Kester Tan

Looking for Software Engineer roles starting May 2026

(+1) 412 9097712 | kestertan@cmu.edu | https://kestertan.github.io/kester/ | KesterTan | tankester

Education

Carnegie Mellon University (Pittsburgh, Pennsylvania) — Graduating May 2026

Bachelor of Science Computer Science + Information Systems (Double Major)

- **Grades/Honors:** Dean's List High Honors
- **Relevant Coursework:** Algorithm Design & Analysis, Distributed Systems, Computer Systems, Parallel Data Structures & Algorithms, Generative AI, Machine Learning, Application Design & Development, Probability for CS, Software Engineering, Functional Programming, Imperative Programming

Skills

- **Programming Languages:** Python, C, C++, TypeScript, JavaScript, Standard ML, Ruby on Rails, Java, Go
- **Technologies:** React, React Native, NextJS, TailwindCSS, ExpressJS, NodeJS, Flask, Socket.IO, MongoDB, MySQL, Docker, Figma
- **ML Technologies:** PyTorch, Numpy, Llama.cpp, TensorFlow, Pandas, Keras, Autogen, LangChain, LangGraph, LlamaIndex
- **Infrastructure:** AWS (EC2, Lambda, CloudFormation, API Gateway), Terraform, Bash, Linux (Ubuntu)

Experience

Microsoft Core AI, Software Engineer Intern (Mountain View, California) — May 2025 – Current

- Developed a custom quantization tool that fine-tunes every tensor of Hugging Face Phi & Llama models to a custom tensor type
- Custom quantization achieved 5–10% performance gain on BabelBench dataset over equivalent-size GGUF quantization from Hugging Face/Llama.cpp
- Built an export pipeline to pack custom quantized tensors into Llama.cpp GGML format and convert models to GGUF files, enabling SIMD & GPU-accelerated inference with <1% accuracy loss
- Designed tensor reshaping logic and implemented nested group quantization for tensors, biases, and scales using constant-time Torch operations
- Created layer-by-layer analysis scripts for quantized models, reporting average bit rate, datatype, and metadata for fine-grained evaluation
- Dockerized the full evaluation workflow in Llama.cpp, allowing researchers to run experiments 3× faster

Amazon Web Services, Software Development Engineer Intern (Seattle, Washington) — May 2024 – August 2024

- Built a standalone, localized feedback & sentiment analysis platform (Java, React, TypeScript) adopted by 10+ teams to improve workflow insights.
- Developed a metrics and event tracking solution using CloudWatch, S3, Athena, and QuickSight, providing real-time analytics for cross-region teams
- Built internal social media platform for Amazon employees to connect and share content, presented to senior leadership for organization-wide adoption
- Implemented Bean validation in Java to enforce API contracts and integrated custom Lambda validators to improve backend reliability

Autolab, Lead Software Engineer (Pittsburgh, Pennsylvania) — Jan 2023 – Current

- Led cloud deployment for one-click installation using Docker, Amazon EFS, Lambda, EC2, CloudFormation & Cognito
- Created File Manager feature from scratch using C & Rails to abstract underlying database file system, used by 100+ professors/instructors worldwide
- Created password-less login feature from scratch using Rails & MySQL, providing temporary logins for 500+ students
- Launched metrics feature used by 100+ professors to monitor & track student performance using MySQL & Rails

IMDA, AI Data Engineering, Software Engineer Intern (Singapore, Singapore) — May 2023 – Sep 2023

- Researched LLM & NLP models for GPTxLegal (summarization & Q&A tool for legal companies)
- Built, designed & hosted the entire API interface & frontend chat interface for GPTxLegal (contributed to 30% of codebase)
- Built slack bot from scratch to automatically create EC2 instances using JavaScript, Terraform & AWS, used by 50+ engineers every day

ZUZlabs, Software Engineer & Product Manager (Pittsburgh, Pennsylvania) — Aug 2022 – Aug 2023

- Led React Native team responsible for end-to-end mobile development & built marketplace & vendor interface from scratch using React, Node.JS & TS

Hypotenuse AI (YC S20), Product Manager (Singapore, Singapore) — Jan 2022 – Aug 2022

- Pioneered & scaled company 2nd largest growth channel & re-engineered auth systems to reduce spam & fake account creation by 30%
- Launched an end-to-end event tracking system to drive in-depth data app analytics, C-Suite business decisions & customer acquisition campaigns
- Designed, improved & conducted SEO optimizations to create webpages that ranked organically on the 1st page of Google using Figma & Webflow

Projects & Clubs

Competitions — June 2016 – May 2025

- **Amazon Alexa AI Social Bot Challenge:** Only undergraduate chosen to represent CMU, evaluated LLMs & wrote data augmentation scripts
- **Project Innervate Founder:** 1st at International Future Problem Solving Conference, built a Parkinson Disease Simulator, used by 1000+ people

Leadership Roles — Aug 2023 – May 2025

- **CMU Blockchain Head of Engineering:** Led 2 teams of 10 engineers to build decentralized applications (Ethereum node validator, decentralized blogging)
- **Singapore Student Association Head of PR:** Automated successful email & marketing campaigns for 100+ students

Additional Projects — Aug 2023 – May 2025

- **Dashseek.AI:** Built a multi-agentic system that abstracts FP&A work (analyzing spreadsheets & building income reports) using Autogen & React
- **Raft Consensus:** Implemented Raft, a state replicated machine protocol in Go
- **Distributed Bitcoin Miner:** Built a bitcoin miner from scratch on top of a custom protocol adapted from UDP protocol with a client & server APIs
- **Enhanced Gorilla API Call Generation through Supervised Fine-Tuning on NESTful Benchmark:** Fine-Tuned Gorilla model using augmented data
- **HowYa:** Built mobile journal entry with sentiment analysis using Swift, Node.JS, Prisma ORM, CockRoach DB & Keras
- **Reversi.AI:** Built a multiplayer AI Reversi game with Monte Carlo & Minimax algorithms using Python & Sockets
- **Malloc:** Dynamic memory allocator implemented via segregated free lists to optimize native C allocations
- **Proxy:** Implemented multi-threaded file caching web proxy that supports HTTP requests
- **CMUD:** Implemented multi-player online game with multiple components communicating via RPCs, used Actor Model for coordination & consistency