John Kitaoka

jckitaok@gmail.com — 507.884.1155 — in/johnkitaoka

AI architect building production systems at scale: fine-tuned LLMs and VLMs, custom evaluations, and agentic architectures. Depth in custom evaluations, AI/ML systems, and public sector vertical software.

Experience

Applied AI Architect II Amazon Web Services (New York, NY) *2024 - Present*

- Design, develop, evaluate, and optimize AI/ML workloads on AWS Cloud for public sector customers, leading 6 end-to-end AI system deployments, from initial design through production, with technical patterns reused and adopted org-wide
- Built end-to-end RAG systems: hybrid retrieval (semantic + keyword) with reranking, agentic query decomposition for multi-hop reasoning and synthesis, and guardrails for PII/policy compliance
- Designed agentic workflows with multi-agent orchestration, custom tooling, data source aggregation, and report generation; for document processing (90% review time reduction) and life science research (60% query writing time reduction)
- Fine-tuned Qwen3-4B with LoRA for long-context QA; built custom industry-specific evaluation benchmark, extended context via RoPE scaling, improved answer accuracy 70% on benchmark at 64K+ context, 90% cheaper than Claude Haiku

Solutions Architect II Amazon Web Services (New York, NY) *2022 - 2024*

- Architected and deployed 25+ production ML workloads across computer vision, NLP, and forecasting; designed for scale, cost efficiency, and operational reliability
- Built first public RAG workshop across major cloud providers; presented at Re:Invent 2023/24/25 to 1,000+ practitioners, with reference architectures adopted in production by multiple organizations
- Led technical evaluations against hyperscaler and frontier lab alternatives; designed proof-of-concept implementations that won 8 competitive assessments including 3 platform migrations
- Architected 100TB data platform migrations, replacing self-managed single-node databases with managed stack: Aurora Postgres, zero-ETL replication to Redshift, BI layer—deployed across 4 organizations, 70%+ reduction in data engineering work
- Created enablement curriculum on AI responsibility for 150 solutions architects; content scaled org-wide to 750+ practitioners

Projects

Multimodal AI Form Filling — GitHub: aws-samples/sample-bedrock-form-filling *December 2025*

- Built a serverless, event-driven pipeline for multimodal-to-text extraction and eForm automation using AWS Step Functions and Amazon Nova LLMs, deployed via SAM/CloudFormation, with CloudFront and Cognito

Small Data, Big Insights — Amazon internal ML conference (AMLC 25) *September 2025*

- Co-first author on internal paper: alignment framework for LLM-based thematic analysis enabling structured handoff between human-labeled samples and LLM-generated analysis for qualitative research at scale

Blind Spots in Vision-Language Models — Amazon internal CV conference (ACVC 24) *August 2024*

- First author on internal paper: introduced spatial awareness evaluation methodology for LLaVA-style VLMs; training modifications improved peripheral region accuracy 30%; framework integrated into production model evaluation pipeline

Fine-tuning and Deploying LLMs in Isolated Environments — AWS blog series *June 2024*

- Led development on a technical blog series on fine-tuning LLMs with QLoRA and deploying an inference server with HuggingFace libraries: PEFT, transformers, and datasets for customers in national security, defense, and federal software

Generative AI in the Public Sector — AWS eBook *January 2024*

- Co-authored official AWS advisory guide on generative AI strategy for public sector; 15K+ downloads, adopted by 3,000+ organizations for production AI planning, writing chapters on foundation model selection and public sector reference architecture

MistralLite — Open-source long-context LLM (HuggingFace: amazon/mistrallite) *October 2023*

- Implemented AWQ quantization achieving 4x memory efficiency; developed long-context evaluation benchmarks for retrieval and reasoning tasks; 450K+ HuggingFace downloads, 3 production deployments

Education

University of Wisconsin-Madison Madison, WI

BSc, Mathematics and Computer Science; Minor in Business *GPA: 3.92/4.0*

- Top 5% of class, Distinction in both majors, D.E. Shaw Nexus Fellowship, Men's Water Polo, Mayo Foundation Scholarship

---

Languages and Tools: Python, Java, C, C++, SQL

Frameworks: pandas, PyTorch, tensorflow, LangChain, CUDA, Neuron, spark, transformers, accelerate

Certifications: AWS Solutions Architect Professional, AWS Machine Learning Specialty, AWS SysOps Administrator