Shawn YAP Zheng Yi

shawnyzy1@gmail.com shawnyap.dev LinkedIn GitHub

Education

Nanyang Technological University, Singapore **Aug 2021 – May 2025**

Bachelor of Computing (Hons), Computer Science

University of Stuttgart (Universität Stuttgart), Germany **Mar 2024 – Aug 2024**

Semester Exchange and German Immersion Programme, Computer Science

Experience

BeyondEdge **Jul 2026 – Present**

Solution Architect, AI Systems *Singapore*

- **Property AI Assistant:** Led the build of a market-leading property agency's consumer-facing assistant, architecting its orchestration pipeline across 9 intents and co-developing backend API endpoints, with scope gating, conversational state management, and memory-aware query rewriting
- Engineered tiered retrieval across 4 Qdrant collections, falling back to partial to semantic matches, with geocoded proximity filters, user-confirmed search slots, and API-sourced numeric facts for grounded property recommendations
- Fine-tuned its Qwen3.5-9B model with LoRA on ∼1k client-curated conversations, lifting reply fidelity 13% over base
- **Enterprise Agentic AI Platform:** Led the build and production deployment of a multi-tenant LLM agent platform for an arm of a major global investment group, with LiteLLM model routing, per-team token budgets, policy-gated MCP tools, and data-classification controls on GKE
- Prototyped a WhatsApp second-brain agent that distils free-text feedback into persistent memory over a 560-page LLM-wiki, productionising it into configurable department-level agents non-technical staff can tune without code
- Grounded agent briefs with per-claim citations to archived sources, prompt guards against forged citations, and watchdog alerts replacing silent run failures

FCS Solutions **Jul 2025 – Jul 2026**

AI Engineer *Singapore*

- Productionised an ASR–LLM–TTS pipeline with tool routing, schema validation, confidence gating, and telemetry
- Fine-tuned Whisper with LoRA, reducing WER from 40% to 5% and improving domain-specific intent recognition
- Built and fine-tuned a 200+ class intent-classification system with calibrated confidence thresholds for production routing
- Built schema-aware NL-to-SQL agents with constrained decoding, query validation, retries, and unsafe-query blocking

Home Team Science and Technology Agency (HTX) **Oct 2024 – Jun 2025**

Data Scientist & Machine Learning Engineer Intern *Singapore*

- Led R&D on audio deepfake detection using speech embeddings and prosodic features, improving EER by 30%
- Built reproducible, seed-fixed benchmarks, evaluating via EER, ROC-AUC, precision-recall, and failure-mode analysis
- Designed denoising and augmentation pipelines, sweeping EER across SNR levels to characterise low-SNR robustness
- Prototyped LangChain LLM and RAG workflows for summarisation and contextual retrieval over internal documents

Projects

Survivor: Evolutionary Multi-Agent Trading System | *Python, PostgreSQL, Docker, LLMs* **Jun 2026 – Present**

- Built a system that breeds, tests, and retires trading strategies by genetic search over 16 families and 8 regimes
- Scored 5k+ point-in-time forecasts against realised outcomes with Brier calibration, demoting overconfident strategies
- Capped the LLM reviewer to raising risk only; unbounded, it rejected all 32 trades (vs. 11) and halted trading
- Ran self-hosted Qwen inference under schema-constrained decoding with Pydantic validation of every model output
- Eliminated look-ahead bias structurally so backtests cannot access future data, verified over 676 backtest sessions

FYP: Transformer-Based Actor Classification & Sentiment Analysis | *PyTorch, BERT* **Jan 2024 – Dec 2024**

- Fine-tuned BERT for multi-label POFMA actor classification over 512-token chunks, at 0.847 accuracy and 0.781 F1
- Built scraping and PDF pipelines with contextual feature engineering, plus LLM sentiment analysis of public reaction

Skills

Machine Learning: PyTorch, Hugging Face Transformers, TensorFlow, LoRA/PEFT, Fine-tuning, BERT, Qwen, Whisper, NLP, ASR, Embeddings, Dataset Curation

Agentic AI & LLM Systems: Tool Calling, Tool Routing, MCP, RAG, Hybrid Retrieval, Qdrant, Structured Outputs, Constrained Decoding, NL-to-SQL, Prompt Engineering, LangChain, OpenClaw

LLM Reliability & Evaluation: Guardrails, Confidence Gating, Schema/Execution Validation, Retry Logic, Calibration, Threshold Tuning, WER, EER, F1, ROC-AUC, Brier Score, Error Analysis, Benchmarking

Data Science: pandas, NumPy, SQL, Feature Engineering, Statistical Analysis, Experiment Design, Backtesting, Point-in-Time Evaluation

AI Engineering: Python, Async Python, FastAPI, REST APIs, SQLAlchemy, PostgreSQL, Docker, Git, AWS, GCP, Azure, Pytest, Automated Testing, Model Deployment, Claude Code

Certifications: DeepLearning.AI TensorFlow Developer Specialization; Google Advanced Data Analytics Certificate

Languages: English (Native), Mandarin Chinese (Native), German (CEFR B1), Cantonese (Working Proficiency)