YIFEI SUN

+1 437-347-0631 | idris.yifeisun@gmail.com | www.linkedin.com/in/yifei-idris-sun

EDUCATION

Bachelor of Science: Computer Science / Mathematics 05/2025

University of Toronto — Toronto, ON

- GPA: 4.0/4.0; Avg 94.1/100
- Arts & Science Internship Program (Co-op): 16 months
- Relevant coursework: Machine learning, Deep learning, Probabilistic machine learning, Computational linguistic

TECHNICAL SKILLS

- **Languages:** Python, C/C++, MATLAB, SQL, {Javascript, HTML, CSS}, Java, Git, Bash
- **Frameworks:** PyTorch, TensorFlow, HuggingFace, MLIR, AutoGPTQ, scikit-learn

PUBLICATIONS

- Zongliang Ji*, **Yifei Sun***, Andre Carlos Kajdacsy-Balla Amaral, Anna Goldenberg, Rahul G. Krishnan. **Can we generate portable representations for clinical time series data using LLMs?** (under review ICLR 2026, rating 8,8,6,4, top 5%)
- **Yifei Sun***, Zhenwei Tang*, Yuran Zhang, Rudraksh Monga, Ashton Anderson. **JTS: Joint Textual and Structural Logical Query Answering Over Knowledge Graphs**. (2024)
- Kosei Uemura, Mahe Chen, Alex Pejovic, Chika Maduabuchi, **Yifei Sun**, En-Shiun Annie Lee **AfriInstruct: Instruction Tuning of African Languages for Diverse Tasks** (EMNLP Findings 2024)

EXPERIENCE

Research Assistant 05/2025 — Current

Vector Institute / University of Toronto — Toronto, Canada (supervised by Prof. Sheila McIlraith under UTEA)

- Designed *First-Order Reward Machines (FORM)* with 6+ formal components (typed quantification, per-object progress vectors, control/object-level product automata) to generalize over arbitrary object sets. **Achieved 6-7x improvement in sampling efficiency** in policy transfer across generalization tasks.
- Bridged *FO-LTL → FORM*: developed a compilation procedure from first-order LTL task specs to executable reward machines
- Extended RM expressivity with *context-free* guards to capture counting and stack-like objectives beyond regular RMs; increased task coverage by **+32%** and reduced automaton size by **90%** for equivalent specs, with no more than **8%** runtime overhead.
- Implemented a Python library integrating FORM with RL environments (Gym-style) and PDDL task generators

Undergraduate Research Student 09/2024 — Current

Vector Institute / University of Toronto — Toronto, Canada (supervised by Prof. Rahul G. Krishnan)

- Evaluated distribution shift in ICU time series across sites; achieved **+19%** in-distribution improvement over simple imputation baselines and **+30–120%** gains under cross-site few-shot settings.
- Built a scalable summarization→embedding (StE) pipeline using **vLLM** and **Hugging Face Transformers** (batched decoding, caching, rate-limit handling) for large-volume clinical data.
- Audited **fairness & privacy**: maintained comparable or lower age/gender leakage vs. baselines; documented privacy risks and mitigation options for downstream use.
- Led data engineering/EDA for ICU datasets (**EHRShot**, **MIMIC-IV**, **HiRID**, private)

Undergraduate Research Student 05/2024 — 10/2024

University of Toronto — Toronto, Canada (supervised by Prof. Ashton Anderson)

- Built joint textual+structural LLM models for knowledge-graph completion and logical QA; reached new SOTA on internal benchmarks with **+7.6 pp Hits@1** and **+9.3% MRR** over strong baselines (3 datasets, 3 seeds).
- Algorithmically optimized inference: cut per-query latency from **14 minutes to 13 seconds** (**65**× speedup) and peak GPU memory by **3.1**× via caching, pruning, and quantization.
- Delivered reproducible pipelines (training/eval scripts, ablations, unit tests); **200**+ experiments across **5** datasets logged with automatic sweep tracking.

Research Associate 05/2024 — 09/2024

University of Toronto — Toronto, Canada (supervised by Prof. David Liu, Alice Gao, Naomi Levy-Strumpf)

- Developed **8** modular units on data science and ML (decision trees, PCA, heatmaps, model evaluation, etc.) with **24** hands-on labs and a **48**+ item exercise bank.
- Produced instructor guides and auto-graded notebooks; reduced facilitation time by **30%** and Git-versioned the curriculum for semester reuse.
- Work shortlisted for the **QS Reimagine Education Awards** (top **7%** of **1300**+ submissions).

Machine Learning Stack Engineer 05/2023 — 04/2024

Cerebras Systems — Toronto, ON

- Built and maintained compiler/runner pipelines for gradient accumulation and layout optimization; increased effective batch size on constrained memory by **2.3**× and improved tokens-per-second throughput by **28%**.
- Co-developed estimator models for compiler pass costs and kernel layouts; achieved <**5%** median error on latency/memory predictions and reduced iteration cycles per optimization by **31%**.
- Designed parameter-sweep infrastructure (lane count, batch size, activation checkpointing, etc.); automated weekly test suites covering **2,000**+ configs and surfaced **70**+ issues, cutting regression detection time by **40%**.
- Integrated results into CI with artifact tracking and dashboards; reduced triage time from >**1 day** to <**2 hours** for performance regressions.

AWARDS

- University of Toronto Excellence Awards ($7500 + $500)
- Summer Undergraduate Data Science ($7200, declined)
- Dr. James A. & Connie P. Dickson Scholarship In Science & Mathematics (2×)
- Fay And David Masson Scholarship
- Howard Ferguson Scholarship (3×$3000)
- Reuben Wells Leonard Scholarship
- University College Alumni Scholarship & Bursary Fund

VOLUNTEER

Undergraduate Student Research Program — Mentor 09/2023 — 04/2025