Daniel He

437-238-5756 | dhe72@uwo.ca | Toronto ON | GitHub | LinkedIn | DanielHe.site

Education

Western University - BSc., Computer Science *September 2024 - May 2028 Expected*
With Admission Scholarship, Scholarship of Distinction ($2500)
Deans Honour List 2024-2025 (**3.93/4.00 GPA**)

Work Experience

OpsGuru - Software Engineering Intern *Incoming May 2026-August 2026*
AWS AI/ML Infrastructure & Agent Systems

Western University - Undergraduate Research Assistant *October 2025 - Present*
LLM Hallucination - Dr. Apurva Narayan

- Co-authored *HalluciBench*, an LLM hallucination benchmarking paper, under review at 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 **1000-sample** dataset for evaluating LLM refusal calibration and hallucination rates

Scotiabank - Developer 1 (Intern) *May 2025 - August 2025*
IT&S Capital Markets/Data Analytics

- 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 6-month production issue to enable scalable AI workflows that support RAG for the bank's internal knowledge bot
- Enhanced internal research summarization pipeline by extending PDF parsing support from 2 to 4 document formats, automated deployment using **CI/CD** build configuration tools, and documented using Jira and Confluence
- Built a **Python** cyclomatic complexity analyzer using SonarQube and Selenium, reducing analysis time by ~85%

Projects

Dex2 (GitHub Link, Video-Demo) *February 2026*

- Built a **TypeScript** (**React**.js) Chrome extension that auto-captures page content to trigger context-aware agent actions (edit slides, open tabs, draft emails) using **MongoDB** for embeddings 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
- Built a Google Slides agent combining **OAuth**, GPT-4o-mini and Gemini vision to generate and edit slides directly in live presentations via the Slides API, using vision-based style extraction to match existing formatting

Mini-v (GitHub Link) *April 2026*

- Built a **C++** LLM inference server with micro-batching and continuous batching, achieving **2.5x throughput** improvement (6.5 to 16.8 req/s) across concurrent load with 100% request success rate

Say Less (Devpost Link, Live-Demo) *November 2025*
Hack Trent 2025 - MLH Best Use of ElevenLabs (21 total 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 (~80% accuracy on ASL to text). Containerized with **Docker** and deployed on **Render**

Skills

Languages: Python, Java, JavaScript, TypeScript, C, C++, SQL
Frameworks and Libraries: Next.js, React.js, FastAPI, MediaPipe, Tensorflow, PyTorch, SQLAlchemy, LangChain
Developer Tools: Docker, Git, Render, Vercel, GCP, Apache Airflow, ChromaDB, SonarQube, Selenium, Ollama, Linux
Databases: PostgreSQL (Supabase), SQLite, MongoDB Atlas