Inesh Chakrabarti

858-925-3059 | inesh33@g.ucla.edu | linkedin.com/in/inesh-chakrabarti | github.com/beesfleas

Education

University of California, Los Angeles Los Angeles, CA

B.S. Electrical Engineering, M.S. Electrical Engineering *Expected Graduation – Jan 2027*

- **Graduate GPA:** 4.0 / 4.0, **Undergraduate Major GPA:** 3.8 / 4.0
- **Coursework:** Large Scale Data Mining, Convex Optimization, Dynamic Feedback Control, Deep Learning, Software Engineering, Embedded Systems, Computer Architecture, GPU Microarchitectures, Numerical Computing, Stochastic Systems, Communications, Signals and Systems, Probability and Statistics
- **Leadership:** American Nuclear Society (President, Founder), Eta Kappa Nu (Mentorship Chair)

Skills

- **Programming Languages:** C, C++, Python (NUMBA, PySpark, Matplotlib, PyTorch, Pandas, Tensorflow), CUDA, PTX, Triton, System Verilog, SQL, x64, C#, Java, MATLAB, R, JavaScript
- **Tools:** Docker, Git, LangGraph, MongoDB, LTSpice, GDB, Unix Shell, OpenMP, Joblib, Django, NVIDIA Nsight Compute, Apache Spark, Fuzzing (AFL), CI/CD

Experience

UCLA Lin Yang Research Group (AI Researcher) *Febuary 2025 - Present*

- *NoWag: A Unified Framework for Shape Preserving Compression of Large Language Models* Lawrence Liu, **Inesh Chakrabarti**, Yixiao Li, Mengdi Wang, Tuo Zhao, Lin F. Yang Publication accepted to **COLM** and **ICLR SLLM Workshop**
- Built dequantization/inference kernels in C (CUDA) for parallelization over multiple GPUs for over 10x speedup while using **48x less calibration data** and matching SOTA VQ method performance
- Implemented Trellis Quantization and benchmarking in Python for NoWag, a set of shape-preserving pruning and quantization algorithms for LLMs

UCLA Complex Networks Group (Paid Machine Learning Researcher) *February 2022 - June 2024*

- Implemented High Frequency Oscillation Detector using Variational Autoencoder for neural signals, **doubling** number of detections with only a **10%** increase in false positives
- Constructed a speech to text pipeline that subtitled recall experiments with precise temporal acc.
- Processed and visualized neural spike data using Python and MATLAB to demonstrate correlation between individual neural spikes and character recognition from animation
- Developed a complete pipeline for EEG data analysis with wavelet transform pre-processing to predict human movement using transformer, LSTM, and CNN models.

Projects

Reinforcement Learning Hearts *September 2024 - January 2025*

- Created RL agent for Hearts using **Counterfactual Regret Minimization** and **Monte Carlo Tree Search** that reaches approximate Nash Equilibrium.
- Enhanced the Hearts project with a Tkinter UI and collaborated in a 3-person team, providing a **real-time interface** allowing for physical gameplay simulation via computer vision.

Vortex GPGPU Dynamic Kernel Scheduler *September 2025 - November 2025*

- Enhanced Vortex GPGPU's Kernel Management Unit by implementing dynamic kernel scheduling, enabling parent kernels to launch child kernels on demand
- Developed and verified the dynamic scheduling functionality in both Verilog and C++ simulation, improving the GPGPU's flexibility for complex workloads
- Implemented and optimized foundation GPGPU kernels within the Vortex framework eg. MatMul

Database Benchmarking Tool *September 2025 - December 2025*

- Engineered a **novel benchmarking tool** by translating TPC-DS SQL queries into **PySpark** via Abstract Syntax Tree (AST) manipulation and injecting realistic User-Defined Functions (UDFs).
- Scraped and analyzed public PySpark workflows from GitHub to create a representative distribution of modern data pipelines, guiding the synthesis of UDFs based on metrics like cyclomatic complexity.