Yu-Chen CHENG

https://rudeigerc.dev · rudeigerc@gmail.com

Highlight

Yu-Chen Cheng is a software engineer specializing in large-scale AI/LLM infrastructure and Kubernetes-native platforms. With 2+ years of experience architecting production AI/LLM systems, he participated critical infrastructure initiatives at Heywhale and drove platform modernization at WizardQuant as a AI platform development engineer. He combines technical expertise in distributed systems with proven ability in cross-functional collaboration, and establishing technical standards. Active contributor to CNCF projects including Kubeflow.

Education

Shanghai Jiao Tong University Shanghai, China
M.Eng. in Software Engineering *Sep. 2019 - Mar. 2023*
Thesis: *Distributed Data Processing Job Scheduling in Multi-Tenant Shared Clusters*

Shanghai Jiao Tong University Shanghai, China
B.Eng. in Software Engineering *Sep. 2015 - Jun. 2019*

Experience

Heywhale Shanghai, China
Software Engineer *May. 2025 - Present*

- **LLM Post-Training.** Co-architected and led implementation of LLM post-training pipeline using Kubeflow Trainer, enabling scalable data curation, model training, and evaluation workflows that support production LLM deployments.
- **LLM Inference.** Participated in the migration from Knative to AIBrix, redesigning inference APIs with advanced prefix-aware routing and time-series based horizontal autoscaling, achieving improvement in performance metrics. Designed and implemented comprehensive LLM inference benchmarking framework to evaluate performance metrics across different model architectures and deployment configurations.
- **Machine Learning Platform.** Partnered with product and research teams to define AI/LLM infrastructure requirements and technical roadmap. Drove platform stability and observability enhancements through comprehensive DevOps pipelines based on GitLab Runner and OpenTelemetry integration for end-to-end distributed tracing.

Kubernetes | Go | Python | LLM | LLMOps

WizardQuant Shanghai, China
AI Platform Development Engineer *May. 2023 - Sep. 2024*[1]

- **AI Platform.** Co-architectured redesign and refactoring of internal MLOps platform on Kubernetes, driving the modernization of Python SDK and CLI, unified Job models architecture, microservices implementation via Protobuf and gRPC, and full-stack dashboard development with React and Vue.
- **DevOps Pipeline.** Designed and built efficient Python and Go DevOps pipelines upon Tekton to accelerate the developing and testing cycles.
- **Technical Leadership. a)** Established team-wide modern Python development standards with PDM. **b)** Introduced Helm templates for hosting static sites on Kubernetes with integrations of Tekton pipelines and OAuth Proxy. **c)** Built a unified documentation generation tool based on VitePress.
- **Mentorship.** Mentored **a)** the development and integration of the Dex identity provider with LDAP, providing an authentication and authorization middleware for internal systems. **b)** the benchmarking framework for Kubernetes schedulers, providing a unified interface for benchmarking and comparing different schedulers.

Kubernetes | Go | Python | Ray | MLOps

Ant Group Shanghai, China
Site Reliability Engineer Intern *Jun. 2021 - Sep. 2021*

- **Federation Observability Platform.** Participated in and developed the internal multi-clusters observability platform, enabling unified monitoring across distributed Kubernetes environments.

[1] I served in military training service in Taiwan from October 2024 to February 2025.

- **Identity Provider Connector.** Led the development of the Dex identity provider plugin connecting to the internal identity and access management system, providing OIDC services for internal systems such as Argo CD and Grafana.

Kubernetes Go React

The Linux Foundation Remote
Mentee of LFX Mentorship Program, CNCF - Volcano *Feb. 2021 - May. 2021*

- **System Stability Enhancement.** Built an efficient testing framework based on Kubernetes and imported unit tests and E2E tests to cover more classical scenarios for Volcano, a Kubernetes native batch scheduling system for compute-intensive workloads.

Kubernetes Go

Services

- Mentor of Google Summer of Code 2025, Kubeflow, **Project 10: Support Volcano Scheduler in Kubeflow Trainer**

Programming Skills

Languages: Go, Python, TypeScript, C++
Skills and Expertise: LLMOps, MLOps, Resource Management, Scheduling
Tools and Frameworks: Kubernetes, Kubeflow, Ray, React, Vue

Licenses and Certifications

Japanese-Language Proficiency Test Level N1 2017

Miscellaneous

Blog: https://rudeigerc.dev
GitHub: https://github.com/rudeigerc
LinkedIn: https://www.linkedin.com/in/rudeigerc/
Medium: https://medium.com/@rudeigerc
Zhihu: https://www.zhihu.com/people/rudeigerc
WeChat Official Account: YC Cheng (@yuchenrcheng)
Languages: Chinese (Native proficiency), English (Professional working proficiency), Japanese (Limited working proficiency)

Open Source Contributions

- Member of **Kubeflow**, **InftyAI**, **Tekton** and **Tensorchord**
- Contributed to repositories: kubeflow/trainer, InftyAI/llmaz, kubernetes-sigs/inference-perf, tektoncd/triggers, volcano-sh/volcano, tensorchord/envd, ray-project/kuberay, pdm-project/pdm

Last Updated: 2025-09-03