

LIN ZHENMING

elz.work22@gmail.com | +65 8124 4215 | [@whatelz.ai](https://whatelz.ai) | [GitHub](#) · [LinkedIn](#) · [Medium](#)

AI engineer / AI product manager who ships production systems end-to-end. Now on my second AI Engineering internship at Prudential, where I led a cost-optimization initiative proving an ~80% token-cost cut on a production AI dashboard for 5,000 consultants, validated through formal AI-safety evaluation. Co-founder of three shipped AI products under AI System Resources. Final-year Fintech student at SIT — cGPA 4.69 (top of cohort), ASEAN Scholar, 6 hackathon podiums, and 1 of 25 in GFTN's inaugural Next Gen Leaders cohort.

RELEVANT WORK EXPERIENCE

AI Engineering Intern · Prudential Assurance Company Singapore

Jan 2026 – Aug 2026

- Led a production token-cost-optimization initiative for the Agent Performance Dashboard — the system driving the majority of Prudential's LLM spend, generating daily AI insights for 5,000 consultants — end-to-end, largely independently after my supervisor moved to spearhead another project
- Demonstrated an ~80% token-cost reduction through controlled experiments (sample sizes 10–64), achieved by migrating GPT-4.1 → Gemini-3.1-flash-lite, splitting prompts into a system/user structure to exploit explicit caching, and optimizing input-data formatting; presented the cost case to product owners
- Validated every change against business and safety bars — ran regression tests so generated insights stayed acceptable to stakeholders, and re-ran a four-dimension Ethical AI evaluation (AI Safety, Language, Actionability, Faithfulness) to confirm the model-family migration preserved insight quality
- Drove changes through enterprise governance — documented via Jira tickets, GitHub PRs, unit tests, and UAT, and presented and defended solutions to the Change Approval Board (CAB), AI Working Group (AIWG), and Local Solution Review Committee (LSRC)
- Prototyped a model-agnostic rebuild of the Product Knowledge Chatbot — after the Head of Department signalled a vendor-agnostic direction, independently re-implemented the chatbot on LangChain (vs. the original Google ADK) to prove the architecture on a different backend and models

AI Engineering Intern · Prudential Assurance Company Singapore

May 2025 – Aug 2025

- Evaluated the Product Knowledge Chatbot through UAT — assessed question–answer pairs against quality bars to certify it was production-ready for financial consultants
- Integrated tool-calling (MCP) and a GraphQL data layer into the chatbot, and implemented PII masking to meet compliance requirements for handling customer data
- Built an LLM-as-a-Judge evaluation framework for the Agent Performance Dashboard, scoring LLM-generated insights across four dimensions — AI Safety, Language, Actionability, and Faithfulness
- Evaluated and validated the dashboard's data pipeline, delivered by Bain — owning QA and handover so the team could maintain it in-house after the engagement closed

Product Management Intern · AsiaVerify

May 2024 – Aug 2024

- Designed request/response API contracts for the core compliance-data product and onboarded SE Asian registry data (Indonesia, Philippines); shipped continuous monitoring as a standalone AML feature

Data Science Intern · Setel Ventures

Jan 2023 – Jun 2023

- Rebuilt customer segmentation models (K-Means) for +25% and +30% business-value gains, built a semi-supervised gig-worker prediction model at 80% accuracy, and refactored the fraud-detection pipeline to cut runtime 20%

PROJECTS

Products of AI System Resources (co-founded)

ATLAS – Agentic Trading System · Capstone, SIT (atlas-broker.vercel.app)

- Designed a domain-specific language ("Ticket Logic") that represents trading strategies as inspectable, versioned artifacts — relocating the AI from placing individual trades to authoring and refining entire strategies, each executable deterministically with no LLM in the trade loop
- Built a model-agnostic MCP server (19 tools) that lets any AI client (Claude, ChatGPT) extract strategies from arXiv research papers, backtest them, and iteratively improve them — driving server-side inference cost to zero by routing reasoning through the user's own AI subscription
- Ran a controlled falsification study across paper-derived strategies: distillation flipped four losing strategies to profit, while an identical methodology could not rescue a standard retail indicator — isolating signal quality as the variable and evidencing that paper-extracted strategies carry real edge
- Re-architected the system after user testing (10 participants) surfaced that the original advisory dashboard was passive and costly — collapsing a five-service stack (Python/FastAPI, MongoDB, Inngest) to a streamlined Next.js + Supabase deployment and cutting inference cost to zero

DOUBLELEAD – AI-first Lead Manager (doublelead.io)

- Architected an operator-first CRM that owns the full chain — capture → pipeline → follow-up → invoicing — on one shared context object, with AI enrichment and automated follow-up at each stage
- Validated in real use: dogfooded daily by the founding team — including a co-founder running his company’s (PLC Automation) entire sales pipeline on it — and in active pitch to two insurance-agency leaders
- Built 5 cloud MCP servers exposing 46 tools on a shared OAuth 2.1 + PAT factory — any Claude-native surface connects to the full CRM with a single token
- Stack: Next.js 16, Supabase, Inngest, Clerk, Stripe, Groq

EMDEE – AI-native context layer for LLMs (emddee.tech)

- MCP server over a plain-markdown vault (single source of truth), with a Postgres-materialized edge graph for fast multi-hop retrieval and multi-user sharing (OAuth, ACL, hierarchy cascade)

EDUCATION

Bachelor of Science in Applied Computing (Fintech)

Aug 2023 – Aug 2026

Singapore Institute of Technology

cGPA: 4.69 / 5.00 (Top in Cohort) · ASEAN Scholarship 2023 – 2026

HACKATHONS

25+ competitions entered · 6 podium finishes

Overall Champion · Hackomania 2026

- Built SAIVERS — an AI-powered energy intelligence platform for SP Group — and won Overall Champion at one of Singapore’s largest student-run hackathons
- Engineered a real-time energy anomaly detection system using LangGraph multi-agent orchestration, surfacing actionable cost-reduction insights from live grid data across a 24-hour build sprint

1st Runner-Up · Pan-SEA AI Developer Challenge 2025

- Built an AI-powered scam-detection browser extension on SEA-LION — Singapore’s first open-source Southeast Asian language model — to flag financial fraud, phishing, and deceptive content in real time
- Placed 1st Runner-Up against developer teams from across Southeast Asia, recognised for the native-language SEA-LION integration and the immediacy of the use case to the regional scam epidemic

2nd Runner-Up · SingHacks 2025

- Built EZBIZ, an agentic payment framework on x402 + Hedera enabling AI agents to transact autonomously on-chain, with the full payment-authorization architecture built and demoed in 24 hours

LEADERSHIP EXPERIENCE

GFTN Next Gen Leaders

Nov 2025 – Present

- Selected as 1 of 25 in GFTN’s inaugural 2025 Next Gen Leaders cohort — chosen from a global pool spanning academia, founders, and builders across the fintech and digital finance ecosystem

President · Fintech Student Council, SIT

Jan 2024 – Dec 2024

- Led an 8-person executive committee delivering academic, welfare, and industry programming for the fintech cohort; set strategic direction and managed cross-functional operations across the full academic year
- Initiated and secured the council’s flagship industry partnership with Hooked Protocol, a Web3 education platform — sustained and renewed across 3 collaborations over 3 years

Secretary · SIT Developer Club

Jan 2024 – Dec 2024

- Ran operations, scheduling, and communications across the club’s 6-person core committee

SIT Ambassador

Jan 2024 – Dec 2026

- Represented SIT at public-facing institutional engagements, including ministerial dialogues

SKILLS

AI/ML: RAG, MCP servers, multi-agent orchestration (LangGraph), vector search (Azure AI Search, pgvector), knowledge graphs, model migration, caching · Gemini, Groq, Google ADK

Languages: Python, TypeScript, JavaScript, Java, C, SQL, HTML/CSS

Frameworks: FastAPI, Next.js

Infrastructure: Supabase, MongoDB, Inngest, Vercel, Databricks, Clerk, Stripe