Kai-Yuan Cheng

j20001020@gmail.com | +886 954 151 913 | github.com/j20001020

Education

National Cheng Kung University Sep 2024 – Jun 2026
M.S. in Computer Science and Information Engineering

- Distributed Systems Laboratory
- Thesis: A Configuration Tuning Enhancement Framework Based on LLM and Bayesian Optimization: A Case Study of RocksDB
- Proposed a hybrid framework combining Large Language Model and Bayesian Optimization to address cold-start and high-dimensional optimization challenges in RocksDB tuning

National Kaohsiung University of Science and Technology Sep 2019 – Jun 2023
B.S. in Computer Science and Information Engineering

Experience

IT Student Assistant, National Cheng Kung University, Office of Sustainable Strategy Development – Tainan, Taiwan Sep 2024 - Jul 2026

- Collaborated with AI Agents to develop a new department website and CMS, decomposed requirements, migrated the legacy website, defined the Agent workflow, and completed deployment and maintenance using Debian, Docker, and Nginx.
- Assisted in introducing Codex and guided colleagues on its usage, driving AI tool integration and automating administrative tasks.

Full Stack Engineer (Intern), Titansoft – Kaohsiung, Taiwan Feb 2023 – Jun 2023

- Designed and developed full-stack systems using ASP.NET Core, Vue.js, and MySQL.
- Designed and implemented RESTful APIs and maintained comprehensive API documentation using Swagger.
- Developed a JWT-based authentication mechanism and an IP Rate Limiting middleware module to enhance API security.
- Introduced an Object Mapper conversion mechanism in .NET Core; the design was adopted by senior engineers as a reference.
- Authored unit tests and integrated GitLab CI pipelines to automate testing and deployment processes.
- Collaborated in an Agile environment, participating in requirement analysis, task estimation, and daily stand-ups within Sprint iterations.

Projects

Cloud-Based Image Processing System (Golang)

- Adopted an Event-Driven Architecture using GCP Pub/Sub to decouple the API upload service from the Worker processing service, enabling asynchronous thumbnail generation and AI label recognition workflows.
- Integrated Google Cloud Vision API for image label recognition.

E-commerce Backend System (Spring Boot)

- Implemented JWT authentication with Redis-based token state management (TTL control and blacklisting), addressing the security limitation of stateless JWTs being unable to be actively revoked.
- Integrated GCP Cloud Storage for image persistence, and used Secret Manager to centrally manage service credentials, preventing sensitive information from being hardcoded.

RAG Knowledge API

- Built a Retrieval-Augmented Generation (RAG) API with an agent-style conversational interface. Used a ReAct loop so the agent could autonomously decide whether to consult the knowledge base.
- Tracked each LLM call's token usage and API cost, aggregating daily statistics for usage monitoring.

Skills

Languages: Golang, Python, C#, Java, JavaScript, C/C++

Frameworks: ASP.NET Core, Spring Boot

Technologies: Linux, Git, Docker, Kubernetes, GCP, PostgreSQL, MySQL, CI/CD, Agile/Scrum

AI Tech: Gemini API, Claude Code, Agentic Development Workflow, AI Agent