TYLER ALIKA GEE

(808) 772-0333 | Tyler.gee13@icloud.com | linkedin.com/in/tyler-gee-ai | github.com/Ikaikaalika

PROFESSIONAL SUMMARY

Machine Learning Engineer with 6+ years building production NLP and document-intelligence systems, currently leading enterprise AI at Sunwest Bank. Specializes in domain-specific language models, agentic architectures, and end-to-end RAG pipelines — from OCR and layout analysis through structured extraction, vector retrieval, and multi-agent orchestration. Direct experience deploying LLM systems in highly regulated environments (BSA/AML, PHI-bearing clinical research) where data privacy, auditability, and structured outputs are non-negotiable.

TECHNICAL SKILLS

Languages & ML Frameworks: Python, PyTorch, TensorFlow, Hugging Face Transformers, JAX, MLX, scikit-learn

Language Models & Fine-Tuning: Llama, Phi-3, Qwen, GPT-4, Claude; QLoRA fine-tuning, RLHF, instruction tuning, prompt engineering, function calling, structured output enforcement, in-context learning

Agentic AI & Orchestration: Multi-agent frameworks, Azure AI Foundry Agent Service, LangChain, MCP (Model Context Protocol), tool/function calling, query routing, structured output enforcement

Document Intelligence & NLP: OCR, document layout analysis, Named Entity Recognition (NER), Relationship Extraction (RE), nested JSON schema design, semantic chunking, Transformers

RAG & Vector Stores: Retrieval-augmented generation, embedding generation, Pinecone, Azure AI Search (vector + hybrid), semantic chunking, hybrid retrieval

Cloud & MLOps: Azure (OpenAI, Foundry, ML, Functions, Cosmos DB, AI Search, Document Intelligence, Key Vault, App Service, Front Door/WAF, App Insights), CI/CD (Azure DevOps), Docker, model versioning, monitoring, cloud-platform agnostic

Compliance & Privacy: PHI/PII masking, NIST AI RMF, SR 11-7 / OCC 2011-12 model risk management, IRB-governed clinical data workflows, BSA/AML

PROFESSIONAL EXPERIENCE

AI/ML Engineer — Sunwest Bank — Salt Lake City, UT *Oct 2025 – Present*

Lead enterprise AI engineer building LLM-powered document intelligence and agentic systems across lending, underwriting, and BSA/AML compliance.

- Designed and implemented an end-to-end document ingestion pipeline that parses PDFs, scanned images, and Excel files using Azure Document Intelligence and Docling for OCR and layout analysis, then performs LLM-based Named Entity Recognition over the extracted text — grounded with years of historical labeled financial data as in-context examples to achieve production accuracy without fine-tuning.
- Built agentic RAG workflows over Azure AI Search, where retrieval is orchestrated by an LLM agent that decomposes complex queries, issues multiple targeted searches across customer profiles, regulatory rule libraries, and historical alert snippets, and synthesizes grounded responses with citation tracking.
- Built a multi-stage LLM pipeline for BSA/AML alert review that ingests transaction data, extracts structured entities, classifies alerts against a regulatory rule library via RAG, and generates SAR narratives — orchestrated across Azure OpenAI with strict structured-output schemas and full audit logging.
- Architected an agentic commercial prospecting system on Azure AI Foundry Agent Service that intelligently routes queries across specialized tools (data extraction, prohibited-business classification, weighted scoring, LOB routing), integrating ZoomInfo, D&B Direct+, Middesk, and IBISWorld via function calling.
- Designed nested JSON schemas to represent financial data extracted from commercial loan documents, transaction records, and counterparty research — including entities, relationships, and decision rationale — enforcing structured outputs from LMs to enable downstream rule-based processing and examiner-grade auditability.
- Currently fine-tuning Phi-3 with QLoRA on Azure ML (A100 80GB) over a curated corpus of historical credit memos that passed underwriting standards and audit, distilling production-quality structured extraction into a smaller, cheaper model. Benchmarking against three baselines (zero-shot Phi-3, few-shot Phi-3, grounded GPT-4) on field-level precision, recall, and JSON schema validity, with Docling as a parsing aid in the evaluation pipeline.
- Built Enclave, a multi-model M&A due diligence platform orchestrating Azure OpenAI and Anthropic models alongside tool-using sub-agents (browser automation, Python sandbox), with streaming chat and structured document export.
- Co-authored Sunwest's AI Governance Framework — a four-tier risk classification (Tier 0–3) built on an Autonomy × Exposure matrix with explicit alignment to SR 11-7 / OCC 2011-12 model risk guidance — establishing PII/sensitive-data handling standards, model evaluation criteria, and

human-in-the-loop boundaries that map directly to PHI handling in healthcare AI.

- Established CI/CD, model versioning, and observability across Azure Functions, App Service, Cosmos DB, AI Search, Key Vault, and Application Insights, with feedback loops from analyst review captured in Cosmos DB for continuous model improvement.
- Stand up a fully local LM evaluation environment (LM Studio + MLX backend on Apple Silicon) running Llama, Phi-3, and Qwen variants for prompt and architecture experimentation before promoting to cloud deployment.

Data Scientist / AI Engineer — Datafy — Ogden, UT *Aug 2024 – May 2025*

- Built and optimized ML models for anomaly detection over large-scale geolocation and event datasets, with production deployment via custom MLOps workflows.
- Implemented reinforcement learning from human feedback (RLHF) fine-tuning pipelines for language model–based forecasting, integrating analyst feedback into reward model training and policy updates — directly aligned with the JD's continuous-improvement and feedback-loop requirements.
- Contributed to large-scale language model research, including reward modeling and preference data curation.

Data Scientist / AI Researcher — University of Utah — Salt Lake City, UT *Aug 2018 – Oct 2025*

Part-time research role across academic terms and internship periods.

- AI team lead on a joint medical diagnostics research collaboration with UCSF, building deep learning classifiers (SVM, CNN) over patient breath and electrochemical sensor data — direct experience with PHI-bearing clinical data, IRB-governed workflows, and HIPAA-aligned data handling.
- Designed GAN architectures to synthesize patient data and address chronic data scarcity in clinical ML — a recurring challenge in regulated healthcare AI.
- Built a custom MLOps pipeline for continuous model retraining as new clinical data became available, and deployed classifiers to NVIDIA Jetson edge devices for in-clinic diagnostic use.
- Engineered SQL/REDCap API data pipelines for clinical data collected in Uganda; co-author on a paper analyzing GC-MS breath data classification (in progress).

Data Science / AI Intern — Micron Technology — Boise, ID *May 2022 – Aug 2022*

- Contributed to a deep learning pipeline predicting metrology statistics, reducing manufacturing decision time by 50%+; built supporting Python data tools (Django, Tkinter) and Tableau dashboards over Oracle/MS SQL pipelines.

SELECTED PROJECTS

mlx-turboquant — Open-Source Quantization Library for Apple MLX *github.com/Ikaikaalika/mlx-turboquant*

- Original implementation of the TurboQuant paper (ICLR 2026) for Apple MLX, delivering both low-distortion MSE quantization (Algorithm 1) and inner-product–oriented quantization via residual QJL (Algorithm 2) across tensor, KV cache, and full model weight pytrees.
- Integrated with mlx_lm via a runtime patcher and context manager to transparently compress KV caches during generation for Qwen, Llama, and Phi model families, with bit-width controls and per-layer compression telemetry.
- Full test suite, benchmark harness, and end-to-end smoke validation against Qwen models from Hugging Face — demonstrates a research-to-implementation pipeline and LLM deployment optimization skills directly relevant to serving specialized models economically at production scale.

Power-Retention-MLX — Linear-Complexity Transformer Alternative with Custom Metal Kernels *github.com/Ikaikaalika/Power-Retention-MLX*

- MLX implementation of Power Retention — an O(n) linear-complexity replacement for attention — including three custom JIT-compiled Metal GPU kernels (feature expansion, gated state update, output) and a dual-mode pure-MLX fallback path that supports autodiff for training.
- Complete LLM training pipeline built on the retention architecture, benchmarked at GPT-2 Small parameter count (125M) against WikiText-103 perplexity, with roughly 3x inference throughput over attention on long sequences on M1 Max.
- Demonstrates deep familiarity with transformer internals, custom GPU kernel development, and the training-vs-inference computation-path tradeoffs that matter for deploying domain-specific language models at scale.

DeepGen — Local-First Agentic Research System *github.com/Ikaikaalika/DeepGen*

- End-to-end agentic research pipeline that decomposes user queries, retrieves evidence across 10+ external APIs plus locally indexed user documents, runs contradiction checks, and synthesizes source-backed proposals for human-in-the-loop review.
- Multi-backend LLM architecture (OpenAI, Anthropic, local MLX), FastAPI service layer, Alembic-migrated data store, macOS Key-

chain–backed secret management, and a packaged native desktop app with CI and notarization pipeline.

- Four-stage research v2 pipeline (retrieval → extraction → contradiction checks → proposal synthesis) with agent-generated clarifying questions — the same agentic pattern that applies to clinical document research, claim denial analysis, and multi-source evidence gathering.

EDUCATION

B.S. Chemical Engineering, University of Utah *GPA: 3.3*

A.S., Weber State University, Ogden, UT *GPA: 3.8*