Krish Agarwal

wonderingkrish@gmail.com
(669) 245-7474

Cupertino, CA
krishagarwal.github.io
www.linkedin.com/in/krishagarwal

EDUCATION

University of Texas at Austin — Aug 2022 - Present

- BS Computer Science, Turing Honors Scholar, Senior (GPA: 4.0)
- Relevant Courses: AI, Principles of ML, Autonomous Intelligent Robotics, Information and Web Retrieval, Algorithms, Data Structures, Computer Architecture, Principles of Computer Systems

TECHNICAL SKILLS

Languages: Python, Java, C, C++, CUDA, OpenAI Triton, NVIDIA PTX/SASS, x86 assembly, Verilog, SQL

ML Skills: PyTorch, NVIDIA Nsight Compute/Systems, high accuracy LLM quantization, vector database (Llama-Index), knowledge graph RAG (retrieval-augmented generation), image segmentation, scikit-learn

EXPERIENCE

Software Engineer Intern at IBM Research — May 2024 - Aug 2024

- Researched and implemented high-accuracy integer quantization for accelerated LLM inference using rotation-based approaches, improving processing speed by 60-100%
- Modified an existing rotation algorithm for specialized hardware, resulting in a GPU kernel that outperforms the state-of-the-art by 10-35% (depending on problem size)
- Used PyTorch-based model compilation with custom GPU kernels to further increase throughput

Research Assistant in Learning Agents Research Group at UT Austin — June 2023 - present

- Implementing RAG architecture using Apache AGE knowledge graph and GPT-4 to track the state of environments from natural language updates
- Using GPT-4o to generate PDDL goals from natural language based on knowledge graph context

Research Assistant in Living with Robots Lab at UT Austin — May - Aug 2023

- Created 3D scene reconstructions for semantic queries using Facebook SAM and OpenAI CLIP
- Improved CLIP model accuracy through custom retraining for mislabeled instances

PUBLICATIONS

- K. Agarwal, R. Astra, A. Hoque, M. Srivatsa, R. Ganti, L. Wright, S. Chen. *HadaCore: Tensor Core Accelerated Hadamard Transform Kernel*. PyTorch Blog Post: pytorch.org/blog/hadacore. Paper: arxiv.org/abs/2412.08832.
- K. Agarwal, Y. Jiang, J. Hu, B. Liu, P. Stone. *L3M+P: Lifelong Optimal Planning with Large Language Models*. Paper: krishagarwal.github.io/projects/l3mp.

AWARDS/CERTIFICATIONS

- UT Austin Distinguished College Scholar
- USA Computing Olympiad Gold-level competitor
- Oracle Certified Associate Java SE 8 Programmer
- National Merit Scholarship Semifinalist