Wyett "Huaye" Zeng

wyettzeng@g.harvard.edu | 519-729-8107 | wyett-zeng.com | LinkedIn | GitHub | Google Scholar

Highlights

- **Open-Source Contributions**: Co-lead of verl-Tinker, contributed RL, VLM, and LoRA functionalities to verl, veomni, and vLLM.
- **500+ citations** for machine learning publications in top venues, including first-author and best paper awards [Google Scholar].
- Languages: Python, Java, C++, C, Cuda, SQL, Go, Bash, JavaScript, HTML/CSS.
- Tools: gRPC, Protobuf, Triton, GraphQL, PyTorch, LLamaIndex, Pandas, NumPy, Slurm, Clickhouse, Docker, Kubernetes.
- Relevant Courses: Object Oriented Programming (95), Data Structure (99), Algorithms (99), OS (96), ML Statistics (97), Intro to AI (100), Reinforcement Learning (91), ML Dev Operations (A+), Unsupervised Learning (A+), Computing at Scale (In Progress).

Experiences

Machine Learning Engineer | Bytedance / TikTok Jun 2026 – Aug 2026 | San Jose, USA

- Optimized Raskel, an internal multi-LoRA training framework, tuning batching and parallelization strategies to improve **MFU by 5%** and optimizing its **Triton multi-LoRA GEMM kernel** to increase H20 throughput by up to **20%** across different workloads.
- Built **verl-Tinker**, an open-source Tinker-compatible API server built on verl, enabling fine-tuning with VeOmni/FSDP actors, vLLM/VeXact rollout, reference engines, and dynamic model offloading for RL training on **120B models** across 64 GPUs [link].
- Implemented importance-sampling and DRO losses in verl; added streaming synchronization for GPT-OSS MoE weights, whose gate and up projections are interleaved across experts, **reducing peak memory** for the affected block to **1%–3% of the original**.
- Added LoRA support for **VLM** training in VeOmni, along with MFU calculation logic for performance analysis and optimization.

Machine Learning Engineer | GPTZero (acq. by Superhuman / Grammarly) Sep 2024 – Aug 2025 | Toronto, Canada

- Built the end-to-end **evaluation** and **data pipeline** for **AI Reviewer**, including metric design, iterative dataset curation and annotation, and scalable third-party labeling workflows; led development of the transformer-based reviewer in PyTorch, achieving **88% accuracy** within ±1.0 IELTS band and **97%** within ±1.5 bands, approaching human-grader agreement on standardized scoring.
- 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%.**
- 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.

Quantitative Developer | Gradient Boosted Investment Inc 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 the investment style matching feature for 150+ institutional clients using **GraphQL, gRPC,** and **Protobuf**. The feature automatically analyzes client portfolios and reports the fitness between the portfolios and their selected investment style.
- Developed and deployed multiple enhancements for the signature Boosted.ai trading algorithm spanning analytics server, database retrieval, and backend server using **Python, Java, gRPC**, and **Protobuf**, increasing client satisfaction scores by over 6%.

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, enabling parallel query execution and caching that accelerated data retrieval by **3×** and improved synchronization throughput by **45%**.
- Developed a market analysis program that integrates streaming data from Morningstar APIs, generates **300+** interactive market trend graphs for team members in minutes, and delivers short-term forecasts using GRU and LSTM models built with **TensorFlow**.

Research Experiences

Text Representation and the Information Content of Financial News (Preprint) **Wilfrid Laurier University**

- Used **unsupervised topic clustering** to isolate **asset-class-specific news signals** from large financial-news corpora.
- Built a high-frequency news-return research pipeline matching **135K+** Dow Jones articles to millisecond **NBBO** mid-quote data across five asset-class ETFs, computing normalized post-news returns over different intervals to measure market reactions.
- Compared **keyword, LDA, and LLM embedding** topic models to quantify return-relevant information in financial news, using **PCA/GMM** clustering, permutation tests, Benjamini–Hochberg correction, mutual information, and multivariate logistic regression.
- Found that richer text representations recover more **market-moving information**, with topic-return associations rising from **31.4%** with keywords to **54.1%** with LDA and **67.8%** with LLM embeddings, while **price discovery** peak at the 60-second horizon.

Education

Harvard University Cambridge, Massachusetts, USA
Master of Science in Data Science (GPA: 3.96/4.0) (Expected) December 2026

University of Waterloo & Wilfrid Laurier University Waterloo, Ontario, Canada
Bachelor of Computer Science and Bachelor of Business Administration Double Degree April 2025
Laurier Alumni Gold Medalist (Major Average: 94.79/100)