GUAN HUANG

APPLIED ML ENGINEER

guanzheng66.huang@gmail.com linkedin.com/in/guan0huang github.com/CoronRing coronring.github.io

Applied ML Engineer with **3 years** building AI systems end to end, from agents and evaluation down to hardware-level work. Shipped **17 production agents** used across hundreds of deployed products, including a tender matcher at **95% accuracy** (GPT-4o with RAG: 71%), and trained or tuned **5 LLMs**. Led teams of 3–5 and created Railtracks, an open-source agent framework (**200+ GitHub stars**).

PROFESSIONAL EXPERIENCE

Aug 2026 – Present

AI Educator & Workshop Lecturer / Built Young PART-TIME

- Teach secondary school students and teachers how to work with AI agents, AI-assisted coding and project management (about 10 hours a month).

Jan 2026 – Present

Teaching Assistant / University of Toronto

- Project supervisor for course CSC301, directly supervising 41 students.
- Technical guidance on course projects and evaluating student performance.

Nov 2025 – Jun 2026

Student Researcher / University of Toronto

- Evaluated LLMs and agents on **suicide risk assessment** over four levels (no risk, ideation, past behaviour, active behaviour): **12 models from 5 providers** on **4,000 cases**, across reasoning levels and popular agent prompts (Hermes, Copilot, LangGraph agents), on a $1,000+ API budget.
- Best result: GPT-5.1 at low reasoning, **F1 0.78**. More reasoning did not help: Gemini 3 Flash fell from F1 0.49 at low to 0.43 at high and shifted its predictions toward the attempt levels, while GPT-5.1 did not over-escalate.
- Methods and results checked with practising clinicians; the findings informed a health system's design and follow-up work. Research group of 2.

Jan 2024 – Nov 2026

Machine Learning Engineer / Railtown AI Technologies Inc.

- Led a team of five to ship Conductr (Dec 2023 – Sep 2025), an early end-to-end **DevOps agent** integrating Azure, Jira, and GitHub. It automated work item triaging, stack trace monitoring, and query responses through source code and Git history. Preferred in 90% of blind, remote pairwise comparisons by engineers and students against a leading commercial coding assistant (two-sided binomial, $p<0.01$).
- In a trial with engineers on development and management tasks, Conductr cut average task time by **34.7%** (about 11 minutes a task; ticket creation from 11 to 3 minutes, pull request creation from 17 to 14). Engineers stopped gathering information from GitHub and Azure DevOps and only confirmed the drafts Conductr prepared. In production it processes at least **10,000 logs a week** (a lower estimate) at **99.98% uptime** (Oct to Nov 2026).
- Led or was the main engineer on **17 major production agents** (100+ agents built in total, counting smaller ones) across legal, commerce, software and healthcare, such as pre-trial defence plan gap analysis, stack-trace analysis, an email agent and medical record summaries, and oversaw more.
- Built **code understanding** into 3+ production agents over about a year: parsed repositories into **ASTs** with tree-sitter and built **call graphs**, so agents navigate code as a structure rather than as flat text.
- Gave production agents **memory across sessions**: long-term memory for a personal assistant, per-developer **preferences** for coding agents, and search across session history, alone or combined.
- Designed and created Railtracks, an Open-source **Agentic Framework** aimed at enhancing the design, building, debugging, and monitoring of AI systems.
- Built **ProcuRing**, a full-stack tender-matching system (React front end) scaled on Google Cloud and shipped to clients: **95% accuracy** against 71% for GPT-4o with RAG; the company used it to bid on 6+ public tenders (see Projects).
- Built **LocateRing**, a site-selection agent that ranks vacant storefronts for a business idea: the true site lands in the top 5 in **54%** of 200 cases against 7% for an OpenHands agent (see Projects).
- Retuned the root cause analysis model on **Qwen3-30B**: a strong model labelled varied data, then supervised fine-tuning with **QLoRA**; served in **Docker** on AWS and Google Cloud. F1 0.94 on 1,200 real, human-labelled historical cases (exceptions with their tickets), up from 0.85 with BERT.
- Ran **vLLM** serving on **8 H200 GPUs** (an Alberta facility) for clients with local deployment needs: **live requests**, with no access to client data, for DeepSeek R1, Qwen models, and embedding and audio generation models. A test with 2 parallel requests gave **91.0 tokens/s** with a 571 ms time to first token.
- Conducted market research and competitor analysis to prioritize features and competitiveness.
- Conducted user studies and requirements engineering for a new AI product.
- Collaborated with Mila researchers to develop NLP models over noisy production stack traces and logs.
- Built **AEval**, the company's agent **evaluation framework**, embedded in hundreds of deployed products. Unlike the end-to-end evaluators of the time, it takes checkpoint submissions, so each stage is scored on its own (speed, number of turns, cache cost) down to individual tool calls. In the legal project it cleaned, categorized and versioned the dataset and evaluated every stage. It ran daily in production as a **batch job**, with a React front end.
- Built **AEval Gate**, attached to each deployed agent: it labels gold stages that must never be skipped (such as source validation) and enforces them in production, which guards against

context drift and hallucination, especially in weaker models. In initial tests it blocked **99% of prompt injection attacks** and cut hallucination by **36%** (11% to 7%). An end-to-end form requires each release to pass a gold set.

- External project consulting and research, selected work presented in the projects section.

Oct 2024 – Apr 2025

Student Researcher / UBC SALT Lab – ECE Department

- Participated in topic "**Control-flow Guided Test Generation**": a method of utilizing execution paths to aid language-agnostic test case generation (88% to 93% branch coverage with Llama3).
- Proposed and led topic: "**SoC Visual Schematic File Generation Using VLM**": a method of interacting with HDL (hardware) file through VLM (Vision-LM) with entity db and feedback loop.
- **Accessible course-portal agent** for students with disabilities: natural-language requests (open a page, check a grade) carried out on Canvas, with speech input and read-out. Preferred by all 11 students over 6 tasks in A/B testing. Research group of 2.

May 2022 – Dec 2023

Data Scientist Intern / Railtown AI Technologies Inc.

- Designed a **multi-agent scoring ensemble** before agent frameworks existed: models score outputs against a rubric, and a reviewer agent is dispatched to reconcile disagreements and consolidate a decision. 91% user preference in A/B testing (two-sided binomial, $p<0.001$). Presented at IEEE UBC.
- Fine-tuned a **BERT** model for root cause analysis (RCA), raising F1 from 0.71 to 0.85 on 1,200 real historical cases and outperforming the OpenAI API baseline of the time; established its fine-tuning pipeline.
- Led a team of five to research and develop an **LLM Professional Knowledge Testing Framework**, which automated the evaluation of AI expert systems, including security and fairness assessments.
- In-depth study and research on language-agnostic Root Cause Analysis (RCA).

LEADERSHIP

Sep 2025 – Present

Vice President / UofT CS Graduate Student Society (CSGSS)

- Represented all CS graduate students (600+) in university governance and industrial outreach, connecting students with over 20 firms.

Oct 2025 – Present

Event Volunteer & Dept. Contact / UofT Graduate Students' Union (UTGSU)

- Facilitated democratic processes at Annual General Meetings and served as the primary liaison between the CS department and the central union.

May 2022 – Aug 2025

Head Secretary / IEEE UBC Student Chapter

- Planned and organized internal meetings and technical workshops. Led recruitment and onboarding.
- 150% increase in new member enrollment; organized 75% more public events.
- Facilitated collaboration between sub-teams and the executive group.
- Leveraged software tools to enhance communication flow and team productivity, such as designing a custom IEEE file management system and an LLM + Whisper meeting summarizer.

EDUCATION

Sep 2025 – Nov 2026

Master of Science in Applied Computing – Artificial Intelligence / University of Toronto

Selected Award: Offered Vector Scholarship in Artificial Intelligence (2025, scholarship)

Affiliation: CS Graduate Student Society **Vice-President**, presenting all graduate students.

Sep 2019 – Apr 2025

Bachelor of Applied Science – Computer Engineering (top 1%) / University of British Columbia

Grade: 93, graduated with honours

Selected Awards: Trek Excellence (2025, scholarship), PMC-Sierra Founders (2024, scholarship), James Yan Award (2024, scholarship), CPEN-SE: First Place (2024, hackathon), EDA ELITE: Second-Class Honor (2022, international competition)

PROJECTS

Jan 2025 – Present

ML Team Lead / FeatherRing COMMUNITY PROJECT

- Designed an **operating-system controlling multi-agent system** that understands all media types.
- The music analysis and creation module will be showcased at the Abu Dhabi Festival, 2025.
- Enables searching and running open-source projects as tools or agents in sandboxed containers.
- Fine-tuned **Qwen3-32B** for two-stage **context compression**: entity selection, then a summary per passage and per entity, leaving a concise summary plus entity memory. Without reasoning, the base model often broke the required output format; the tuned model follows it **99%** of the time (validation loss 3.3 to 1.3).
- Engineered a **steerable LLM** that accepts live user instruction during generation.
- One of the idea proposers of this invite-only project for multi-platform synchronized AI memory and operation; leads the Intelligence Core team (AI and agents), which builds both the framework and the resulting agent.

Mar 2024 – Present

Railtracks / Agentic Framework with Visual Emphasis OPEN SOURCE

- An agentic framework to enhance the design, build, debug and monitor flow of AI systems, designed and engineered from scratch by my team at Railtown AI Technologies Inc, and the first and most popular Canadian agentic framework. **200+ GitHub stars**, about **1,500 downloads a month**, and **40+ contributors**, including engineers from Amazon, TD and top universities; 67 pull requests and 66 new issues in Oct 2026.
- Plain-Python design: each part (loop, tools, sub-agents, budgets, run record) is an object you can read and replace; any function becomes a tool with one decorator, and every run is recorded and replayed locally in a visualizer. Includes a Swift port.
- Profiled latency and memory overhead against LiteLLM, LangChain and the raw OpenAI SDK. Agent runs are **23% faster** (an 8-tool agent on SWE-bench with a harness), and 8 parallel agents run **41% faster** on a 2-vCPU Azure server.
- Original developer, community engagement leader and maintainer. Source: github.com/RailtownAI/railtracks

May 2025

LocateRing / Site-Selection Agent RAILTOWN AI

- An agent that takes a city map and a business proposal and ranks vacant storefronts to rent.
- Tool pool of image understanding, browsing, search and map APIs, database queries and sub-agents; evaluation loops with dynamic rubrics and comparative scoring keep each location's score comparable across candidates.
- Benchmark of 200 cases built from real shops (coffee shops, jewelry stores and more) re-listed as vacant and described as a business idea: the true site ranks in the **top 5 in 54%** of cases, against **7%** for an OpenHands agent with similar tools.

Jan 2025

ProcuRing / Public Tender Matching RAILTOWN AI

- Gathers North American public tenders and matches them against a company's inventory, for example a paper maker matched to a Nova Scotia government order for woven paper bags.
- Structure-aware RAG, rubric scoring, an agentic workflow and verification loops: **95% accuracy** on customer-labelled samples, against **71%** for GPT-4o with RAG.
- Scaled on **Google Cloud** in production; the company used it to bid on 6+ tenders. Team of 3.

Apr 2025 – Sep 2025

Advisor/Engineer / ClickCut PARTNERING WITH NUCHAIN

- Built an automated **AI video editing** tool capable of planning and producing hour-long videos through uploaded segments or AIGC. Established deployment, fine-tuning and data pipeline.
- Developed and deployed the system using a formal engineering process, which included proposal development and iterative design. Authored comprehensive design doc and engineering reports.
- Project lead in a group of 6, collaborating with top scholars around the world.

May 2024 – May 2025

Team Lead / MapChat PARTNERING WITH EAGLEEYES SEARCH

- Led a team of four in designing a **voice-interactive map** powered by LLMs for wildfire response.
- Designed a geo-data structure suitable for parallel operations by agentic systems.
- Managed workflows and timelines, drafting project proposals and design documents.
- Designed a framework for mass discrete data querying, aiding first responders in map and inventory operations.

Sep 2024 – Jan 2025

GCNN4C / Generative CNN for Classification

- Converted an unlabeled generative model (PixelCNN) into a predictive model by adding a label channel and forcing the final generation probability to map to classes.
- Explored moving from a single-class to a class-based generative model, reusing generative weights in a classification task through a new classification layer, and the impact of input transformations (rotation, flip, filter) on model performance.

Sep 2021 – Aug 2024

GPS System Designer / UBC Unmanned Aircraft Systems

- Led a four-member team in designing GPS systems for a payload autonomous rover for the AUVSI international competition.
- Developed drone and rover firmware for the AUVSI Student Competition.
- Designed the GPS module for an autonomous rover, incorporating dynamic location calibration, multi-positioning system sampling, component selection, hardware placement, and testing.

Oct 2026 – Present

NewsRing / Daily AI News Agent OPEN SOURCE

- Daily AI news agent built on Railtracks that picks only from sources it collected, published every day to the web and by email.

Sep 2026 – Present

EvalRing / Evaluation Harness OPEN SOURCE

- Ranks models, prompts or pipelines on one task, with isolated, reproducible runs. Published on PyPI as evalring.

Apr 2023 – Present

Emalia / Email Operating Agent OPEN SOURCE

- An **event-driven agentic system** for email-related operations: it lives inside a mailbox and works with any sender by email, like a personal secretary.
- Agentic RAG-based database enables intelligent mail search, with summarization across email chains and sessions.
- Remote file access and VM configuration through email. Published on PyPI as emalia, with Gmail and Microsoft mailbox sign-in.

COMPETITION

May 2022 – May 2023

ML Lead Designer / Integrated Circuit EDA Elite Challenge INTERNATIONAL

AWARD Second-Class Honor (Top 20% globally) and Special Award of Recognition (2 award per category for novel solution) (2,000$)

RANK 10th of 57 teams in the topic FIELD 3000+ Graduate Students TEAM 3

- Competed in graduate level participants to implement a Generic Electronic Device Modeling Algorithm for TCAD simulation using a modified ANN.
- Built **SimuLax**, a SoC hardware property simulator that uses a modified ANN to predict system properties such as the voltage and current of elements inside an SoC during operation.

Oct 2023 – Nov 2023

Hacker / IEEE University Hackathon UBC

AWARD Second Place (400$) FIELD 300+ Students TEAM 1

- Solving various programming question and challenges, in particular regarding algorism and systems.

Oct 2023 – Nov 2023

Hacker / RT Hacks BC

AWARD Second Place (500$) FIELD 200+ Professionals TEAM 1

- Designing AI solution and developer tools using the given tech stack of LLM and packages.

Dec 2024 – Jan 2025

Hacker / **NwHacks** BC

AWARD top 1% FIELD 700+ Students TEAM 1

- Designing Learning system to aid student in learning a new skill through LLM and Machine Learning systems.

Dec 2023 – Mar 2023

Project Team / **Intro Software Engineering Annual Competition** UBC

AWARD First Place FIELD University Wide TEAM 2

- Designing a LLM operated website that enables voice control and async data access through gesture.

OTHERS

- 2018 Junior Math Olympics top 5% (international)
- 2020 UBC Junior APSC Competition First Place (850 Participants)
- 2018 Junior Badminton third place (Province Wide)
- 2019 Green Earth Pitch Competition Top 10 (Canada)

SKILLS

LANGUAGE	Python (**PyTorch, JAX**), JS/TS, C++ (CUDA), Triton, Rust, Java, Swift, Dart, SQL (PostgreSQL, SQLite)
ML AREA	**LLM**, VLM, **evals**, RAG, hybrid retrieval (BM25, dense, RRF), context compression, multi-model system, multi-agent systems, AI security, Human-AI interactions
EVALUATION	Held-out splits, blind pairwise and A/B tests, rubric and LLM-as-judge scoring, ablations, bootstrap CIs, McNemar tests, regression evals
INFERENCE	TTFT, throughput and latency benchmarking, prefix caching, batching, cost accounting, custom CUDA and Triton kernels, torch.compile and torch.export
MLOPS	CI/CD (GitHub Actions, Azure DevOps), Docker/Kubernetes, Linux, FastAPI, async services, Git, local deployment, data engineering, data cleaning, fine-tuning
CLOUD	Azure (AzureML, Azure DevOps), Google Cloud, AWS
ML TOOLS	vLLM, ADK (LangChain, Google, RailTracks), Vector Store (Pinecone, Chroma, Faiss), Microsoft Machine Learning Server; Claude Code, Codex and Antigravity for agentic coding
COMPILERS	C compilers, a compiler for a custom FPGA CPU, program analysis (ASTs with tree-sitter, call graphs), research on execution path analysis and optimization for ML compilation (unpublished)
HARDWARE	GPU programming, Verilog, FPGA, SoC, embedded systems, firmware, AI hardware tuning, Robotics (ROS), Hardware Simulation
OTHER SOFTWARE	full-stack development (React), web applications, HCI, software DevOps, blockchain, cybersecurity, API design, Open-source project management, distributed computing
DESIGN	video production, 3D modeling, industrial design, graphics art.

RESEARCH PROFILE

DISCIPLINES	Computer Science, Applied Mathematics, Computer Engineering and Software Engineering, Electrical Engineering and Electronic Engineering

AREAS	Software Development, Computer Science and Statistics, Software (Tools), Computer Systems, Data mining, Algorithms
APPLICATIONS	Communication and Information Technologies, Education, Foundations and Knowledge Acquisition, Industrial Manufacturing and Production, Security
KEYWORDS	AI System, Artificial Intelligence, Electronic Design Automation, Human Computer Interaction, Knowledge Base, Large Language Model, Machine Learning, Software DevOps

ORGANIZED EVENTS

Sep 2025 – Nov 2025

Lead Organizer / AI x Finance: Agent-aided decision SEMINAR

- Organizing a professional seminar at the intersection of AI and finance. Attendees include professors and industry leaders. Duties include sponsor engagement, social media engagement, general event planning and leading volunteers.

Oct 2025 – Nov 2025

Lead Organizer / Railtracks x AMII hackathon WORKSHOP

- Organize and plan a Hackathon with AMII and UofA to attract professional and students to use the first Canadian agentic frameworks to build solutions. Event include topic discovery, workshop, judging session and talk from industry professionals. Duties include general planning, external contact, social media engagement and content preparation.

Nov 2025

Facilitator / Anthropic AI Hackathon WORKSHOP

- Helping to organize the event and giving talk on the important of opensource project and encourage students from UofT to engage in more project for their growth.

Oct 2025

Organizer / AI Friend: AI project showcase SEMINAR

- Invited 120+ attendees from academic and industry to discover AI project in Toronto. This repeated event include presenters from university, start-ups and individual developers.

Oct 2025

Organizer / UofT MScAC Interview prep night WORKSHOP

- Gathering 50+ graduate student from Master of Computer Science program and past alumni for interview preparation. Responsibilities include communication, event setup, documentation preparation, event proposal drafting and hosting the event.

Jan 2025 – Jul 2025

Event planning and execution / IEEE CCECE 2025 CONFERENCE

- Help plan and execute the IEEE CCECE 2025 conference at UBC. Leading volunteers, airport welcoming group and student engagement initiative. The event have 1000+ attendees and numerous workshops on a wide variety of topics.

Nov 2024

Organizer / Electrical + Computer Engineering Career Fair from IEEE UBC WORKSHOP

- Led the planning and execution of a career fair with over 500+ student registrations and participation from 20+ industry-leading companies, fostering professional connections between students and employers.
- Managed all aspects of event planning, including budgeting, logistics, vendor negotiations, and securing company participation to ensure a seamless event experience.
- Collaborated with cross-functional teams to market the event through social media, email newsletters, and physical promotions, achieving strong attendance and engagement.
- Analyzed post-event feedback and delivered actionable recommendations for improvement to inform future IEEE UBC events.

Sep 2023 – Apr 2025

Student Mentorship Leader / UBC IEEE Student Mentorship Program COURSE

- Leading Mentorship program for fellow students by providing academic support and guidance across multiple courses, fostering a collaborative and inclusive learning environment.
- Example courses include ELEC 201, CPEN 212, MATH 256

VOLUNTEER

Nov 2025

Volunteer / University of Toronto

- Volunteer at the Master of Computer Science Open house

Sep 2025

Volunteer / University of Toronto

- Volunteer at various student events at the program, including game nights, resume review session, and interview prep nights. Upto 100+ hours in total

Sep 2024 – Mar 2025

Volunteer / University of British Columbia

- Volunteer at the UBC student console review nights (attended 16 weeks)

Sep 2024

Volunteer / University of British Columbia

- Volunteer at the UBC Applied Science Career fair.

Oct 2020

Volunteer / Berton College

- Guest speaker and mentor for youth (ages 6–18) on stress management and academic tips.

Sep 2018 – Aug 2021

Member, Green Team / A.R. MacNeill Secondary School

- Promoted environmental awareness and sustainability through school-wide events, publications, and interactive demonstrations to engage and educate the student body.
- Played a key role in maintaining the school's organic farm, managing the planting, cultivation, and harvesting of seasonal vegetables to encourage eco-friendly agricultural practices.
- Collaborated with team members to organize initiatives such as waste reduction campaigns, recycling drives, and eco-friendly workshops, fostering a greener school community.
- Designed and contributed to environmental content for school publications, sharing practical tips and inspiring action towards sustainability.

May 2018

Volunteer / South Arm Community Centre

- Regularly volunteered at the community center, supporting a variety of events and programs, including roles as a summer camp counselor, translator, and seasonal event assistant.
- and supervised activities for children during summer camps, ensuring a safe, engaging, and educational environment for participants.
- Assisted in planning and executing seasonal events, such as holiday festivals and community celebrations, contributing to their success and positive outreach.
- Developed strong communication, leadership, and interpersonal skills while building meaningful connections within the community.

PUBLICATIONS

REPORT

- **Guan Zheng Huang**, Logan Underwood, Marwan Haddad, Amir Refaee, Aryan Ballani, Elliot Holtham, Jaime Bueza, Levi Varsanyi, Tristan Brown. (2025). *Railtracks Agentic Framework Public Report.* 51. Railtown AI Technologies. Co-Author, IEEE Member. Number of Contributors: 9

WORKING PAPER

- **Guan Zheng Huang**. (2025). *MapRing: A method of operating digital maps through agentic system for first responders.* First Listed Author.
- **Guan Zheng Huang**, Elliot Holtham, Logan Underwood. (2025). *LORAG: LLM Operated Retrieval Augmented Generation.* First Listed Author.
- **Guan Zheng Huang**, Yiping Liu, Hanze Li. (2025). *VSFL: A Framework for SoC Visual Schematic File Generation and Interpretation Using LLMs.* First Listed Author.

PATENT

- TICKET GENERATION SYSTEMS AND METHODS. Patent Status: Pending
- Root Cause Analysis Copilot and the Release Notes Copilot. Canada. Patent Status: Pending
- CO-PILOTS FOR ENHANCED ACCESS TO DEVOPS BASED KNOWLEDGE AND RELATED SYSTEMS AND METHODS. Patent Status: Pending