Jui-Hung, Cheng

+886 932 395 392 | Taipei, Taiwan | ryan98153@gmail.com | github.com/ryancheng98153

Education

National Chenhgchi University (NCCU) Taipei, Taiwan

Bachelor's of Science in Computer Science, Double major in Electro Physics *Sep 2021 — Jun 2025*

- Teaching Assistant of *Natural Language Processing* Mar 2024 — Jun 2025
- Teaching Assistant of *Distributed System* Mar 2025 — Jun 2025

National Yang-Ming Chiao-Tung University (NYCU) Hsinchu, Taiwan

Institute of Data Science and Engineering *incoming, Sept 2025 —*

Personal Skills

Programming Languages: C/C++, Python, Bash, Typescript, Golang, Latex

Software Development: React JS, Rest API, MySQL, MongoDB, gRPC, MQTT

Tools: Git/Github, Docker, Unity, Postman, Llama-index, Grafana k6

AI skills: Deep Learning, PyTorch, LLMs, RAGs, Image Processing

Research & Work Experiences

Academia Sinica Jul 2024 — Dec 2024

Scholarship Research, Advisor: Associate Research Fellow Huang, Hen-Hsen"

- First authored the research paper "Don't Do RAG: When Cache-Augmented Generation is All You Need for Knowledge Tasks" at The ACM Web Conference 2025,
- Co-authored the research paper "Efficient Beam Search for Large Language Models Using Trie-Based Decoding"
- Recently researching LLM Reasoning with GRPO and exploring and context compression during multi-step inference.

National Science and Technology Council Mar 2022 — Mar 2025

Scholarship Researcher, Advisor: Scientist Lin, Yu-Cheng

- Contributed to the research proposal "Multifunctional Quantum Annealing Toolkit and Its Applications" and successfully selected for the College Student Research Scholarship, which has a 30% to 40% selection rate over the years.
- Researched the technology of quantum annealing; used Canada's D-wave quantum computer to simulate annealing.
- Co-developed a multi-functional quantum annealing tool using C/C++, optimizing parallel computing with C++ MPI and creating lattice structures like Triangular, Kagome, and MapleLeaf for faster construction and analysis.

National Chengchi University (NCCU) Mar 2024 — Feb 2025

Student, Advisor: Assistant Professor: Sun, Shi-Sheng

- Contributed to the research proposal "Optimizing Electric Bus Charging Schedules for Extended Battery Life" and successfully selected for the College Student Research Scholarship, which has a 30% to 40% selection rate over the years.
- Developed EV-scheduling system simulating charging process and optimizes energy consumption using Genetic Algorithm.

Academic Reports

Don't Do RAG: When Cache-Augmented Generation is All You Need for Knowledge Tasks Jan 2025

First Author, 5 Pages, The ACM Web Conference 2025

- This paper proposes Cache-Augmented Generation (CAG) as an alternative to RAG, eliminating retrieval latency and errors by preloading knowledge into LLMs with extended context and KV-cache optimization. For constrained knowledge bases, CAG offers a simpler, more efficient approach with comparable or superior performance.

Multifunctional Quantum Annealing Toolkit and Its Applications May 2024

Contributed Researcher, National Science and Technology Council

- The project focuses on developing an open-source C++ software package that combines simulated annealing, simulated quantum annealing, and quantum annealing algorithms for solving combinatorial optimization and theoretical physics problems. The software will be modeled after existing interfaces and designed for use on traditional computers.

Optimizing Electric Bus Charging Schedules for Extended Battery Life May 2024

Contributed Researcher, National Science and Technology Council

- The project develops an electric bus charging optimization system using Greedy and Genetic Algorithms to improve energy efficiency. Reviewed prior research and refined system variables by interviewing Taipei Hsin-Hsin Bus Company. Built a simulation framework to model charging processes and optimize scheduling for better fleet management.

Awards & Honors

5 points in AIME, (PR 91 in AMC 12) American Invitational Mathematics Examination Mar 2020

4 out of 7 (PR 91), Collegiate Programming Examination (CPE) Mar 2022

Certificate of Participation 2022 Game Design Hackathon, BlackHole Creative Co., Ltd. Oct 2022

Enter the finals, National Collegiate Programming Contest (NCPC) Oct 2022

Merit Award, Quantum Hackathon, Applied Physics Dept., NCCU Nov 2023

Outstanding Award, 2024 Blockchain Hackathon Sui Track, XueDao organization Aug 2024

Extracurricular Activities

Tech Team Member, Google Developer Groups on Campus Sep 2021 — Sep 2022

- Cooperated with Google; Built club's website; served as lecturer for club's classes and giving technical support to members.

Frontend member, PeoPo Citizen Journalism Platform Project Sep 2022 — Jun 2023

- Built an online news website including the front-end using ReactJS, microservice, back-end and database.

Teaching Assistant, NCCU CS Camp Jul 2023 — Feb 2024

- Served as Lecturer for Unix commands and shell script; Teaching Assistant for Unity Game Development

Projects

CAG (Cache-Augmented Generation), Main Contributor github.com/hhhuang/CAG

- Research Code: The source code implementation for my research paper "Don't do RAG" which provide the example usage of Cache-Augmented Generation and comparison between RAG and CAG.
- It has gained **1,100+ GitHub stars** and highlighting its impact and interest in retrieval-free knowledge augmentation.
- Tool Used: Llama-index, Hugging-face, PyTorch

Digital Annealer, Co-developer github.com/GNITOAHC/DigitalAnnealer

- Side project: A multi-functional annealer contains several algorithm, such as simulated annealing, quantum annealing and optimizing with Message Passing Interface (MPI).
- Tool Used: C/C++, Bash, MPI

Dwave quantum annealer tools, Developer github.com/RyanCheng98153/dwave-triangular

- Side project: A tool for handling D-Wave quantum computer, allowing users to input common lattice or custom graph.
- Tool Used: Python, D-wave quantum computing, Bash Script, Networkx

Electric Bus Scheduling, Developer github.com/RyanCheng98153/EV-charge-scheduling

- Side project: An EV bus simulation tool that optimizes energy consumption and determines the best charging schedule.
- Technique Used: Greedy Algorithm, Genetic Algorithm, OOP, Factory Design Pattern

Flight simulator, Co-developer github.com/GNITOAHC/2024-DistributedSystem

- Course project: A Distributed-System project control by a airplane model IOT device simultaneously storing device data and synchronizing flight status display.
- Tool Used: IOTDB, MQTT, FastAPI, threeJS, Golang

Meta CRAG, Co-developer gitlab.aicrowd.com/chaoting_chen/meta-comphrehensive-rag-benchmark-starter-kit

- Course project: A contest project that manually implements RAG including query classification, fine-tuning reranking.
- Technique Used: BeautifulSoup, Llama-index BGE-reranking, Fine-tuning, Chain of Thought

Omniverse Camera Capture, Developer github.com/RyanCheng98153/Omniverse-camera-capture

- Side project: A simple real-time scene capture tool for NVIDIA Omniverse USD.
- Tool Used: Python, NVIDIA Omniverse USD