

# GAL COHEN

[linkedin.com/in/Galc3882](https://www.linkedin.com/in/Galc3882) [gal.cohen@mail.utoronto.ca](mailto:gal.cohen@mail.utoronto.ca)  
[github.com/Galc3882](https://github.com/Galc3882) 647-208-9330

Engineering Science student (AI & Robotics) graduating May 2026. Experience includes building LLM tools for 16k+ users at Intuit and leading a computer vision team to 7 international competition wins. Seeking a full-time New Grad position to leverage my passion for building robust, scalable AI solutions.

SKILLS	LANGUAGES	LIBRARIES & FRAMEWORKS
<ul style="list-style-type: none"><li>Generative AI</li><li>Computer Vision</li><li>Model Optimization</li></ul> <ul style="list-style-type: none"><li>Cloud Architecture</li><li>Autonomous Systems</li><li>Reinforcement Learning</li></ul> <ul style="list-style-type: none"><li>Algorithm Design</li><li>Technical Leadership</li><li>Deep Learning</li></ul>	<ul style="list-style-type: none"><li>C++</li><li>C</li></ul> <ul style="list-style-type: none"><li>Python</li><li>GoLang</li></ul>	<ul style="list-style-type: none"><li>OpenCV</li><li>PyTorch</li><li>TensorFlow</li></ul> <ul style="list-style-type: none"><li>ROS</li><li>NumPy</li><li>scikit-learn</li></ul>

## EXPERIENCE

### INTUIT | SOFTWARE ENGINEER INTERN, AI & FULL-STACK

SEP 2024 – AUG 2025

- Architected an AI incident agent for 1,000+ engineers in Go, AWS Lambda, API Gateway, Step Functions, CDK; cutting MTTR by 30%+.
- Engineered a multi-tool conversational AI using OpenAI GPT-5 & Gemini 2.5 Pro for incident diagnosis, reducing engineer hours by 15%.
- Developed a GenAI service from scratch to generate RCA summaries, questions, and action items from Slack and ServiceNow data, saving 1000+ engineering hours per quarter and achieving 85%+ semantic similarity with human-authored RCAs.
- Built AI risk & quality modules using both LLMs & algorithms, projected to cut high-risk change failures by 25% for 8,000+ engineers.
- Refactored a critical service from Go-routines to a scalable, asynchronous Lambda invocation, eliminating 100% of timeout errors.
- Created a data pipeline in Go using AWS S3 for model evaluations, enabling data-driven improvement loop through semantic analysis.
- Improved service latency & accuracy by 20%+ by migrating AI models to Gemini 2.5 and shipping critical features for HCAP management.

### AUTORONTO UOFT | 2D VISION TEAM LEAD (GM-SAE AUTODRIVE CHALLENGE)

SEP 2023 – PRESENT

- Spearheaded a 9-engineer team to 7 first place wins at the SAE AutoDrive Challenge by architecting a perception system with YOLOv10 boosted model mAP by 12% and accuracy by 18% while slashing latency by 35% on embedded NVIDIA Jetson via TensorRT optimizations.
- Engineered a high-throughput data infrastructure featuring a semi-automated labeling pipeline with SAMv2 that tripled dataset preparation speed for over 200k frames; simultaneously improved model robustness in adverse weather by 22% using advanced augmentations like synthetic rain and CutMix, which slashed false positives by 40% and increased validation consistency by 15%.
- Scaled the 2D vision stack by expanding traffic sign classes to 7 types with 95%+ accuracy & integrated the whole perception module with ROS2 for localization & planning; validated performance & reliability across a suite of 500+ real-world and simulated driving scenarios.

### SWAP COMMERCE | SOFTWARE ENGINEER INTERN

May – SEP 2023

- Architected a high-performance admin dashboard in Flutter & Dart to manage operations for 50+ enterprise clients (e.g., Sirplus, Aspiga), reducing manual workflow time by 30% via digitizing legacy onboarding processes.
- Led a backend optimization initiative that reduced server requests by 90% via query caching and state management fixes, directly unlocking \$1M in additional revenue by stabilizing the platform for high-volume sales events.
- Refactored legacy codebases into modular REST APIs, increasing system reliability and code coverage by implementing end-to-end unit testing pipelines that enabled safer scaling for new partner integrations.

## PUBLICATIONS

### INCLUDE: EVALUATING MULTILINGUAL LANGUAGE UNDERSTANDING

ICLR 2025

- Co-authored an ICLR 2025 Spotlight paper (top 5% of accepted papers) on reasoning-centric benchmark of 197K QA pairs across 44 languages to evaluate LLMs performance in regional contexts.
- Addressed critical gaps in multilingual AI deployment by demonstrating significant performance disparities in state-of-the-art models when processing region-specific cultural knowledge versus standard translated benchmarks.

## EDUCATION

### BASC IN ENGINEERING SCIENCE + CO-OP

University of Toronto

Sep 2021 – May 2026

- Major: AI & Robotics + Minor in Business
- 4.0 GPA
- 5 x Dean's Honours List

- Thesis:** Geometry-Aware Domain Adaptation for 3D Lane Detection in Adverse Weather | Supervisor: Prof. Steven Waslander.
- Achieved a grade of 100% in Data Structures and Python course.
- Relevant courses:
  - CSC401: Natural Language Computing
  - CSC412: Probabilistic Learning and Reasoning
  - ROB501: Computer Vision for Robotics
  - APS360: Applied Fundamentals of Deep Learning