Yuzhi (Paul) Tang

Toronto, ON yuzhi.tang@mail.utoronto.ca +1 778-798-8289 /yuzhi-tang /paulslss300

Education

University of Toronto *Sept 2024 – Dec 2025*
Master of Science in Applied Computing - AI Concentration

- **Coursework:** Deep Learning Theory & Data Science, Large Models, LLMs & GPTs for Ubiquitous Computing, Geometric Deep Learning

University of Toronto *Sept 2020 – Jun 2024*
HBS (High Distinction), Specialist in Computer Science (Focus in Artificial Intelligence), Major in Cognitive Science, Minor in Mathematics

- cGPA: 3.92/4.0
- **Coursework:** Neural Networks and Deep Learning (A+), Intro Machine Learning (A+), Probability and Statistics (A+), Intro Artificial Intelligence (A+), Intro Image Understanding (A+), Algorithm Design & Analysis (A), Software Design (A+), Intro Databases (A)

Experience

LLM Researcher *Toronto, ON*
University of Toronto - Machine Learning Group - Prof. Chris Maddison *Jun 2024 – Present*

- Developed an **automated LLM agent evaluation framework** utilizing LLMs for **tool emulation and risk evaluation**, with support across **328+** agent scenarios and **36+** high-stakes tools
- Developed LLM agent workflows and designed probing tests for decomposing and benchmarking LLM agent tool-use risks at the **knowledge, verification, and generation levels**, uncovering **significant gaps in risk awareness** of SOTA LLM agents (e.g. GPT-4o, Claude-3.5, Llama-3.1), with a pass rate of 20%
- Developed a **reflection agent framework** using scenario extraction and iterative self-refinement to **significantly reduce risky tool-use behaviour** by over 60%

Technologies: Python, PyTorch, Langchain, vLLM, HuggingFace, Seaborn, Slurm

Student Researcher in SE4AI *Toronto, ON*
University of Toronto - Software Engineering Group - Prof. Marsha Chechik *May 2023 - May 2024*

- Designed and implemented a framework utilizing **neuron activation patterns (NAPs)** to assess and improve the reliability of deep neural networks (DNNs) against distributional changes
- Developed coverage metrics for diverse NAPs to drive **coverage-guided fuzz testing**, enabling interpretable evaluations of DNN performance and uncovering hard-to-find DNN defects
- Evaluated the framework on **MNIST, CIFAR10, SVHN**, and **ImageNet** datasets, achieving accurate and interpretable results with fewer tests than existing baselines
- Research supported by the **NSERC USRA award ($7500)**

Technologies: Python, PyTorch, Tensorflow, Torchvision, NumPy, Matplotlib, Slurm

ML Engineer *Toronto, ON*
Sunnybrook Research Institute *Jan 2023 - Apr 2023*

- Developed a machine learning pipeline to classify sleep stages (e.g. REM, Non-REM, Wake, etc.) from **multi-modal biometric data** collected from the ANNE One wearable sensors
- Implemented and trained a **Convolutional Recurrent Neural Network (CRNN)** architecture and experimented with different architecture modules including **CNN, RNN, GRU, LSTM, and Transformer**
- Addressed data imbalance with **weighted CE loss**, developed **auxiliary training objectives** and a **decision tree ensembler** to improve classification F1 by **11% (0.72 macro-F1)**. Paper accepted to SLEEP2024 (top-tier)

Technologies: Python, PyTorch, Scikit-learn, NumPy, Matplotlib, ONNX, Slurm

Publications

Understanding and Mitigating Risk Causes in Large Language Model Agents — In Preparation for ICML2025

Y Tang, T Li, E Li, Y Ruan, H Dong, C Maddison

DeFeaT: Feature-based Reliability Testing of Deep Neural Networks through Feature-specific Neurons — In Preparation

Y Tang, C Hu

Mamba-based Deep Learning Approaches for Sleep Staging on a Wireless Multimodal Wearable System without Electroencephalography — In Submission for SLEEP2025

A Zhang*, A He-Mo*, R Yin*, C Li, ***Y Tang***, D Gurve, N Ghahjaverestan, M Goubran, B Wang, A Lim

https://doi.org/10.48550/arXiv.2412.15947

A Deep Learning Model for Inferring Sleep Stage from a Flexible Wireless Dual Sensor Wearable System Without EEG — SLEEP2024

A Zhang, C Li, ***Y Tang***, A He-Mo, N Ghahjaverestan, M Goubran, A Lim

https://doi.org/10.1093/sleep/zsae067.01122

Asynchronous Detection of Erroneous Behaviors in Human-Robot Interaction with EEG: A Comparative Analysis of Machine Learning Models — Oct 2023

Z Ren*, X Xia* ***Y Tang****, B Zhao, C Wong, D Xiao

https://doi.org/10.1101/2023.09.30.560271

Projects

Intrinsic Error Evaluation during Human-Robot Interaction Competition [IJCAI2023] - Winner — *IJCAI2023*

- Built a soft voting **ensemble model** combining **MLP, SVM, XGBoost, and Random Forest** to detect anomalies in EEG signals. Leveraged **SMOTE** to address data imbalance. **Achieved over 23% lower cumulative error than the 2nd-place team**

From Frustration to Personalization: Revolutionizing Course Recommendations with LLMs — *Report*

- Developed a **Retrieval-Augmented Generation (RAG)** pipeline using **Faiss** for efficient vector search and **MongoDB** for structured storage, enabling accurate retrieval of relevant courses from natural language queries of student needs. Implemented a **reranker** which improved retrieved course relevancy **by 21%**

Text-Conditioned 3D Object Generation with Latent Diffusion Prior in NeRF — *Report*

- Developed a **text-conditioned Diffusion-NeRF** architecture for 3D object generation using **CLIP** as the text encoder and a **U-Net** diffusion model. Achieved **19%** and **25%** improvements in FID and KID respectively over SOTA baseline

Self-Supervised Pretraining For Improving Segmentation Performance of Ultrasound Medical Images — *Poster*

- Adapted **dense contrastive self-supervised pertaining recipe** to a **U-Net** for ultrasound knee effusion segmentation. Achieved **3.9 mIoU improvement (79.1 mIoU)** and **10x faster training convergence**. Poster presented at the **Undergraduate Research Conference 2022** at the University of Toronto. Supervised by Prof Pascal Tyrrell

Skills

Programming Languages: Python, SQL, C, Flask, Java, C++, Swift, R, MatLab

Machine Learning Libraries: NumPy, PyTorch, TensorFlow, Scikit-learn, Seaborn, LangChain, vLLM, HuggingFace, Weights & Biases, OpenCV, MMSegmentation, ONNX

Tools: MySQL, XCode, Git, Conda, SSH, Slurm, Linux