

Wyett “Huaye” Zeng

wyettzeng@g.harvard.edu | 519-729-8107 | wyett-zeng.com | [LinkedIn](#) | [GitHub](#) | [Google Scholar](#)

Summary of Qualifications

- Languages: Java, Python, C++, C, C#, SQL, Go, Bash, JavaScript, HTML/CSS
- Tools: gRPC, Protobuf, GraphQL, AWS, Azure, PyTorch, LLamaIndex, LangChain, Pandas, NumPy, Slurm, Clickhouse, Docker
- Relevant Courses (Grade out of 100): Object Oriented Programming (95), Data Structure (99), Algorithms (99), OS (96), Machine Learning Statistics (97), Intro to AI (100), Reinforcement Learning (91).
- **200+ citations** for machine learning publications in top venues (ACL & COLM), including first-author papers [[Google Scholar](#)].

Education

Harvard University

Master of Science in Data Science

Cambridge, Massachusetts, USA

Sep 2025 – Dec 2026 (Expected)

University of Waterloo & Wilfrid Laurier University

Bachelor of Computer Science and Bachelor of Business Administration Double Degree

Laurier Alumni Gold Medalist (Major Average: 94.79/100)

Waterloo, Ontario, Canada

Sep 2020 – Apr 2025

Experiences

Software Developer (AI/ML) | GPTZero

Sep 2024 – Aug 2025 | Toronto, Canada

- Migrated the writing feedback system from Prompt Flow and **Flask** to LLamaIndex and **Quart**, enhancing scalability and achieving **30%** reduction in processing time on internal API endpoints.
- Led the AI-Grader product using transformer-based architectures with **PyTorch**. On IELTS, a widely used standardized English tests, the model achieves **88% accuracy** within ± 1.0 band (out of 12) and **97%** within ± 1.5 bands, rivaling human graders.
- Spearheaded the writing coach product, implemented a new **AWS Lambda** function to parse large user documents and used RAG techniques with LLamaIndex to generate feedback that aligns with the provided material, resulted in over **30%** increase in user screen time and **15%** increase in satisfaction rating.

Software Developer | Boosted.ai

Jan 2024 – Apr 2024 | Toronto, Canada

- Rewrite the factor model algorithm which reduces 5,000+ customer models' scheduled inference time by over **90%**, with **weekly 500+ hours** less computation time on **AWS EC2**. The algorithm uses **NumPy**, **Clickhouse**, and **PostgreSQL** to efficiently compute economic factor values for every publicly listed securities (20,000+) each day.
- Developed the investment style matching feature facing 150+ institutional clients using **GraphQL**, **gRPC**, and **Protobuf**. The feature analyzes client's portfolios and reports the fitness of their selected investment style.
- Developed and deployed enhancements for Boosted.ai's trading algorithm spanning analytics server, database retrieval, and backend server using **Python**, **Java**, **gRPC**, and **Protobuf**, increasing client satisfaction by over **6%**.
- Developed AI commentary features leveraging **LangChain** for prompting LLMs to deliver tailored portfolio analysis on macroeconomic topics for 150+ institutional clients.

Data Scientist | Canadian Imperial Bank of Commerce

Jan 2023 – Apr 2023 | Toronto, Canada

- Set up and deployed a **PostgreSQL** server on **AWS** to centralize financial data storage for alternative investments, significantly accelerating team members' data retrieval and synchronization processes.
- Developed a client report generation application using **PyQt5**, integrating data from multiple external partners through **REST API** calls, reducing monthly processing time from **25 to 6 hours**.
- Developed the market analysis report that presents hundreds of market trend graphs to team members and make short-term future predictions using various methods such as GRU and LSTM using **TensorFlow**.

Software Engineer | Siemens Healthineers

Jan 2022 – Apr 2022 | Ottawa, Canada

- Developed and maintained the Android-based NXS application using **Java** and **Android Studio**, strictly adhering to **object-oriented design** patterns such as MVC and Singleton to ensure modularity, maintainability, and reliable hardware interfacing.
- Fixed the Glucose conversion issue and improved its performance by **70%**. This issue arose from incorrect asynchronous saving logic with **Realm database** and **Reactive Java**.

Publications

- **(First Author, ACL 2025)** AceCoder: Acing Coder RL via Automated Test-Case Synthesis [[paper](#)][[website](#)][[huggingface](#)].
- **(COLM 2025)** ScholarCopilot: Training LLMs for Academic Writing with Accurate Citations [[paper](#)][[website](#)][[demo](#)].
- **(TMLR 2024)** MANTIS: Interleaved Multi-Image Instruction Tuning [[paper](#)][[website](#)][[demo](#)].