine Learning for Networking Conference.[Link]*

AI Collaborator, Inc. *January 2021 – May 2022*
Head of AI / CTO *Los Angeles, CA*

- Drove AI strategy and end-to-end product lifecycle for AI-driven solutions; managed cross-functional team of engineers, data scientists, and product managers.

Nabafat.AI *January 2013 – March 2019*
Head of AI and Lead Technical Program Manager *Sacramento, CA*

- Led AI/ML department developing supervised and unsupervised models; achieved 30% improvement in predictive accuracy and 25% gain in operational efficiency.
- Established robust data infrastructure and pipeline architecture; reduced model training times by 40% through automation of ingestion, cleansing, and feature engineering.

AWS CERTIFICATIONS

- **AWS Certified Machine Learning Engineer – Associate** (MLA-C01)
- **AWS Certified Solutions Architect – Associate** (SAA-C03)
- **AWS AI Foundational** (L100)

FEATURED PUBLICATIONS & TALKS

- **Mohammed, B.**, Krishnamoorthy, B., Fina, G., Stewart, D., Persons, M. *"Accelerate custom LLM deployment: Fine-tune with Oumi and deploy to Amazon Bedrock,"* AWS Machine Learning Blog, March 2026. [Link]
- AWS Speaker — *"Deep Dive into Agentic AI" workshop*, StartX & Friends, Palo Alto, CA. Hands-on session on Amazon Bedrock, AgentCore, Kiro, Amazon Q Developer CLI, and multi-agent systems. [Link]
- Presented research to U.S. Congress and participated in White House Office of Science and Technology Policy (OSTP) sessions on AI and Quantum policy.[Link]

EDUCATION

University of Bradford, UK *November 2014 – July 2019*
Ph.D. in Computer Science — Advisors: Prof. Hassan Ugail, Prof. Irfan Awan *United Kingdom*

- **Thesis:** *"A Predictive Framework for Efficient Management of Fault Tolerance in Cloud Data Centres and High-Performance Computing Systems."*

University of Sheffield, UK *September 2010 – February 2012*
M.Sc. in Control Systems — Advisors: Prof. Peter Fleming, Dr. Andy Mills *United Kingdom*

- **Thesis:** *"Integrated Combustor Temperature Measurement and Health-Aware Control Framework for Gas Turbine Engines."*

Federal University of Technology, Minna
B.Eng. in Electrical and Computer Engineering *Nigeria*

HONORS & AWARDS

- **British Council Study UK Alumni Awards 2026 — Finalist, Science and Sustainability Award.** [Link]
- Board of Advisors, **Institute for Mathematical and Statistical Innovation (IMSI).** [Link]
- **SIAM Science Policy Fellowship Award (2023):** Member, Committee on Science Policy (CSP). Contributed to AI and Quantum policy memos representing the SIAM community to policymakers in Washington, D.C. [Link]
- **Exceptional Talent in Digital Technology — UK Government Endorsement (2020):** Endorsed by Tech Nation as a World-Leading Exceptional Talent in Digital Technology.
- **Black & Brilliant / Codecademy AI Accelerator Coaching Award (2021):** Selected as Data Science and AI Coach for the Black & Brilliant AI Accelerator.
- **Berkeley Lab Research SLAM (2019):** Second place winner; selected as finalist among 42 scientists.
- **IYPT Elemental SLAM on Capitol Hill (2019):** Selected to represent Berkeley Lab; presented research to U.S. legislators including Senators Lisa Murkowski, Bruce Westerman, and Randy Weber. [Link] [Link]
- Best Paper Award, Machine Learning for Networking Conference. [Link]

TECHNICAL SKILLS

- **GenAI & Agentic AI:** LLMs, LVMs, SLMs, RAG, Fine-Tuning (LoRA, QLoRA, full FT), RLHF, DPO, Prompt Engineering, Eval Frameworks (MMLU, LLM-as-judge, custom). Frameworks: LangChain, LlamaIndex, Haystack, CrewAI, LangGraph, Strands, MCP, A2A.
- **AWS GenAI Stack:** Amazon Bedrock (Claude, Llama, Custom Model Import), Bedrock AgentCore (Runtime, Identity, Gateway, Memory), Amazon Q Developer CLI, Kiro, SageMaker, Trainium/Inferentia, Comprehend.
- **Cloud & Infrastructure:** AWS (EC2, S3, IAM, VPC, KMS, EKS, Lambda, CloudWatch), Kubernetes, Slurm, Lustre, BeeGFS.
- **HPC & Networking:** InfiniBand, NVLink, RDMA, NCCL, NVSwitch, Distributed Training (FSDP, DeepSpeed, DDP).
- **Performance Optimization:** NVIDIA Nsight Systems, CUDA kernels, Distributed PyTorch, Model Parallelism, Pipeline Optimization.
- **Programming Languages:** Python, C, C#, Java, MATLAB, SQL.

- **Databases & Vector Stores:** SQL, MongoDB, InfluxDB, PostgreSQL, ChromaDB, Intel VDMS, Weaviate, pgvector.
- **ML/DL Libraries:** PyTorch, TensorFlow, Keras, Theano, Caffe, scikit-learn, NumPy, SciPy, Pandas, NetworkX, CUDA.
- **NVIDIA Stack:** NeMo, NIM, RAPIDS, CUDA, Triton.
- **General Tools:** Linux, Git, Shell Scripting, Docker.
- **Mathematics:** Engineering and Industrial Mathematics, Control Systems, Differential Equations, Probabilistic and Statistical Modeling, Time-Series Forecasting.
- **Leadership:** Technical roadmaps, cross-functional alignment, AI governance, partner ecosystem enablement, cross-team mentoring.