

Aarmen Sidhu

asa618@sfu.ca | linkedin.com/in/aarmensidhu | github.com/aarmens702-hub

EDUCATION

Simon Fraser University, Burnaby, BC

Sept 2024 - Apr 2028 (Expected)

Bachelor of Science in Data Science

Relevant Coursework: Computational Data Science, Linear Algebra, Statistics, Data Structures and Algorithms, Data Analytics and Visualizations

EXPERIENCE

Supply Chain Analyst, SFU Beedie Analytics and BC Hydro, Burnaby, BC

Apr 2025 - Jun 2025

- Built a composite risk model across 100+ suppliers and 15+ procurement categories in a 3-person team
- Modeled a 25% tariff scenario in Python and Tableau, presenting to BC Hydro and KPMG; placed 2nd in division

API Research Volunteer, TeejLab, Burnaby, BC

Dec 2023 - Aug 2024

- Wrote Python scrapers extracting endpoint, authentication, and schema metadata from 500+ APIs across 10+ industries
- Classified APIs by industry, function, and compliance, structuring the data for a production discovery platform

PROJECTS

Mailo, Multi-Agent LLM System

AWS Bedrock AgentCore, Python, Strands, Cedar

- Built a multi-agent GenAI system on AWS Bedrock AgentCore, a thin LLM orchestrator coordinating 3 domain specialists over the A2A protocol
- Engineered per-agent tool scoping over MCP and isolated each specialist's context, cutting hallucinations and token cost vs. a monolithic prompt
- Built an LLM-as-judge evaluation pipeline (custom rubric plus built-in evaluators) that scores a 25% sample of agent responses to catch quality regressions
- Integrated long-term memory, namespaced and isolated per user, to personalize campaign planning across sessions

GreenLeaf, Agentic RAG Assistant (2nd place)

Python, scikit-learn, sentence-transformers, LLMs, Streamlit

- Built an agentic RAG assistant where an LLM chains 12 tools, semantic search over a local MiniLM index plus pandas analytics, citing every source
- Compared gradient-boosting and logistic regression, shipping the interpretable model (0.78 ROC-AUC) as a watch-list that caught 2.3x more stress cases than the baseline
- Hardened the RAG pipeline with a versioned index and a 34-question test of retrieval accuracy, gated in CI to catch quality drops

GenAI-Stack-Coder, LLM Fine-Tuning Pipeline

Python, PyTorch, QLoRA, Hugging Face, Ollama, FastAPI

- Fine-tuned Qwen2.5-Coder-7B with QLoRA on a single 12GB GPU, raising accuracy on current OpenAI and Anthropic API code from 12% to 72%
- Designed the 50-prompt benchmark behind that result, scoring syntax validity and API currency automatically; the tuned model produced zero deprecated API calls
- Curated a 6,700-pair instruction-tuning dataset from permissively licensed code with scripted harvesting, filtering, and deduplication
- Wrote a coding agent harness that let the fine-tuned model write, run, and debug its own chat app end to end

Image Generation on CIFAR-10, Generative Deep Learning

Python, PyTorch

- Built a class-conditional CIFAR-10 image generator in PyTorch with a VAE latent space, particle-based drift loss, EMA weights, and automated FID evaluation
- Cut FID by about 70% (330 to 99) by diagnosing high-dimensional mode collapse and adding a diversity-matching loss, exploring kernel-gradient drift as the recovery mechanism
- Designed a controlled, multi-seed ablation framework and audited the pipeline, catching inverted evaluation metrics that had invalidated earlier headline results

SKILLS

Languages: Python, SQL, R, JavaScript/TypeScript, C++

ML and GenAI: PyTorch, scikit-learn, Hugging Face Transformers, QLoRA fine-tuning, quantization, sentence-transformers, RAG, LLM agents and tool-calling, LLM evaluation

Tools and Infra: FastAPI, Docker, ONNX Runtime, AWS Bedrock, Ollama, Git, GitHub Actions, CI/CD, pandas, NumPy, Streamlit, DuckDB