

Sai Mun Lee

SENIOR SOFTWARE ENGINEER, MACHINE LEARNING

☎ (+65) 9336 4288 | ✉ saimun92@icloud.com | 🏠 saimunlee.com | 🌐 lee-sai-mun-8a12726

Summary

Senior Software Engineer with 8+ years building ML infrastructure and large-scale distributed systems. Built multi-tenant serving platforms for tabular foundation models and recommendation engines from the ground up — owning model inference orchestration, lifecycle management, and inference optimization at enterprise scale (8B+ recommendations served). Pairs deep distributed-systems expertise with core ML (recommender systems, foundation models, PyTorch) and a track record of leading engineering teams through zero-to-one, high-ambiguity initiatives.

Skills

ML & AI	Foundation Model Serving, Model Inference & Optimization, Recommender Systems, PyTorch, LLM & RAG (prototyping)
Languages	Python, Go, TypeScript, JavaScript, Bash
Backend	FastAPI, Flask, Django, PostgreSQL, HANA DB, ElasticSearch
Infrastructure	Kubernetes, Docker, Cloud Foundry, Jenkins, Terraform
Cloud	AWS (Solutions Architect Pro), GCP (Associate Cloud Engineer)
Practices	Distributed Systems, System Design, ML Platform Architecture, TDD, Code Review
Spoken	English, Chinese

Experience

SAP Asia

Singapore

SENIOR SOFTWARE ENGINEER | CO-PRODUCT OWNER

Dec 2020 - Present

- Co-architected SAP's Tabular AI platform with a team of architects — a highly scalable, multi-tenant serving layer for the RPT-1 tabular foundation model (few-shot classification and regression without per-task fine-tuning), targeting 10+ enterprise tenants at 100K users each on Kubernetes via SAP AI Core.
- Owned the end-to-end ML serving architecture: API contracts, tenant isolation, model inference orchestration, and model lifecycle management across the platform.
- Optimized the model inference serving path to improve latency and throughput for concurrent multi-tenant workloads.
- Led a 15-person engineering team delivering Personalized Recommendations, a recommender engine that generated 8B+ job recommendations in 2025 for enterprise customers on SAP SuccessFactors.
- Architected People Like Me — a microservice that trains collaborative-filtering models on employee clickstream data and serves precomputed batch recommendations via REST APIs.
- Built multi-tenant onboarding APIs and background workers on Cloud Foundry, orchestrating concurrent model training cycles at scale.
- Instrumented the platform with Dynatrace APM and Kibana logging (correlation-ID tracing) and established CI/CD (Jenkins) with unit/integration/e2e test coverage, reducing mean-time-to-resolution for production incidents.

TAU Express

Singapore

LEAD SOFTWARE ENGINEER

Apr. 2020 - Dec. 2020

- Led development of a knowledge management system enabling analysts to retrieve and correlate insights from cybersecurity documents using graph-based relationships.
- Owned end-to-end delivery: gathered requirements directly from users, designed system architecture, and facilitated communication between users and the development team.
- Tech: React, Django, Docker, Neo4J, ElasticSearch, Cytoscape.js

SPIRIT Research Center (NTU)

Singapore

RESEARCH ENGINEER

Feb. 2018 - Mar. 2020

- Developed the Intelligent Case Retrieval System (ICRS) for Singapore's smart nation initiative — a full-stack application (Django, Vue.js) for efficient retrieval of legal documents.
- Built web crawlers (Selenium, BeautifulSoup) to index Chinese news articles and social media content for document retrieval pipelines.

Certificates

Dec 2021 **AWS Certified Solutions Architect – Associate**, Amazon Web Services (AWS)

Feb 2022 **GCP Associate Cloud Engineer**, Google Cloud Platform (GCP)

May 2023 **Hashicorp Certified: Terraform Associate (003)**, Hashicorp

AWS Certified Solutions Architect – Professional, Amazon Web Services (AWS)

In Progress

Nanyang Technological University (NTU)

B.S. IN COMPUTER SCIENCE (HONOURS) — MACHINE LEARNING & DATA SCIENCE SPECIALISATION

Singapore

2013 - 2017