

Daniel He

437-238-5756 | dhe72@uwo.ca | [LinkedIn](#) | [GitHub](#)

EDUCATION

Western University

London, ON

Bachelor of Science in Computer Science, Honours Specialization

Sept. 2024 – May 2028 Expected

- Admission Scholarship & Scholarship of Distinction (\$2,500) | **GPA: 3.50/4.00**

EXPERIENCE

Software Engineering Intern

May 2026 – Aug. 2026

OpsGuru (AWS Premier Partner), AWS AI/ML Infrastructure & Agent Systems

Toronto, ON

- Built a reusable **AWS Bedrock AgentCore** accelerator with 9 AI agents using Python, **LangGraph**, and Strands, integrating Bedrock Knowledge Bases, S3, DynamoDB, and **Terraform** infrastructure used by **15 Engineers**
- Designed and developed a Deep Agent accelerator with multi-step planning and token-aware memory management, reducing retained memory tokens by **~40%** through threshold-based compression

Undergraduate Research Assistant

Oct. 2025 – May 2026

Western University, LLM Hallucination — Dr. Apurva Narayan

London, ON

- Co-authored *HalluciBench*, an LLM hallucination benchmarking paper submitted to COLM 2026
- Designed and implemented an LLM-as-a-judge benchmarking pipeline in Python with **Ollama** to evaluate hallucination rates across **29 LLMs**, incorporating hyperparameter tuning and **Prompt Engineering**
- Implemented a **Reinforcement Learning**-inspired **Synthetic Data** generation pipeline from a research paper to create a **1,000-Sample** dataset for evaluating LLM refusal calibration and hallucination rates

Software Engineering Intern

May 2025 – Aug. 2025

Scotiabank, IT&S Capital Markets/Data Analytics

Toronto, ON

- Refactored a broken **Airflow** DAG pipeline to store vector embeddings in **ChromaDB** with persistence to **Google Cloud Storage**, migrating embedding logic into a Cloud Function; cut p95 latency by **~90%** and resolved a six-month production issue, enabling a scalable RAG knowledge base for the bank's internal knowledge bot
- Enhanced an internal research summarization pipeline by extending PDF parsing support from 2 to 4 document formats, automating deployment through **CI/CD** pipelines, and documenting the system using Jira and Confluence
- Built a **Python** cyclomatic complexity analyzer using SonarQube and Selenium, reducing analysis time by **~85%**

PROJECTS

Dex2 ([GitHub](#), [Video Demo](#)) | *TypeScript, React.js, Python, FastAPI, MongoDB, Supabase, LangChain* Feb. 2026

- Built a TypeScript/React.js Chrome extension that automatically captures page content to trigger context-aware agent actions, including editing slides and drafting emails, using MongoDB and Supabase (**PostgreSQL**)
- Implemented AI **Agent Orchestration** using FastAPI for hybrid retrieval and **Tool Calling**, with background workers and concurrency to keep chat non-blocking; used LangChain for chunking and retrieval

Mini-v ([GitHub](#)) | *C++, LLM Inference*

Apr. 2026

- Built a C++ LLM inference server with micro-batching and continuous batching, achieving a **2.5× Throughput** improvement from 6.5 to 16.8 requests per second under concurrent load with a 100% request success rate

Say Less (Devpost, [Live Demo](#)) | *Python, FastAPI, MediaPipe, Docker, Render*

Nov. 2025

- Won **MLH Best Use of ElevenLabs** at Hack Trent 2025 among 21 submissions
- Created a website that converts American Sign Language (ASL) to speech and speech to text in real time
- Spearheaded development of a FastAPI backend with RESTful API endpoints, integrating a fine-tuned MediaPipe model with approximately **80% ASL-to-Text Accuracy**; containerized the application with Docker and deployed it on Render

TECHNICAL SKILLS

Languages: Python, Java, JavaScript, TypeScript, C, C++, SQL | **Spoken Languages:** English, Chinese (Mandarin)
Frameworks: Next.js, React.js, FastAPI, MediaPipe, Tensorflow, PyTorch, SQLAlchemy, LangChain, LangGraph
Developer Tools: Docker, Git, Vercel, GCP, Apache Airflow, AWS, Ollama, Linux, Claude Code, Cursor, Terraform
Databases: PostgreSQL (Supabase), SQLite, MongoDB, ChromaDB