

Zhen Ee Wong

AI Systems & Backend Engineer

✉ hi@zhen.ee ☎ +65 86990919 📍 Singapore 🔗 zhen.ee 🌐 linkedin.com/in/zhenee

SUMMARY

Backend engineer with 8 years of experience building high-throughput proxy infrastructure and production AI systems: local LLM inference, computer-vision QA, and RAG pipelines. Tech lead for a team of 4 engineers, owning the architecture of an autoscaled reverse-proxy platform serving 10M+ requests/day at 75% lower infrastructure cost. Currently designing a closed-loop debugging system that converts verified fixes into reusable knowledge, cutting recurring manual work. Ships production systems from zero to one.

EXPERIENCE

Senior Software Engineer

10/2022 – Present

Vanguard Software

Singapore

- Led a team of 4 engineers, overseeing system architecture, technical direction, pull request reviews, and mentoring in debugging and skill development. Managed reverse proxy platform using OpenResty edge infrastructure.
- Architected a resilient multi-layer reverse proxy stack (OpenResty/Lua, Couchbase, TypeScript) handling 10M+ requests/day on a GCP fleet behind Cloud Load Balancing.
- Engineered load-testing infrastructure to sustain 60,000 concurrent TCP connections at full throughput with minimal latency via Linux kernel and NGINX kernel-level tuning.
- Reduced per-node infrastructure cost by 75% by optimizing from high-RAM VMs to 2 vCPU / 2 GB nodes, increasing connection capacity.
- Established horizontal autoscaling on GCP Managed Instance Groups with real-time alerts sent to Discord, enabling immediate visibility into scaling events.
- Automated client onboarding with a service generating per-client proxy configs from a generic template, ensuring scalability and consistency.
- Developed a proxy fleet rotation service for dynamic upstream IP management, maintaining uptime across distributed nodes.
- Created an automated visual QA system with custom YOLO models and local LLM to validate proxied sites.
- Built an OpenSearch-based log aggregation service to verify deployed fixes, bridging the gap between deployment and error resolution.
- Designed a self-learning proxy debugger using HAR diff engine and LLM + RAG pipeline for automated config patch proposals and validations.
- Integrated Apache Kafka for event-driven architecture for real-time data processing.
- Managed SQL Server, MySQL, Couchbase, and Elasticsearch for improved storage and data integrity.
- Authored architectural documentation, technical specifications, and drove code reviews and continuous improvements across the team.

Software Developer

05/2018 – 09/2022

Domain Connect

Malaysia

- Led full-lifecycle development for client platforms including online car auctions, LIMS, HR, POS, RFQ, and invoicing systems.
- Built auction bidding flows & admin verification workflows, enhancing platform security and curbing fraudulent listings.
- Engineered LIMS features for workload scheduling, equipment QC, and report data compilation/approval.
- Mentored junior developers and led code reviews to raise overall engineering standards.

Backend Developer

Hermo Creative

09/2017 – 05/2018

Malaysia

- Enhanced and extended a production warehouse management system: automatic bin location allocation, Levenshtein-based bin replenishment matching, barcode/QR scanning, cross-dock management, stock take and cycle counts, static/dynamic slotting, and RFID tracking.
- Built an integrated e-commerce channel letting outlets raise sales orders directly against warehouse stock; system improvements contributed to a 70% boost in warehouse KPIs and 50% reduction in operating costs.

TECHNICAL SKILLS

Languages

Python, TypeScript, Lua, PHP, Java, JavaScript, C/C++ (llama.cpp internals)

Data & messaging

Couchbase, Elasticsearch / OpenSearch, Microsoft SQL Server, MySQL, Apache Kafka, Liquibase

Frameworks & delivery

Spring Boot, Laravel, automated QA pipelines, CI / auto-deploy, GitHub Apps, SaaS licensing automation

Infrastructure

OpenResty / NGINX, distributed reverse proxies, Docker, GCP (Managed Instance Groups, Cloud Load Balancing, Cloud Functions), AWS (Lambda), Linux performance tuning, REST/gRPC

AI / ML

Local LLM inference (llama.cpp, GGUF quantization), RAG design & evaluation, computer vision (YOLO training), static analysis (Python AST)

SELECTED PROJECTS

EU AI Footprint Scanner

Argus Intelligence

AST-based static analysis engine that scans Python codebases for AI/ML dependency usage and maps libraries to EU AI Act risk tiers. Shipped as a commercial GitHub App with automated licensing and payments via Lemon Squeezy, built end to end from engine to monetization.

Aegis-RAG

Modular framework for benchmarking retrieval strategies (dense vector vs. BM25/keyword vs. hybrid with rerankers) to select optimal context-injection pipelines before LLM prompting.

Local LLM inference benchmarking

Empirical quantization trade-off analysis on consumer hardware (16 GB RAM) with llama.cpp and GGUF models (Qwen, Llama variants); measured logic accuracy vs. speed (~50 tokens/sec on Qwen 4B) and characterized memory/swap ceilings for 8B+ quantizations.

EDUCATION

Bachelor of Information Technology (Hons), Software Engineering

Southern University College

2014 – 2017

Johor Bahru,
Malaysia