

James Chen

(770) 309-2207 | jameschen0831@gmail.com

EDUCATION

Georgia Institute of Technology

Atlanta, GA

B.S./M.S. Computer Science, GPA: 3.96 (Concentration: Machine Learning)

Expected Graduation: December 2026

EXPERIENCE

Amazon

May 2025 - August 2025

Software Development Engineer Intern

Sunnyvale, CA

- Architected ML-powered event monitoring system processing 14.4TB daily data, serving 50+ developers
- Developed multi-source data ingestion pipeline processing 50K+ documents across Quip, Wiki, and dashboards for RAG-based AI assistant integration
- Developed real-time Grafana dashboards pipeline, decreasing report generation time from 10+ hours to 15 minutes
- Built data pipeline with 99.9 % reliability using AWS CDK and CI/CD, processing 100GB+ event data per 10-minute interval

Transformation Continuum - Google Cloud Commission

November 2023 - May 2024

Full Stack Software Engineer Intern

Austin, TX

- Delivered novel sales tooling to be used by Google employees using ReactJS
- Developed Node.js REST APIs, improving system integration efficiency by 40%.
- Implemented Quantitative Assessment Module with 500+ standardized questions using React components
- Streamlined training by deploying LLM-powered chatbot to website for reviewing performance

Georgia Tech LIDAR Lab

January 2024 - Present

Reinforcement Learning Research Assistant

Atlanta, GA

- Conducted research on bipedal navigation in human-crowded environments, addressing challenges of nonlinear dynamics in bipedal locomotion
- Developed K-means clustering algorithm and single shot ego agent, reducing processing time by 15%
- Delivered re-targeting pipeline for humanoid datasets with 100K+ motion frames (MoCap, Mujoco, AMASS) for RL training

PROJECTS

Globally Shared KV Cache for LLMs | Python, Golang, HuggingFace, LLaMA 2, AWS, Redis

May 2025

- Developed 3x cache-hit rate, 30% memory usage and 2.8x speedup in inference latency compared to no-cache
- Integrated semantic similarity logic using cosine distance and embeddings for reusable queries
- Achieved ~900ms reduction in response time on ~0.5MB memory overhead
- Simulated on 1000+ real and synthetically generated queries across 3 datasets

Sano Gym Social Media App | Swift (Frontend), Python (Backend), SwiftUI, Flask

August 2024

- Developed a full-stack social media app tailored for fitness enthusiasts, with a Swift-based frontend and a Python Flask backend.
- Launched at Georgia Tech, achieving over 1,000 downloads within the first month.
- Integrated workout tracking, routine sharing, and community engagement features to enhance user interaction and motivation.

HackGT Contractly AI | Python, Flask, React, MongoDB, LangChain

October 2023

- Full-stack web app Contract Analysis tool with Large Language NLP model trained on a dataset of over 500 contracts spanning NDAs, mergers, acquisitions, and more
- Implemented Chrome Extension which fetches websites privacy statement to rate using ChatGPT-3

TECHNICAL SKILLS AND AWARDS

Programming Languages: Java, C, Python, C++, SQL, JavaScript, TypeScript, Swift, HTML/CSS, Dart

Languages: English, Mandarin

Libraries: pandas, NumPy, Matplotlib, React, Angular, PyTorch, TensorFlow, Flask, Django, Spring, OpenCV

Technologies: AWS, Docker, Kubernetes, Git, REST APIs, SQL, NoSQL, MongoDB, Firebase, CI/CD, Agile, Scrum

Awards: Highest Honors, Dean's List, Faculty Honors, AIME qualifier, National Merit Finalist