Duy Nguyen

Seattle, WA | dnguyen44@seattleu.edu | duyng-portfolio.com | linkedin.com/in/duwe-ng | github.com/dcnguyen060899

Professional Summary

I am an M.S. Data Science candidate at Seattle University with two research roles. In one I designed a 24-cell controlled experiment on retrieval-augmented mammography report generation and first-authored a paper accepted at PSB 2027; its lesson is that a train/test split alone is not enough: a retrieval metric must clear the prevalence floor and its gain must propagate downstream. In the other I built the 3NF SQLite database that turned a neuroscience lab's selection criteria into reproducible SQL. I bring experimental design, reproducible evaluation, and the habit of checking a number before I report it.

Education

Seattle University — *Seattle, WA*
M.S. in Data Science. GPA 4.0; Dean's Graduate Student Honor Roll, Winter 2026. — *Expected Jun 2027*

UC Berkeley College of Engineering (Online)
Professional Certificate, Machine Learning & AI. Capstone selected as a program exemplar. — *Jul 2024*

Simon Fraser University — *Burnaby, BC, Canada*
B.A. in Economics, Data Analysis Concentration. — *May 2023*

Experience

Graduate Research Assistant, Medical Vision-Language RAG — **Feb 2026 – Present**
Seattle University · Advisor: Wenjing Yang, PhD — *Seattle, WA*

- Replicated and audited a published mammography-RAG benchmark on VinDr-Mammo (5,000 studies; patient-level 80:20 split) in Python and PyTorch; found the headline scores mislabeled, with weighted metrics reported as macro, which motivated a controlled re-evaluation of the full pipeline.
- Designed and ran a 24-cell controlled experiment: 4 encoder substrates (OpenCLIP, BiomedCLIP, BioCLIP, Mammo-CLIP) x frozen/fine-tuned x 3 frozen VLM decoders (MedGemma, Qwen2.5-VL, LLaVA-Med), with one identical supervised-contrastive recipe across all arms; HuggingFace Transformers, ChromaDB, scikit-learn.
- Showed that a train/test split alone is not sufficient evidence of generalization here: three fine-tuned out-of-domain encoders inflated train-set margin 9 to 23x while held-out P@1 rose only from below the 0.487 majority-class floor to a tie with it; only mammography-native Mammo-CLIP cleared it (0.585, p = 3.3e-10).
- Verified the gain propagates to the report with every decoder frozen: fine-tuning Mammo-CLIP alone raised suspicion macro-F1 by 0.11 (Qwen2.5-VL) and 0.06 (LLaVA-Med) and five-class BI-RADS accuracy by 10.5 and 6.6 points (23.2% and 14.6% relative), mostly on majority classes; retrieval-insensitive MedGemma was the control.
- Distilled this into a component-wise auditing protocol (held-out P@1 vs the prevalence floor, train/test margin ratio, paired within-decoder tests under multiplicity control) and self-verifying Python generators that recompute the paper's 100+ metrics byte-exact from raw predictions; first-author paper accepted at PSB 2027, oral presentation.

Research Data Engineer, Computational Neuroscience Research Group — **Mar 2026 – Present**
Seattle University · Supervisor: Prof. Brian Fischer — *Seattle, WA*

- Built the research database for an NIH CRCNS-funded barn-owl sound-localization project: curated a 53 GB, 30,147-file raw archive to 1,579 staged files and loaded 1,325 recording passes, 195 neurons, 7 owls, and 13 paradigm types into a 3NF SQLite schema (7 tables, 5 views, CHECK/UNIQUE rules) with a read-only pandas loader.
- Found what hand curation could not see: keying every recording on filename plus header resolved all 77 of the PI's hand-selected neurons and surfaced 118 recorded-but-never-analyzed neurons, 49 of which carry the exact ITD x ILD measurement his selection criteria require.
- Rewrote those criteria as SQL over all 195 neurons at the PI's request, replacing a per-neuron file loop from a hard-coded laptop path with one reproducible query that screens the whole set in about a second; added a depth_source column recording whether each depth came from the instrument header or the curated position file.
- Engineered the ETL for failure across 5+ formats (XDPHYS binary, MATLAB, headerless CSV): copy-only staging with SHA256 manifests never mutates the raw tree, checksum dedup kept three same-named but different recordings a filename rule would have dropped, and an atomic build-then-swap publish survived a mid-fill disk I/O abort.
- Audited the team's pipeline output against the lab's 8 file-naming and content rules, found 164 violations across all 16 uploaded date folders and traced the largest class to one hard-coded line in the shared renaming notebook; automated integrity checks also caught all-zero rows and weak-tuning exclusions before analysis.

Earlier Experience

Research RAG Engineer **Jan 2024 – May 2024**
SFU Faisal Lab (medical-imaging AI), Simon Fraser University *Burnaby, BC*

- Built a Chainlit chatbot (Python, LlamaIndex, GPT-4, Pydantic) that turns a clinician's plain-English CT / PET-CT request into the deterministic slab and ROI JSON the lab's DAFS tool expects, asking a follow-up when slab or ROI is missing or unclear.
- Compared two designs; the simpler won: a system-prompt agent with typed output gave correct JSON more reliably in the demo than a tool-calling OpenAIAgent with query-engine tools, which I diagnosed as emitting malformed output; the supervisor adopted it.
- After the role (Dec 2025 to Jan 2026), refactored the single-file scripts into a modular deployed app (Claude Sonnet 4.5 via the Anthropic API, Pydantic, Chainlit, Render) with a live demo, applying the modular structure learned at Blueprint.

AI Engineer **Mar 2024 – Oct 2024**
SFU Blueprint · Client: MOSAIC (immigrant and refugee settlement services) *Burnaby, BC*

- Built features, with a tech lead, a senior developer and a junior engineer, for MOSAIC's settlement-services assistant (Neo4j knowledge graph and vector index over 50+ programs, LangChain, GPT-4o mini, Flask); delivered in the team's production handoff.
- Owned a fallback ticket from a senior developer in the modular RAG pipeline: a second LLM call judges whether the primary Cypher answer is specific and, if not, triggers vector-search fallback, as no parser can enumerate every phrasing of "no result".
- Traced a defect (backend emitted hyperlinks, React showed raw text) to missing rendering logic and fixed it with two engineers and the tech lead; shortlisted (top four) for the SFU Computing Science Diversity Award on a signed MOSAIC endorsement.

Awards & Honors

- Winner, Graduate Division, CAUSE Student Data Scrollytelling Contest 2026: "Will AI Make Human Work Worthless or Priceless?", a nine-act data story (D3.js, Scrollama.js) with a reader-controlled parameter widget, translating an original nested-CES general-equilibrium model of AI and cognitive labor for non-specialists.
- 12th of 111 teams, Computer Vision track, ISODS World Championship in Data Science and AI (Kaggle-hosted, Oct 2023): embryo-quality classification from day-3/day-5 microscopy with a fine-tuned ResNet50 in PyTorch, class-weighted loss, 5-fold stratified CV, and F1 threshold tuning.

Skills

ML / AI: PyTorch, HuggingFace Transformers, scikit-learn, RAG, vision-language models, supervised contrastive fine-tuning, QLoRA / parameter-efficient fine-tuning, LLM-as-judge evaluation, LlamaIndex, LangChain, Pydantic, OpenAI and Anthropic APIs
Experimentation & Causal Inference: experimental design, factorial designs, controls and negative controls, prevalence floors and confidence intervals, hypothesis testing, effect-size estimation, multiplicity control, replication and reproducibility auditing
Data Engineering: pandas, NumPy, SciPy, SQLite, ETL design, atomic publish and checksum manifests, relational schema / ER design (3NF), data quality and provenance, ChromaDB, Neo4j / Cypher
Programming: Python, SQL, JavaScript
Tools: Jupyter, Git, LaTeX, D3.js, Scrollama.js, Chainlit, Flask, Docker, Render, Google Colab (A100 GPU), Ollama