

Keith Steward, Ph.D.

Lead AI Engineer | Foundational LLM & Retrieval Systems

Boston, MA | 508-318-8035 | keith.steward@gmail.com

[LinkedIn](#) | IntelligentDataWorks.ai | [IDW LinkedIn](#)

PROFILE

Hands-on AI engineer and architect who builds reliable foundational services for LLM inference, retrieval, agents, evaluation, governance, and production operations. Combines Python, AWS, APIs, distributed systems, vector databases, guardrails, observability, and performance-cost optimization with deep applied AI experience. Built the AssistX core and first 80% of its initial application, enabling 70+ automations across more than 20 workflows with 99.9% availability. Brings the technical judgment to turn emerging research into scalable capabilities while improving latency, reliability, safety, and developer leverage.

CORE TECHNOLOGIES

Python • LLM Inference • Similarity Search • Vector Databases • RAG • AI Agents • Guardrails • AI Evaluation • Experimentation • Governance • Observability • AWS • Amazon Bedrock • SageMaker • Distributed Systems

SELECTED CAREER HIGHLIGHTS

- Architected and built the AssistX core and first 80% of its initial application hands-on using Python, FastAPI, LLMs, RAG, agents, vectors, APIs, and AWS.
- Delivered 70+ automations across more than 20 workflows with 99.9% availability and selected work accelerated up to 50x.
- Established evaluation, guardrails, human review, model routing, caching, observability, security, cost controls, and release practices.
- Advised 200+ AWS enterprise customers on AI, ML, data, and cloud architecture and delivery.
- Led Responsible AI, privacy, IAM, auditability, bias, drift, and model-governance work in financial services.
- Built and operated production AI services and data-intensive ML systems spanning ingestion, transformation, feature preparation, LLM invocation, retrieval, model evaluation, APIs, deployment, monitoring, and support. Delivered optimized data pipelines for forecasting and optimization services, an enterprise data lake across 11 ERP environments, and agentic AI and RAG pipelines that crawled and embedded labor-law content from all 50 states. Used Python, SQL, Spark/PySpark distributed computing, TensorFlow, SageMaker, AWS Bedrock, PostgreSQL/PGVector, FastAPI, event-driven AWS services, containers, CI/CD, and automated controls for reliable enterprise use at scale.

PROFESSIONAL EXPERIENCE

Technical Co-Founder | Intelligent DataWorks | Boston, MA | Aug 2023 to Oct 2026

Built and operated a production AI platform while personally owning architecture, implementation, evaluation, delivery, governance, security, and cloud operations.

- Architected and built the AssistX core platform and first 80% of its initial application hands-on using Python, FastAPI, LLMs, RAG, agents, PostgreSQL/PGVector, APIs, Docker, AWS, and CDK.
- Recruited and technically led 11 engineers while owning a \$900,000 budget, product priorities, architecture, releases, security, and operations.
- Delivered 70+ automations across more than 20 workflows, accelerating selected work by up to 50x and sustaining 99.9% availability.
- Established evaluation, guardrails, human oversight, observability, model routing, cost controls, incident response, and release practices.
- Completed planned transition from operating CEO responsibilities to pursue a long-term senior AI/engineering leadership role.

Senior Director of Engineering | SuccessKPI | Remote | Jun 2022 to Aug 2023

Built and led engineering for an AI-powered workforce-management component of a cloud-native contact-center SaaS platform.

- Recruited and led eight engineers and one manager across software, data, platform, security, BI, and quality engineering.
- Owned Agile execution, architecture, operational priorities, performance reviews, bonus allocation, team capacity, and a \$600,000 budget.
- Delivered forecasting, scheduling, optimization, analytics, APIs, and integrations for enterprise and government customers.

Senior Director, Data Analytics & Operations | TriMark USA | Mansfield, MA | Aug 2021 to May 2022

Established the enterprise data strategy, engineering team, architecture, roadmap, and operating capability across a fragmented multi-division environment.

- Aligned executives around more than 50 analytics use cases and secured funding for an AWS data-lake strategy and delivery roadmap.
- Upskilled an analytics team with no prior AWS experience and expanded it to seven cloud data engineers.
- Delivered the first enterprise data lake across 11 divisions in seven months, reducing cross-ERP reporting from roughly two months to one day.

Technical Lead, Cloud Data Privacy & AI Governance | MassMutual | Boston, MA | Nov 2020 to Aug 2021

Led architecture and hands-on implementation across five enterprise Data Privacy and AI Governance initiatives.

- Led architecture and engineering across five enterprise Data Privacy and AI Governance initiatives spanning privacy rights, enterprise metadata, policy automation, and governance APIs.
- Advised the AI Governance and Ethics organization on model transparency, bias, drift, ownership, risk classification, monitoring, human oversight, and auditability.
- Established AWS security, IAM, API, event-driven, and CI/CD patterns while coordinating Product, Enterprise Architecture, SRE, governance, and delivery teams.

VP of AI/ML | NorthBay Solutions | Boston, MA | Jan 2019 to Nov 2020 | (Role ended in company-wide COVID layoff).

Recruited out of AWS by the NorthBay Solutions CEO to establish a new AI/ML practice and reusable production data-and-ML capability.

- Built an AWS Data Lake and ML accelerator hands-on with Python, Django, React, APIs, and AWS, then directed a 24-person remote team toward delivery.
- Delivered an AWS Neptune knowledge graph integrating complex hierarchical enterprise data and expanded reusable cloud and AI architectures for customers.
- Advised enterprise customers on AI strategy, data platforms, knowledge graphs, solution architecture, and production implementation.

Senior Specialist Solutions Architect, AI/ML & Big Data | Amazon Web Services | Boston, MA | May 2015 to Dec 2018

Advised enterprises and AWS teams on production machine learning, distributed analytics, data platforms, and cloud adoption.

- Advised 200+ enterprises on AI/ML strategy, production architecture, SageMaker, data lakes, distributed systems, security, and scalable adoption.
- Created and led global Big Data and AI/ML communities, developing 180+ specialists through standards, mentoring, workshops, and reusable assets.
- Built a SageMaker and TensorFlow forecasting hackathon for Morgan Stanley data scientists and created 170+ technical enablement assets.
- Delivered 50+ technical presentations and workshops and earned AWS ML Elite Gold and Platinum recognition.

EARLIER RELEVANT EXPERIENCE

- Built an AWS data-and-ML accelerator and delivered hands-on AI/ML consulting solutions.
- Developed commercial AI, bioinformatics, and scientific software, including patented knowledge-discovery systems.

EDUCATION & RECOGNITION

Ph.D., Molecular Biology / Microbiology & Immunology, University of Western Ontario. AWS ML Elite Gold and Platinum; AWS re:Invent Rockstar; named inventor on multiple patents.

TECHNOLOGIES

Python • LLM inference • similarity search • vector databases • RAG • AI agents • guardrails • AI evaluation • experimentation • AI governance • Responsible AI • observability • model routing • caching • latency optimization • throughput optimization • cost optimization • AWS • Amazon Bedrock • SageMaker • FastAPI • REST APIs • distributed systems • microservices • PostgreSQL • PGVector • LangChain • LangGraph • MCP • prompt engineering • embeddings • semantic retrieval • MLOps • CI/CD • Docker • Kubernetes • AWS CDK • security • IAM • auditability • human-in-the-loop • machine learning • TensorFlow