

Henry Vu

469-649-6967 | Vuducdung28@gmail.com | linkedin.com/in/henry-vu27 | github.com/HenryVu27 | henryvu.io | henryvu.blog

EXPERIENCE

Founding Machine Learning Engineer

Aug 2025 – Present

eXRealityAI

Dallas, TX

- Built the **knowledge graph** behind a Meta Quest 3 automotive repair app for an OEM pilot: parts, symptoms and procedures as typed entities merged from five sources, with a build pipeline that blocks release on bad data.
- Built the **retrieval layer** that finds the right repair context for a technician's question, filtered to their vehicle. Correct procedure first **92%** of the time on an expert-written test set, with no steps from unrelated repairs.
- Built its **evaluation harness**: automated scoring in CI plus an LLM judge. Tested the scoring against random baselines, found it caught a broken ranker but could not separate two good ones, and raised that limit.
- Traced bad production answers to a vendor **LLM proxy silently discarding our images**: identical request sizes with and without a photo, and a confident plant ID for a blank pixel. Separately cut a request from **194,000 to 1,900** tokens.
- Shipped **GardenXR** to the **Meta Horizon Store**, passing Meta's review for restricted camera access. Fixed unreliable plant IDs by constraining the model to a strict output schema and lowering the provider-default temperature.
- Wrote **nine API specifications** for a partner team, citing the exact code behind each defect. Handed the knowledge layer to a backend team outside the repo, excluding every data source whose licence did not allow it.

Computer Vision Engineer

Feb 2025 – Present

ThorMed Innovation

Dallas, TX

- Led domain-aligned transfer learning for an NIH-funded bladder segmentation project, pretraining **U-Net** and **SimSiam** SSL encoders on **9.2K** ultrasound images to reach **96%** Dice.
- Enabled edge deployment via **4-bit** quantization, holding up **7x** better than starting from ImageNet weights.

Research Assistant

Apr. 2022 – May 2024

Alberta Machine Intelligence Institute (Amii)

Edmonton, AB

- Implemented online optimization algorithms with the **primal-dual** framework, improving competitive ratios beyond worst-case analysis using ML predictions.
- Surveyed and benchmarked adversarial, Markovian and restless **multi-armed bandits** (UCB, Exp3, Gittins Index) on real-world data. [\[REPO\]](#)

SELECTED PROJECTS

DatDai: Deployed Legal RAG Assistant | *FastAPI, Gemini, Qdrant, Langfuse, Cloud Run*

- Vietnamese land-law assistant live on **Cloud Run**, deployed by GitHub Actions with Workload Identity Federation and Secret Manager.
- **Langfuse** traces every request, and a batch job replays real production traffic through an LLM judge and writes the scores back, so it is **evaluated on live usage** rather than a fixed test set.
- Models the amendment graph between statutes so retrieval resolves which decree amends which, plus a citation verifier that needs no LLM call.
- Ran my own pre-launch deployment audit, which found and fixed cascading timeouts in a model fallback chain. [\[REPO\]](#)

ADHDAgent: Guardrailed Agentic RAG | *LangGraph, Gemini, Qdrant, React 19*

- **LangGraph** ReAct agent with guardrails as graph nodes: crisis and jailbreak detection on the way in, medication and scope validation on the way out.
- Hybrid RAG over **Qdrant** dense + sparse with **RRF** and a local ONNX **cross-encoder reranker**, plus 4-tier memory and complexity-based model routing.
- Offline eval package with golden retrieval and memory sets, and LLM judges for coherence, tool use and pairwise response quality. [\[REPO\]](#)

EDUCATION

University of Texas at Dallas

Richardson, TX

M.Sc. Computer Science, Intelligent Systems (GPA 4.0/4.0)

Sep. 2024 – May 2026

University of Alberta

Edmonton, AB

B.Sc. Computing Science with Honors (GPA 3.83/4.0, Summa cum laude)

Sep. 2019 – May 2024

TECHNICAL SKILLS

LLM & Agents: LangGraph, ReAct agents, tool calling, structured output / JSON schema, guardrail gates, context engineering, prompt versioning; Gemini, Claude, OpenAI.

Retrieval & Evaluation: knowledge graphs (typed ontologies, graph-scoped traversal), hybrid dense+sparse (Qdrant, RRF), cross-encoder reranking, query rewriting, BM25/IDF; golden sets, LLM-as-judge, citation containment, scorer ablation, online eval on production traces (Langfuse).

ML & Infrastructure: PyTorch, HuggingFace, ONNX, quantization (PTQ), Python, FastAPI, Docker, GCP (Cloud Run, Cloud SQL, GCS, Secret Manager), Postgres, GitHub Actions, C#/Unity, TypeScript.