KARTIK SINHA

ksinha45@gatech.edu ⋄ linkedin.com/in/kartik-sinha-gt/ ⋄ kartikyz.github.io ⋄ github.com/kartikyz ⋄ (470)·833·0971 ⋄ US Citizen

EDUCATION

Georgia Institute of Technology August 2020 - December 2024
B.S./M.S. in Computer Science *Atlanta, GA*

GPA: 4.00, Faculty Honors
Specializations: Machine Learning, Systems and Architecture
ML, CV, Compilers, Adv. Operating Systems, Algorithms Analysis, Data Structures, OOP, Complexity Theory

WORK EXPERIENCE/RESEARCH

Citadel June 2024 - August 2024
Software Engineer Intern, Post-Trade Engineering - Corporate Actions *New York, NY*

- Automated biweekly prime broker tax reclaims process to book transactions (∼2MM total PnL) crediting the firm's trading desks, eliminating ∼98 % of manual touches by Operations teams for a process accounting for 6 percent of all global corporate action operations.
- Implemented and deployed an E2E transaction generation and auth-protected booking system using Java gRPC, SQL, server-side view engine, and internal services with browser-based user interface now used in production.

ML Systems Research January 2024 - Present
Student Researcher, *Atlanta, GA*

- Researching workload-aware adaptive serving of mixture-of-expert LLM models by varying the number of active experts during inference to realize latency gains.
- Implemented scheduling and expert activation policies atop vLLM for Mixtral 8x7B MoE model along with distributed LoRA full-precision finetuning on 8xH100 GPUs on PACE clusters using FSDP and DeepSpeed via Accelerate and HuggingFace libraries.
- Researching efficient training of graph neural network-large language model hybrid architectures.

Amazon Web Services May 2023 - July 2023
Software Development Engineer Intern, AWS Cryptography – Secrets Manager *Seattle, WA*

- Implemented and deployed new APIs for an AWS-critical internal service to manage secrets on a distributed system deployed to every AWS host, reducing customer tickets by 90 percent and 4 manual operations hours per week.
- Built APIs using Java Spring with Mockito unit tests and integration tests. Built rate-limiting mechanism to handle large scale traffic (1M+ hosts). Extended Perl CLI and integrated internal IAM service to automate verification.

Embedded Pervasive Lab, Georgia Tech August 2022 - Present
Student Research Assistant, UROP, College of Computing *Atlanta, GA*

- Researching scheduling methods for DAG workloads and function invocations to design an edge-native FaaS system meeting edge resource constraints and latency-critical objectives in a geo-distributed environment.
- Benchmarked transfer-learning ML models on edge cluster comprising RaspPis and Google Coral TPU accelerators.

Amazon May 2022 - August 2022
Software Development Engineer Intern, Amazon Ops Finance – Fusion *Seattle, WA*

- Delivered an end-to-end scalable native-AWS system prototype for financial report-generation (100+ pages), collating multiple financial and BI data ingestion sources (Redshift, SQL, etc.) for org-wide use (20+ teams).
- Implemented cross-team real-time collaboration using CRDTs and WebSockets. Used NodeJS, DynamoDB, IAM, STS, Fargate, S3, and the AWS CDK.
- Deployed internally to Beta. Presented to stakeholders, senior engineers, and an Amazon Finance VP.

College of Computing, Georgia Tech January 2022 - Present
Undergraduate Teaching Assistant, Project Lead *Atlanta, GA*

- Taught CS 3510 Algorithms Analysis and CS 2110 Computer Org. and Programming (digital logic, assembly, C).
- Managed teams of 5 TAs as Project Lead to create, revise, and test course projects and test suites on Java, C, and assembly for a course of 500+ students.
- Taught twice-weekly 1.5 hour labs for 50+ students, held office hours, received Thank-A-Teacher awards.

PROJECTS/LEADERSHIP

Song Transformer January – May 2023
CS 7643 Graduate Deep Learning Final Team Project *Atlanta, GA*

- Created an ML music generation pipeline by tuning and re-training 4 open-source ML models to take in as inputs a song and voice sample and reproduce a new song in the target voice. Wrote a paper in CVPR format with results.
- Used Demucs for waveform domain source separation to separate music and vocals. Used Spotify's U-Net and attention-based BasicPitch model for audio music transcription. Processed MIDI using Google's transformer-based model for music generation. Recombined resulting track with the SoftVC VITS SVC model's results.
- Used FFT convolutions to calculate cross-correlation between original and output tracks for similarity metrics.

Office Hours Booking System May 2022 - December 2022
Founder and Team Lead, Student Government Association IT Board, Georgia Tech *Atlanta, GA*

- Pitched to and secured funding from SGA leadership to create a unified cross-course queueing system for the College of Computing to improve existing office-hours and course administration logistics.
- Kickstarted a new project, interviewed other students to recruit 8 team members. Led software development, working directly with leadership, course staff and project team.
- Used Websockets, AWS Lambda for SSO and backend service, DynamoDB, with frontend hosted on S3.

Graph Algorithms Animations Visualizer May 2021 - July 2021

- Created website using React and TypeScript to visualize animations of algorithms for directed/undirected and weighted/unweighted graphs. Hosted on GitHub Pages.
- Allows users to build custom graphs or use random graphs or network/grid graphs. Implemented animations for depth-first search and breadth-first search algorithms.

TECHNICAL SKILLS

Languages	Java, Python, C, TypeScript, JavaScript, Assembly, HTML/CSS, Perl, Bash
Frameworks	AWS services, NoSQL, SQL, PyTorch, Kubernetes, scikit-learn, Pandas, Tensorflow, gRPC, Node.js Git, Docker, UNIX, React, Flutter