

Wyett “Huaye” Zeng

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

Skills

- Languages & Tools: Python, C++, C, Java, SQL, Bash, Pytorch, Tensorflow, DeepSpeed, vLLM, SGLang, LoRA, Scikit-Learn, Pandas, NumPy, Slurm, Clickhouse, LlamaIndex, Prompt Flow, LangChain
- Research Topics: NLP, Transformers, Computer Vision, Reinforcement Learning, AGI, Semantic Trading, Reward Modelling
- **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
Aug 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

Selected Publications

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

Work Experiences

Machine Learning Developer | GPTZero

Sep 2024 – Aug 2025 | Toronto, Canada

- 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, resulted in over **30% increase** in user screen time.
- Optimized the grammatical error correction model, reducing edit distance by **50%** and improving GLEU score from 0.7 to **0.8**.
- 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.

Quantitative Developer | Boosted.ai

Jan 2024 – Apr 2024 | Toronto, Canada

- Rewrote 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 **4 trillion+** economic factor values for every publicly listed security each day.
- 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

- 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**.
- Developed the quantitative portfolio builder, which can construct a portfolio whose return is within **$\pm 2.8\%$ of the benchmark return** using QSolver to provide insight into the more “obscure” alternative investment hedge funds.

Research Experiences

AceCoder | Tiger Lab | University of Waterloo

Nov 2023 – Feb 2025

- Developed a fully automated pipeline for large-scale synthesis of (question, test-case) pairs using **vLLM** and **SGLang**, enabling efficient inference generation with Qwen Coder 2.5 7B to create both preference and inference-accuracy datasets.
- Trained reward models using **distributed training** in **PyTorch** with Bradley-Terry (BT) loss, and conducted reinforcement learning via **PPO** and **REINFORCE++**, scaling experiments across **16 NVIDIA A100 GPUs**.
- The finetuned reward model showed an average of **10%** improvement for Llama-3.1-8B-Ins and **5%** for Qwen2.5-Coder-7B-Ins through best-of-32 sampling across benchmarks, **making the 7B model on par with 236B DeepSeek-V2.5**.

MANTIS | Tiger Lab | University of Waterloo

Nov 2023 – Nov 2024

- Investigated the multi-image reasoning capabilities of LLMs by interleaving image tokens from **vision encoders** such as CLIP and SigLIP with textual instructions, enhancing tasks such as co-reference, comparison, and temporal understanding.
- Fine-tuned the Fuyu model on Mantis-Instruct, achieving a **13%** performance improvement over the **SoTA baseline**, Idefics2-8B.

Transformer Trader | Wilfrid Laurier University

Sep 2023 - Present

- Completed an Honours Thesis on training LLMs to predict market sentiment from economic news by fine-tuning custom LLMs.
- Combined the Llama base model with custom heads using **PyTorch**, **DeepSpeed**, and **QLoRA**, then transformed the model to a Transformer-based trading algorithm; back tested trade shows **over 15% IRR**.
- Studied financial news' topic compositions over time using methods such as Latent Dirichlet Allocation and Gaussian Mixture Model.