




Dav Vrat Chadha

✉ davvrat.chadha@mail.utoronto.ca  in/davvratchadha  github.com/davvratchadha  davvratchadha.com

Education

B.A.Sc. in Engineering Science, University of Toronto

Major: Machine Intelligence

Toronto, ON, Canada

Sep 2020 – Apr 2025

Skills & Tools

• Python • Tensorflow • PyTorch • NumPy • NLP • AI • ML • Sklearn • Keras • CUDA • Jax • Objax • spaCy • Cython • C/C++ • CI/CD • Git • MATLAB • Linux • Java • MySQL

Professional Experience

Teaching Assistant - CSC401/2511 Natural Language Computing

University of Toronto

Toronto, ON, Canada

Sept 2024 - Present

- Responsible for grading assignments, quizzes, and exams, ensuring consistency and fairness in evaluation.
- Collaborating with the Head Teaching Assistant and course instructor to uphold grading standards and policies.

Research Student - Computational Linguistics Lab

University of Toronto

Toronto, ON, Canada

Sept 2024 - Present

- Part of Prof. Gerald Penn's Computational Linguistic lab, working on a thesis project focusing on mechanistic interpretability of large language models using circuit discovery.

Software Engineer Intern, Memory Team

AMD - Datacenter GPU

Markham, ON, Canada

May 2023 - Aug 2024

- Developed a memory validation and debug tool, SuperScript, as a **Python** library with CLI capabilities, significantly improving data integrity and reducing debugging time. Enhanced user efficiency and productivity, has been used over **100,000 times**, and is now a dependency for several other AMD tools.
- Engineered a 4pt HBM-PHY margin testing tool for filtering bad MI325 parts **during manufacturing**, aiming to reduce customer RMAs.
- Designed and created a memory characterization tool, Char_wizard, reducing characterization time by **70%** (or 650 hrs) with **innovative grid search algorithms** and **decision trees**. Improved runtime complexity from $O(n^2)$ to $O(n)$. Incorporated **concurrency** to run characterization on multiple processor dies simultaneously. The tool has been used for over **2,000 hours**.
- Designed and developed a **Python** tool, amd-address-mapper, to translate physical addresses from ROCm workloads to failing locations in HBM, currently used internally and at customer sites.
- Contributed to AMD **automation frameworks** and infrastructure. Developed memory automation tests for all MI300 platforms.
- Contributed to HBM3E bring-up for MI325 product offering, providing memory validation task execution, automation, and validation tools support. Additionally, contributed to the successful HBM3 debug process during the El Capitan supercomputer bring-up phase.
- Implemented a scalable Bit Error Rate testing tool for the HBM-PHY interface on all MI300 platforms, enabling efficient data collection at scale.
- Developed a tool to simplify patching custom components over Integrated Firmware Images (IFWIs) during validation.

ML Engineer - FINCH Satellite Mission

University of Toronto Aerospace Team

Toronto, ON, Canada

Sept 2023 - Aug 2024

- Implemented a novel **diffusion model** conditioned on neighboring spectral frames for **destriping** hyperspectral images resulting in PSNR = 39.2274, LPIPS = 0.2214, SSIM = 0.8817, and SAM = 0.0423, for the ICVL-HSI dataset.
- Worked on the hyperspectral data augmentation pipeline in **PyTorch**, utilizing a modified image patch extraction technique to create new images out of existing ones to increase the size of the training and validation dataset.
- Operated in an **agile** environment, contributing to continuous improvement and innovation within the team.

Publications

- Dav Vrat Chadha et al., *Optimizing Shmooing for AI HPC HBM Memory Characterization Using Decision Trees*, AMD Internal Publication, July 2024. Reviewed by AMD Fellows Committee.
- Ian Vyse et al., *Beyond the Visible: Jointly Attending to Spectral and Spatial Dimensions with HSI-Diffusion for the FINCH Spacecraft*, Presented in 38th Annual Small Satellite Conference, 2024. DOI: 10.48550/arXiv.2406.10724

Projects

ML Engineer - NoPunIntended

Toronto, ON, Canada

Repository; Try API

Jan 2023 - Apr 2023

- Utilized an ensemble of **transformer**-based **LLMs DeBERTa** and **RoBERTa** to detect and locate puns with contextual masking using K-means.
- Built upon research done by Amazon to improve the existing methods and achieved **75.58%** test accuracy, which is competitive to GPT-4 performance (82.77%).

Honors & Awards

- Two-time AMD Executive Spotlight Award winner - Aug 2024, Dec 2023
- AMD Intern Innovation Showcase Award - Mar 2024
- Two-time AMD Spotlight Award winner - Dec 2023, Aug 2023
- Dean's List, University of Toronto - Dec 2020