

KEVIN KAKOLLA

Senior Agentic AI Engineer · AWS / Cloud AI · LLM Orchestration · Production ML Systems

Seattle, WA · kkakolla@andrew.cmu.edu · github.com/torontodeveloper · huggingface.co/torontodeveloper · youtube.com/@KevinKakolla · Open to relocation

PROFESSIONAL SUMMARY

Senior AI Engineer with 15+ years of software development experience and 3+ years designing, deploying, and optimizing production ML/GenAI systems across financial services, energy, healthcare, and media. **Forward-deployed delivery throughout** — embedded directly with customers and stakeholders (HP/Riyadh, Navigating Cancer, Dish Network, and regulated enterprise clients) to take systems from requirements to shipped, integrated production. Hands-on builder of agentic AI systems, RAG pipelines, tool-calling workflows, LLM evaluation/guardrails, and model fine-tuning (QLoRA/DPO). Recently fine-tuned and published a DeBERTa-v3 web-agent grounding model on HuggingFace (93% recall@10 on unseen domains, Mind2Web benchmark). Cloud experience across AWS, Azure, and Vertex AI, with Terraform/IaC, Docker/Kubernetes, CI/CD, Databricks, and MLflow. Currently completing a Graduate Certificate in GenAI & LLMs at Carnegie Mellon University (2026), researching multimodal web agents.

TECHNICAL SKILLS

- **Languages / Backend:** Python, Java, Go, C/C++, FastAPI, Django, Node.js, Spring Boot, PostgreSQL, MongoDB, Oracle
- **Agentic AI / LLM:** Tool-calling agents, function calling, secure tool execution, LangChain, LangGraph, RAG, Pinecone, Playwright browser agents, prompt engineering, guardrails
- **Models / Fine-tuning:** GPT-4o/GPT-5, Claude, Gemini, Llama 3.1/4, Qwen, Flan-T5, DeBERTa, HuggingFace, QLoRA, DPO, vLLM, LLM-as-a-Judge
- **ML / Statistics:** PyTorch, scikit-learn, XGBoost, TensorFlow, Keras, CNNs, RNNs, SHAP, regression, classification, anomaly detection, model evaluation
- **Cloud / MLOps:** AWS SageMaker, AWS deployments, Azure ML, Azure OpenAI, Databricks, Vertex AI, MLflow, Airflow, Terraform, CI/CD
- **Containers / DevOps:** Docker, Kubernetes/AKS, Azure DevOps, GitHub Actions, Jenkins, Locust, monitoring/evaluation pipelines
- **Data / Analytics:** Pandas, NumPy, PySpark, Dask, BigQuery, Azure Synapse, SQL Server, data warehousing, data pipelines

SELECTED AI PROJECTS (PUBLIC, VERIFIABLE)

- **Mind2Web Computer-Use Agent** — trained DeBERTa-v3 candidate ranker (93% recall@10 on unseen domains; 95% recall@20) + FLAN-T5/GPT action prediction + Playwright execution. Model published: huggingface.co/torontodeveloper/mind2web-candidate-ranker; code: github.com/torontodeveloper/mind2web-computer-use-agent.
- **Sales Voice Agent** — Llama 4 voice agent (Pipecat, Deepgram STT, Cartesia TTS, Twilio), QLoRA/DPO fine-tuning, GPT-5 LLM-as-a-Judge evaluation, Locust load-tested to 1,000 concurrent users.
- **AI Interview Coach** — voice mock-interview agent with JD-ingestion RAG pipeline (Pipecat + Azure AI Foundry GPT-5.4).

PROFESSIONAL EXPERIENCE

Sr. Machine Learning / GenAI / LLM Engineer | Brooks Running | May 2026 – Present

- Hired to build ML and LLM-agent solutions across Brooks Running's digital product and customer-experience initiatives.
- Built a production chatbot using Cohere API, FastAPI, tool calling, and vector database retrieval for customer-facing AI workflows.
- Built historical chat endpoints with Redis-backed user conversation history for persistent, personalized chat experiences.

GenAI / LLM Engineer | Fleetworthy | Oct 2025 – Apr 2026

- Built production-oriented Llama 4 sales voice agent using Pipecat orchestration, Deepgram STT, Cartesia TTS, Twilio, FastAPI microservices, and SmartHub CRM workflow integration.
- Designed agentic AI workflows with RAG, LangChain, LangGraph, tool orchestration, domain-specific reasoning paths, and guardrails for fleet operations and sales automation.
- Implemented LLM adaptation workflows using supervised fine-tuning, QLoRA, DPO, vLLM inference, quantization, and customer/sales data preparation for production-quality responses.

- Created evaluation and governance controls using GPT-5 LLM-as-a-Judge, DeepEval, and Locust load testing up to 1,000 concurrent users.
- Built two-stage computer-use web agent to automate truck/trailer inspection submissions across web portals: fine-tuned DeBERTa-v3 element ranking trained on Mind2Web (evaluated with recall@k, MRR, element accuracy, action F1 across cross-task/website/domain splits), GPT-4o/FLAN-T5 action prediction, and Playwright browser execution.
- Developed Django/PostgreSQL and FastAPI backend APIs for fleet inspection, user management, vehicle lifecycle workflows, and operational data flows.
- Deployed open LLMs (Llama family) using AWS SageMaker and Databricks.

Machine Learning Engineer | T6 Health Systems | May 2025 – Sept 2025

- Designed ML pipeline for radiation exposure risk prediction using XGBoost, ridge/lasso regression, biomedical data, environmental sensor inputs, and explainability-driven evaluation.
- Applied SHAP-based explainability to translate model outputs, feature impact, and uncertainty into clinically interpretable insights for healthcare stakeholders.
- Built biomedical image classification workflows using ResNet50, VGG16, and HALO AI tools for whole-slide imaging and clinical decision support.
- Built React/TypeScript real-time dashboards for health monitoring across mobile and desktop, connecting model outputs to operational product workflows.

Machine Learning Engineer | Ontario Power Generation | Jun 2023 – Mar 2025

- Built enterprise GenAI platform on Azure using Databricks, Azure ML Studio, Azure OpenAI, Cognitive Search, Azure Synapse, and SQL Server for document search, report generation, and predictive maintenance.
- Designed and deployed production ML services for anomaly detection, image-based inspection, and radiation permit confidence scoring in regulated operational environments.
- Implemented containerized ML services and repeatable CI/CD delivery using Terraform, Docker, Kubernetes/AKS, Azure DevOps, and production data/ML pipelines.
- Embedded on-site at the client (OPG) as a forward-deployed engineer: navigated requirements directly with business and engineering stakeholders, and communicated model limitations, data-readiness gaps, deployment blockers, and operational risk in clear delivery terms.

Machine Learning Engineer | Scotiabank | Jan 2023 – Jun 2023

- Built credit card fraud detection pipeline using K-NN, logistic regression, SVM, and supervised ML evaluation for financial-risk classification workflows.
- Applied OpenAI LLM workflows with zero-shot/few-shot prompting for NLP classification in financial services; designed FastAPI microservices with authentication, monitoring, and Azure DevOps CI/CD.

Machine Learning Engineer | Bell Canada | Jul 2021 – Jan 2023

- Developed NLP classification, object detection, and customer churn prediction pipelines using scikit-learn and sentence transformers; deployed model training toolchains end-to-end.
- Built Wi-Fi checkup and video-feed features in React/Redux, deployed to production for desktop and mobile.

ML Engineer | Goodfood | Jan 2021 – Jun 2021

- Built food recommendation system (clustering/classification) deployed via MLflow and Vertex AI; CNN-based recipe and grocery image classification; sentiment analysis with HuggingFace.

Machine Learning Engineer | Bell Media | Sept 2019 – Dec 2020

- Built movie/TV sentiment analysis, clustering-based recommendations, and video genre classification (CNNs); developed Crave/CTV media application in ReactJS with Apollo GraphQL; deployment via Vertex AI.

ML Engineer – Oncology & Biomedical Imaging | CBRE / Navigating Cancer | Jan 2019 – Oct 2019

- Embedded on-site as a forward-deployed engineer with the client's oncology teams: predicted patient appointment no-shows, drug risk scores, and high-risk patient triage; integrated HALO AI for whole-slide image analysis; built clinical decision-support backends in Go and Python (Django/Flask) across multiple oncology projects, working directly with oncologists to turn their workflows into shipped systems.

Earlier Engineering Career | 2000 – 2019

- Frontend/full-stack engineering (Angular, React, Java/Spring) at TD Bank, Equifax, Scotiabank, RBC, and Rogers (2015–2019) — trading UIs (MIFID II/FIX), mortgage applications, microservices, accessibility, CI/CD.

- Earlier: Java/J2EE and systems roles at JPMorgan Chase, Infosys, Hewlett-Packard (Riyadh), and Fujitsu Network Communications (San Jose) — including RIPv1/v2, OSPF, IS-IS routing protocols in C++ (2000–2014).

EDUCATION

- Graduate Certificate — Generative AI & LLMs, Carnegie Mellon University (Jan–Dec 2026) — researching multimodal web agents
- Data Science & Machine Learning, University of Toronto — Data Sciences Institute (2025); additional AI/LLM programs: UC Berkeley (2024), University of Waterloo (2024), University of Toronto (2023–24)
- MBA — Business Intelligence & Data Analytics, University of Calgary (2014)
- B.Eng. — Computer Science, S.V.U. College of Engineering, India (2000)