

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, Slurm, Docker
- 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](#)].

Selected Publications

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

Work Experiences

Machine Learning Engineer Intern | GPTZero

Sep 2024 – Aug 2025 | Toronto, Canada

- Spearheaded the writing coach product, implemented a new **AWS Lambda** function to parse large user documents and used RAG with LlamaIndex to generate relevant feedback, increased **user screen time by 30%** and satisfaction rating by **15%**.
- 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.
- 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**, redesigning API routing and introducing asynchronous request handling, enhanced scalability and reduced processing time on internal API endpoints by **30%**.

Quantitative Developer Intern | Boosted.ai

Jan 2024 – Apr 2024 | Toronto, Canada

- Rewrote the core factor model algorithm, significantly reducing 5,000+ customer models' scheduled inference time by over **90%**, resulting in weekly savings of **500+ hours** of computation time on **AWS EC2**. The algorithm leverages **NumPy**, **ClickHouse**, and **PostgreSQL** to efficiently compute and update daily economic factor values for **20,000+** publicly listed securities.
- Developed new AI commentary features leveraging **LangChain** for prompting LLMs to deliver tailored portfolio analysis and insights on macroeconomic topics for 150+ institutional clients worldwide, achieving over **85%** user adoption within two weeks.

Data Scientist Intern | Canadian Imperial Bank of Commerce

Jan 2023 – Apr 2023 | Toronto, Canada

- Developed a market analysis program that integrates streaming data from Morningstar APIs, delivers short-term forecasts using GRU and LSTM models built with **TensorFlow**, and generates 300+ interactive market trend graphs for team members in minutes.
- Developed the quantitative portfolio builder, which formulates and solves an optimization problem using **QSolver** to construct a portfolio whose return is within $\pm 2.8\%$ of the benchmark, uncovering insights into “obscure” alternative investment hedge funds.

Research Experiences

AceCoder (ACL 2025) | Tiger Lab

University of Waterloo

- 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 **DeepSpeed ZeRO Stage 3** and **LlamaFactory** with Bradley-Terry (BT) loss, and trained reinforcement learning models via **PPO** and **REINFORCE++**, scaling experiments across **8 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 (TMLR 2024) | Tiger Lab

University of Waterloo

- 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 (In Progress)

Wilfrid Laurier University

- Completed an Honours Thesis on training LLMs to predict market sentiment from economic news by fine-tuning custom LLMs.
- Finetuned LLaMA with custom classification and regression heads using **PyTorch**, **DeepSpeed**, and **QLoRA**, then transformed the architecture into a Transformer-based trading algorithm, back testing on historical market data shows over **15% IRR**.
- Studied financial news' topic compositions using methods such as Latent Dirichlet Allocation and Gaussian Mixture Model.

Education

Harvard University

Master of Science in Data Science

Cambridge, Massachusetts, USA

December 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

April 2025