Bashir Mohammed, PhD

+1-510-356-7983 | bashirmx6@gmail.com | personal-website

bashir-mohammed | Github | Google-Scholar | BashirMohd100

San Francisco, CA 94105 USA

Summary

I'm a hands-on Senior Lead GenAI Solutions Architect at AWS Startups, where I work with some of the world's most innovative founders to turn bold ideas into scalable, cost-efficient GenAI and agentic applications, from prototype to production. I specialize in architecting reliable, high-performance AI infrastructure, with a strong focus on model behavior, evaluation rigor, and responsible deployment to help startups unlock speed, efficiency, and real-world impact on AWS.

Previously, I was a Senior Staff AI Architect at Intel, where I led anti-hallucination and guardrail initiatives for LLMs, vision-language models, and multi-agent systems, designing evaluation frameworks and prompt-based safety controls that measurably improved reliability across enterprise and industrial deployments. I hold a Ph.D. in Computer Science and bring a strong research foundation from my time at Lawrence Berkeley National Laboratory, where I conducted applied research in intelligent networks, quantum communication systems, and high-performance distributed systems.

At the intersection of deep science and real-world deployment, I focus on understanding how AI systems behave in practice, building AI solutions that are safe, responsible, and ready to scale globally.

Experience

- **Amazon Web Services(AWS)** — *March 2025 - Present*
  Senior Lead GenAI Solutions Architect at AWS Startups, — San Francisco, CA
  - Senior GenAI Solutions Architect at AWS Startups where I partner with elite founders to turn breakthrough ideas into production-grade GenAI and agentic systems fast, scalable, and cost-smart on AWS.
- **Intel Corporation** — *Feb 2023 - Feb 2025*
  Principal Staff AI Solutions Architect, Office of the Chief Technology officer (OCTO) — Santa Clara, CA
  - Currently specializing in Natural Language and Vision models, with a strong focus on Large Language Models (LLMs), Large Vision Models (LVMs), and Small Language Models (SLMs). Building Retrieval-Augmented Generation (RAG) pipelines utilizing frameworks such as LangChain and LlamaIndex.
  - Led empirical research on hallucination, over-confidence, and instruction-following failures in large language and multimodal models, designing evaluation frameworks to systematically measure and characterize unsafe or misleading model behaviors in real-world scenarios.
  - Designed and deployed prompt-engine and guardrail systems as safety control layers, including dynamic instruction steering, uncertainty-aware responses, and retrieval-augmented workflows, resulting in measurable reductions in hallucination and improved reliability across production deployments.
  - Designed and executed proof-of-concept solutions for theft detection and video understanding using LVMs, LLMs, and multi-agent frameworks, tailored for retail industry applications.
  - Lead architect and Inventor of SEAL: A SmartEdge Agent and LLM-Powered Conversational Control for Advanced Edge Manageability - a novel solution designed to revolutionize edge management through conversational command and control.
  - Led the Visual-RAG Theft Detection Video Summarization Project to Address Extensive Shoplifting Challenges for a Major Retail Clients, delivering a real-time, multi-modal solution on low-cost Intel hardware while ensuring strict compliance with safety guardrails and leading efforts to minimize AI hallucinations for accurate and reliable performance
  - Experienced with NVIDIA's software libraries, platforms, and frameworks, including Neural Modules (NeMo), NVIDIA Inference Microservices (NIM), RAPIDS, and CUDA, among others.

Lawrence Berkeley National Laboratory — *June 2022 - Jan 2023*
Computational Research Engineer/Scientist — Berkeley, CA
Key projects:

- CRD-NERSC Supporting Workflows: Focused on advancing intelligent scientific workflow data management at the National Energy Research Scientific Computing Center (NERSC), with an emphasis on real-time stream processing and data provenance, contributing to optimized and efficient scientific computing processes.
- QUANT-NET (Quantum Application Network Testbed for Novel Entanglement Technology): Developed a proof-of-concept quantum network linking Berkeley Lab and UC Berkeley, featuring entanglement swapping over optical fiber and managed by a quantum network protocol stack. Collaborated with leading experts from Berkeley Lab, UC Berkeley, and Caltech to demonstrate entanglement between small-scale quantum computers.
- Securing Automated, Adaptive Learning-Driven Cyber-Physical Systems: 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 system processes.

Lawrence Berkeley National Laboratory *April 2019 - May 2022*
Postdoctoral Research fellow, Berkeley, CA

- Worked on the "Large-scale Deep Learning for Intelligent Networks" project at Berkeley Lab, funded by the US Department of Energy, where I led and developed AI and ML algorithms to optimize the control of distributed network resources, enhance high-speed data transfers, and minimize network downtime for exascale scientific workflows. Achieved the Best Paper Award at the Machine Learning for Networking Conference.

AI Collaborator, Inc *Jan 2021 - May 2022*
Head of AI and CTO, Los Angeles, CA

- Spearheaded the development and execution of the AI strategy, driving innovation across products and services, and ensuring alignment with business objectives and market trends.
- Oversaw the end-to-end product lifecycle, from ideation to launch, for AI-driven solutions, ensuring timely delivery, market fit, and customer satisfaction.
- Managed and mentored a cross-functional team of engineers, data scientists, and product managers, fostering a collaborative environment that maximized productivity and innovation.

Nabafat.AI *Jan 2013 - Mar 2019*
Head of AI and Lead Technical Program Manager, Sacramento, CA

- Led the AI/ML department in developing cutting-edge machine learning algorithms, including supervised and unsupervised models, resulting in a 30% improvement in predictive accuracy for key business metrics.
- Led technical program management for AI initiatives, including resource allocation, risk assessment, and stakeholder communication, ensuring smooth execution of large-scale AI/ML deployments.
- Spearheaded the end-to-end design, development, and deployment of AI-driven solutions across multiple domains, including natural language processing (NLP), computer vision, and predictive analytics, enhancing operational efficiencies by 25%.
- Managed and delivered high-impact AI/ML projects, coordinating cross-functional teams of data scientists, engineers, and stakeholders to achieve project goals on time and within budget.
- Established a robust data infrastructure and pipeline architecture, automating data ingestion, cleansing, and feature engineering processes, reducing model training times by 40%.
- Provided technical leadership and mentorship to a team AI/ML engineers and data scientists, fostering a collaborative environment that accelerated innovation and knowledge sharing.

EDUCATION

- **University of Bradford, UK** *Nov 2014 - July 2019*
  PhD in Computer Science, Advisors: Prof. Hassan Ugail and Prof. Irfan Awan UK
  - Thesis: "A Predictive Framework for Efficient Management of Fault Tolerance in Cloud Data Centres and High-Performance Computing Systems"
- **University of Sheffield, UK** *Sep 2010 - Feb 2012*
  MSc in Control Systems Advisors: Prof. Peter Fleming and Dr. Andy Mills (Rolls Royce, UTC, Sheffield, UK) UK
  - Thesis: "Integrated Combustor Temperature Measurement and Health-Aware Control Framework for Gas Turbine Engines: A Holistic Approach to Fault Tolerance, Prognostic and Diagnostic Algorithms"
- **Federal University of Technology,Minna** *Nov 2006*
  Electrical and Computer Engineering Nigeria

SKILLS

- **Programming Languages:** Python, C, CSharp, Java, MATLAB, SQL.
- **Databases:** SQL, MongoDB, InfluxDB, Postgres, ChromaDB, Intel VDMS, Weaviate, pgvector
- **Packages and Libraries:** NumPy, SciPy, Pandas, TensorFlow, Keras, Theano, Caffe, PyTorch, NetworkX, PyTorch, SciKit-Learn, CUDA.
- **General Tools and Platforms:** Linux, Git, Shell Scripting.
- **Mathematics:** Strong foundation in Engineering Mathematics and Industrial Mathematics, with expertise in Control Systems, Differential Equations, Probabilistic and Statistical Modeling
- **Artificial Intelligence and Machine Learning:** Expertise in Gen-AI and Deep Learning - LLMs, LVMs, SLMs, RAG, Fine-Tunning, Prompt-Tunning, Prompt Engineering, Langchain, LlamaIndex,Haystack, Multi-Agent frameworks, CrewAI, LangGraph.
- **HPC & Networking:** InfiniBand, NVLink, RDMA, NCCL, NVSwitch, GPUDirect, Slurm, Kubernetes, Lustre, BeeGFS.
- **Performance Optimization:** Nsight Systems, CUDA kernels, Distributed PyTorch, Model Parallelism, Pipeline Optimization
- **Predictive Modeling and Forecasting:** Time Series Forecasting and Statistical Modeling.
- **Leadership Strategy:** Technical roadmaps, cross-functional alignment, AI governance, partner ecosystem enablement.

HONORS AND AWARDS

- **SIAM Science Policy Fellowship Award 2023** *Jan 2023*
  Society of Industrial and Applied Mathematics (SIAM) - Part time [LINK]
  - As a Science and Technology Policy Fellow for SIAM, I serve on the Committee on Science Policy (CSP), actively representing the SIAM community to policymakers in Washington, D.C. In this role, I contribute to the development of AI and Quantum policy memos and white papers. Click the following links for more details: [LINK] [LINK]

- **Black and Brilliant and Codecademy AI Accelerator Coaching Award** *Feb 2021*
  Codecademy [LINK]
  - Selected as a Data Science and AI Coach for the Black and Brilliant AI Accelerator Course with Codecademy.

- **Exceptional Talent Digital Technology UK Government Endorsement Award** *Feb 2020*
  Tech Nation - Global Exceptional Talent program [LINK]
  - Endorsed by the UK government as a World-Leading Expertional Talent in Digital Technology.

- **Berkeley Lab Research SLAM Award Winner** *Sep 2019*
  Lawrence Berkeley National Lab [LINK]
  - I am honored to have earned second place in the prestigious Berkeley Lab Research SLAM contest.

- **Berkeley Lab Research SLAM Finalist** *Sep 2019*
  Lawrence Berkeley National Lab [LINK]
  - Selected as a finalist in the Berkeley Lab Research SLAM competition with 42 Scientist.

- **The IYPT Elemental Slam Award on Capitol Hill** *Oct 2019*
  US Department Of Energy/ UC Berkeley [LINK]
  [LINK]
  - Winner of the Berkeley Lab Research SLAM and selected to represent Berkeley Lab at the IYPT Elemental Slam on Capitol Hill, where I had the privilege of presenting my research to legislators and a Capitol Hill audience. Notable attendees included Senators Lisa Murkowski (Alaska), Bruce Westerman (Arkansas), and Randy Weber (Texas).