

CHUA SIOK CHENG

AI / MACHINE LEARNING ENGINEER

Singapore Citizen | +65 9618 9621 | siokcheng.chua96@gmail.com
linkedin.com/in/chua-s-51a194143 | github.com/94399345sc | 94399345sc.github.io

SUMMARY

AI / Machine Learning Engineer who ships production LLM systems end to end. Sole architect and engineer of a multilingual HR AI assistant (RAG + LLM), taken from a one-line business requirement to a production Phase 1 launch in six weeks. Earlier work spans OT cybersecurity anomaly detection, GenAI localization and dubbing pipelines, and time-series forecasting. Comfortable across the full delivery path: retrieval architecture, FastAPI services, Next.js interfaces, Docker deployment, automated testing and observability.

CORE SKILLS

AI / ML: RAG architecture, LLM application design, prompt engineering, vector embeddings (BGE-M3), semantic and hybrid retrieval, multi-agent systems, multilingual NLP (English / Chinese), anomaly detection, time-series forecasting, model evaluation, DeepSeek, OpenAI API

Languages & Frameworks: Python, FastAPI, Pydantic, TypeScript, React, Next.js, SQL, C, C++, PyTorch, TensorFlow, Keras, scikit-learn, XGBoost, Pandas, NumPy, Streamlit

Platforms & Tooling: Docker, Docker Compose, Redis, Google Cloud Platform (Cloud Run, Cloud Build, Cloud Workflows, Artifact Registry, GCS), REST API design, OAuth 2.0, webhook integration, Google Workspace APIs, pytest, CI/CD, Git, GitHub, GitLab, OpenSearch, Linux

Practice: Product ownership, technical roadmapping, PII and data governance, observability, automated testing

PROFESSIONAL EXPERIENCE

OlaParty / Qoretex, Singapore — AI Product Manager Trainee Jun 2026 – Present

- Designed and single-handedly built a production multilingual HR AI assistant answering employees in English and Simplified / Traditional Chinese, with governed policy coverage across 8 jurisdictions — owning product scope, architecture and implementation as both product manager and sole engineer.
- Architected a governed RAG pipeline (FastAPI, Python, DeepSeek LLM, BGE-M3 multilingual embeddings, hybrid jieba / scikit-learn retrieval) across 8 retrieval generations, adding a need-to-skill router, citation-bound answer composition and a deterministic grounding guard that blocks ungrounded numbers, currencies and durations — raising useful-answer rate 67% (10/12 vs 6/12) over the prior architecture.
- Integrated the assistant with enterprise messaging (signed-webhook Open Agent protocol), the company HRIS (multi-entity OAuth, live leave balances and employee profiles) and Google Drive / Sheets / Docs for automated knowledge-base sync — 164 API endpoints across 30+ HR skills and 4 domain agents (HRBP, payroll, recruiting, shared services).
- Shipped a 38-page operations console (Next.js 15, React, TypeScript) covering A/B answer-version comparison with an offline LLM judge, engagement analytics, RAG review, policy-coverage and question-gap tracking, and an approval workflow — giving non-engineer HR staff direct control over knowledge quality.
- Engineered privacy and multi-tenant governance into the core: a PII gateway, self-only record boundary, 3-employee privacy floor on all analytics, fail-closed organization scoping between business entities and an immutable audit trail.
- Built the quality and observability backbone — 4,579 test cases across 275 modules (6,484 assertions passing), a pinned-frame regression harness and per-turn structured logging with trace-joined model-call telemetry — converting subjective quality complaints into reproducible, root-caused defects.
- Owned Dockerised deployment of an 8-service stack (API, isolated tenant API, Redis, 3 schedulers, web) with production overlays and one-command operator tooling; cut the API image 3.88 GB to 2.42 GB (-38%) and resolved OOM failures in the containerised test runtime.
- Contributed to the team AI workflow track alongside the solo build: improved Dify-based workflows for livestreamer tooling, and did the prompt engineering for copyright and sensitive-content detection on social media posting.
- Led the build-vs-buy evaluation of commercial and open-source RAG platforms (Onyx, RAGFlow) and delivered the management brief behind the decision to migrate onto Onyx — trading a bespoke pipeline for a maintained platform as the codebase grew, cutting long-term maintenance load without losing the governance the assistant depended on.

Lifegame — AI Engineer

Dec 2025 – Feb 2026

- Built AI-driven localization and dubbing workflows for short-form film and video, improving translation quality and consistency while reducing manual editing effort.

- Set up Google Cloud infrastructure and orchestrated multi-step localization and content-generation pipelines with Google Cloud Workflows; managed releases and version control in GitLab.
- Delivered avatar-based ad creation and AI-generated script workflows, including automated script generation from Google Play Store content for marketing use cases.

First Watch OT Cybersecurity — Machine Learning Engineer

May 2025 – Dec 2025

- Architected and implemented an end-to-end ML pipeline for OT network anomaly detection across traffic volume, protocol behaviour and communication patterns, as the sole ML / AI engineer alongside a senior developer and a cybersecurity specialist.
- Automated ingestion and transformation of OpenSearch telemetry into structured time-series datasets, with preprocessing and feature engineering to improve anomaly signal quality across packet, byte and protocol metrics.
- Integrated LLM components to generate analyst-ready anomaly summaries and explanations, improving triage and reporting efficiency.
- Built Streamlit dashboards for anomaly investigation, trend analysis and system-health visibility.

Yanmar Singapore — Data Analyst Intern

Sep 2024 – Mar 2025

- Built and compared time-series forecasting models (LSTM, Prophet, XGBoost, ARIMA) to predict product sales and service demand across multiple product series.
- Cleaned, merged and analysed business datasets to extract KPIs, delivering an end-to-end analytics project independently from data preparation through evaluation and presentation.

EDUCATION

DigiPen Institute of Technology Singapore — B.Sc. Computer Science in Real-Time Interactive Simulation, Minor in Mathematics

Singapore Polytechnic, Singapore Maritime Academy — Diploma in Marine Engineering