Yichen Cai

yichen.charles.cai@outlook.com · (647) 868-5037 · https://cycv5.github.io/ · linkedin.com/in/yichen-charles-cai/

‣ **Summary:** AI/Machine Learning Researcher & Engineer with MSc & BSc from University of Toronto and experience building enterprise AI (LLM) systems, completed with publications in vision and multi-modal models - seeking a role to utilize my skillset in ML and software engineering to deliver state-of-the-art AI products and solutions.

EDUCATION

University of Toronto, St. George | MSc in Applied Computing | Toronto, ON. **Sept 2024 - Dec 2025**
Master of Science in Applied Computing. Computer Science Stream. **Grade: A**

University of Toronto, St. George | Computer Science Specialist | Toronto, ON. **Sept 2019 - Jun 2024**
Honours Bachelor of Science. Specialized in Artificial intelligence (AI). **GPA: 3.99/4.0**
Awards: Dean's List Scholar 2020, 2021, 2022 and 2024. University of Toronto Scholar.

PROFESSIONAL EXPERIENCE

AI/Machine Learning Research Intern - The Vanguard Group Inc. (Toronto, ON.) **Apr 2025 - Dec 2025**
‣ Project w/ Enterprise AI & Research: Multi-agent Orchestration Framework with Agentic LLMs for Financial Advisors.
- Built a multi-agent system using OpenAI Agent SDK and in-house LLMs with MCP and A2A for communication.
- Designed agent tools, protocols and RAG systems for financial advisors with accurate, real-time information.
- Benchmarked LLMs vs. SLMs with fine-tuning, and explored hierarchical combinations of different model sizes.
- Contributed to agentic AI orchestration research and internal documentation with Vanguard's Enterprise AI team.

Software Engineer - Zebra Technologies (Mississauga, ON.) **May 2022 - Jul 2023**
‣ Zebra Android System Applications - Main Developer:
- Created Zebra Antenna setting app, enabling 2 wireless communication modes with Zebra forklift docks.
- Developed the heater control app that regulates device temperatures, making devices suitable for frozen warehouses.
- Collaborated with the PM and hardware team (AGILE) to use the GPIO drivers to control the antenna and heaters.
- Expanded the applicability of Zebra devices under more scenarios, broadening Zebra's target customer base.

PROJECTS & PUBLICATIONS

Side-Channel Attack on Keyboards with LLM Corrections, USENIX WOOT'25 **Sept 2024 - Mar 2025**
- Developed a multi-modal model to predict keyboard presses from varying distances with overall 98% accuracy.
- Designed a new pipeline using STFT to segment audio and video of the keyboard presses dynamically.
- Fine-tuned Llama 3B model for context-aware error corrections with 90%+ the performance of GPT-4o.
- Co-authored the paper with the core results, published in USENIX WOOT'25 Conference proceedings.

Reward-Driven Role Allocation with Multi-Agent Reinforcement Learning **Sept 2023 - Jan 2024**
- Designed a coordinator-executor model with Q-learning to obtain an optimized policy in a sequential social dilemma.
- RL agents matched the performance of the best human-coded agents, demonstrating its effectiveness.

NVIDIA AICity Challenge: Retail Product Recognition and Counting, CVPR 2023 **Dec 2022 - Jun 2023**
- Created a project plan, then designed and implemented a YOLO + StrongSORT backbone algorithm.
- Fine-tuned the model to recognize 116 distinct products from retail stores with an F1-score of 0.82.
- Achieved top-3 finish at CVPR 2023 NVIDIA AICity Challenge, published in CVPR 2023 proceedings.

SKILL SETS

- **Programming languages:** Python (w/ PyTorch), C/C++, Java, SQL, Shell, Racket.
- **Tools:** Git, AWS, CI/CD tools, GCP, JetBrains.