TANG KEN YI

Tel: +65 86618162 | Email: tangkenyi2001@gmail.com | LinkedIn: https://www.linkedin.com/in/tang-ken-yi-05629421a/ | Personal Portfolio: https://tangkenyi2001.github.io/

EDUCATION

NANYANG TECHNOLOGICAL UNIVERSITY Aug 2022 – Present
Bachelor of Engineering (Computer Engineering)
Expected Graduation Date: December 2026

- Second Class (Upper) Honours/ Honours (Distinction), Current CGPA: 4.43/5.00
- Dean's List AY 24/25
- Relevant Coursework: Data Structures and Algorithms, Operating Systems, Computer Networks, Object Oriented Programming, Machine Learning, Neural Networks, Natural Language Processing

EXPERIENCE

Apple Feb 2026 – August 2026
Software Engineer Intern

•. Working on the Apple Online Store team to design and engineer AI-powered data pipelines processing 80+ GB of production traffic daily.

PayPal July 2025 – December 2025
Software Engineer Intern

- Engineered an agentic framework that ensures deterministic code quality by orchestrating Claude Code within autonomous TDD feedback loops, driving the model to iteratively refine outputs until validation passes (20–30% faster development).
- Developed, containerized, and deployed remote MCP (Model Context Protocol) servers on AWS using Docker and Kubernetes.
- Led exploration and deployed the team's first OAuth 2.0 authentication system with SSO for MCP servers.
- Presented solutions to 30+ stakeholders across multiple time zones and authored internal technical blogs to increasing adoption and usage of newly developed tools across engineering teams.

SAP Jan 2025 – July 2025
Software Engineer Intern

- Collaborated with cross-functional teams to design and release a full-stack, end-to-end internal ChatGPT platform, focusing on backend services.
- Engineered key backend features in Node.js, including file upload integration with Amazon S3, persistent chat history storage, and document-based conversational querying.
- Led and coordinated the development of a PowerPoint generation feature leveraging LLMs to dynamically create presentation content; integrated into the internal productivity platform.
- Integrated robust observability by logging metrics and application data to OpenSearch, enabling stakeholders to monitor and evaluate usage and performance trends.
- Deployed and optimized Large Language Models (LLMs) for production utilizing Docker and vLLM, reducing inference costs and improving response latency through efficient model serving.

PROJECTS

Reducing Cold-Start Latency in LLM Inference (Final-Year Project) May 2025 – Present

- Forked and extended the vLLM inference engine to prototype a disaggregated inference architecture separating executors from persistent GPU workers to mitigate cold-start latency.
- Implemented a Worker Controller to manage GPU workers, model lifecycle, and resource allocation under dynamic workloads.
- Reduced cold-start engine overhead by 50% per model (from 2.6s to 1.3s) by avoiding repeated CUDA initialization.
- Achieved 10% faster end-to-end cold start over baseline vLLM, with 6.5s savings across 3 models, validated via controlled benchmarks.

SKILLS & CERTIFICATIONS

Languages**:** Proficient in English and Chinese, conversant in German
Technical Skills**:** Python, C, Java, SQL, React, JavaScript, TypeScript, Docker, Git, Kubernetes
Certificates: Amazon Solution Architect (Associate), Generative AI with Large Language Models by DeepLearning.AI
Hackathons: EasyA x BCG Hackathon (4th Place), NTU Techfest Hackathon (Participant)