Sanfu Yee (Pengfei Li)

Beijing, China | sanfuyee@gmail.com | github.com/sanfuyee | x.com/lewiscooper1024

Software engineer with 10+ years of experience in backend development, distributed systems, and AI infrastructure. Specializing in building scalable, high-availability platforms that bridge the gap between cutting-edge AI models and production-ready services.

EDUCATION

University of Science and Technology of China
M.S. in Computer Applied Technology, 2013 - 2016

Wuhan University
B.S. in Computer Science and Technology, 2009 - 2013

CORE COMPETENCIES

- **Backend & Architecture -** 10+ years across fintech, e-commerce, and AI; building high-availability systems with K8s, Istio, and Prometheus
- **Distributed Systems -** Golang / Java / Python; microservices, DDD, TiDB, Kafka, gRPC
- **AI Engineering -** MaaS platform for DeepSeek-V3/R1 and text-to-video models; inference optimization with vLLM and TensorRT-LLM; RAG/Agent deployment
- **CS Fundamentals -** Strong foundation in algorithms, networking, and systems; first-principles debugging

WORK EXPERIENCE

HPC-AI Tech

MLOps Engineer | Nov 2024 - Present

- Built GPU cloud platform using Karmada to manage multi-cluster Kubernetes, providing instances, jobs, and APIs for customers to run training workloads
- Built MaaS (Model as a Service) platform from scratch as the core middleware between inference engines and customer-facing products
- Shipped DeepSeek-V3/R1, VideoOcean text-to-video, Hunyuan text-to-video, and Flux.dev text-to-image models to production
- Designed flexible billing system (usage/time/per-request), model marketplace, and API key management
- Built TPM/RPM rate limiting and multi-tenant authentication for system stability
- Established observability dashboards, alerting pipelines, and standardized error code taxonomy

Tech: Golang, Go-Zero, Kubernetes, Karmada, Helm, Kafka, Redis, PostgreSQL

WPS (Kingsoft Office)

Expert Software Engineer | Jul 2023 - Sep 2024

- Productionized LLM features: intent recognition, reading comprehension, PPT auto-outline (RAG), and PPT-Agent
- Built high-performance model serving with FastAPI/Tornado
- Applied TensorRT-LLM and vLLM for inference acceleration, reducing latency and compute costs

Tech: Python, FastAPI, Tornado, TensorRT-LLM, vLLM, RAG

IDEA Institute

AI Full Stack Engineer | 2021 - 2023

- Contributed to GTSFactory framework for generative AI research and engineering

Shopee

Senior Software Engineer | 2018 - 2021

- Built CRM push platform from scratch serving 600K+ sellers with 30M+ daily pushes during peak promotions
- Designed membership system with Bloom filter validation, TiDB optimistic transactions, and idempotent APIs
- Developed real-time WebChat instant messaging system for sellers and buyers

Tech: Golang, Gin, GORM, MySQL, Redis, gRPC, etcd, TiDB

WeBank

Software Engineer | 2016 - 2018

- Developed mobile banking authentication and security modules
- Built KYC (Know Your Customer) identity verification and compliance workflows

Tech: Java, Spring Boot, MyBatis

TECH STACK

Languages: Golang, Java, Python

Frameworks: Go-Zero, EntGo, Gin, gRPC, Spring Boot, Spring Cloud, FastAPI, Django, Celery

Infrastructure: Kubernetes, Docker, Helm, Istio, Prometheus, Alertmanager, CI/CD

Databases & Storage: PostgreSQL, MySQL, TiDB, Redis Cluster, etcd, Elasticsearch, Milvus

Messaging: Kafka, RabbitMQ

AI/ML: vLLM, SGLang, TensorRT-LLM, Kubeflow, RAG, LLM, VLM