Bruce Cheng

AI Application Engineer | Enterprise AI Architecture & LLM Systems
Hsinchu, Taiwan | bruce.zheng0814@gmail.com | brucelcz.com | linkedin.com/in/bruce-lcz | github.com/bruce-lcz

Experience

Information Engineer (Generative AI / LLM Projects) Aug. 2025 – Present
PixArt Imaging Inc. *Hsinchu, Taiwan*

- Independently architect, develop, and deliver internal GenAI / LLM systems end-to-end, partnering with R&D, Legal, HR, and engineering teams to assess existing workflows and evaluate technical feasibility.
- Designed a shared internal GenAI workspace and LLM Gateway (vLLM, OpenAI-compatible APIs, model routing, API key governance, LLMOps telemetry) for coding, document review, and technical analysis.
- Delivered internal analytics and document systems: natural-language SQL (NL2SQL) restricted to read-only manager scope with audit logs, multimodal IC datasheet review, and page-cited ChromaDB RAG.
- Built legal, patent, meeting, and engineering tools via Prompt Engineering (clause extraction, staged review, validation), OCR/ASR, and Python — freeing R&D and legal/admin teams from manual proofreading.
- Drove GenAI enablement, knowledge transfer, and LLM optimization via telemetry and 20+ 1-on-1 consultations, authoring company-announced Prompt Engineering, OpenCode, context-compression, and KV-cache (TTFT) guides.

AI Engineer / Project Leader Aug. 2021 – Aug. 2025
AUO Corporation - Advanced Manufacturing Center *Taichung, Taiwan*

- Served as Project Leader (2023–2025), leading engineers and interns from scoping and feasibility analysis to fab deployment.
- Evolved an Azure OpenAI–powered manufacturing assistant from RAG into a LangGraph tool-using agent workflow with SQL/API access and process-knowledge reasoning.
- Improved yield by 1.43%, reduced defects by 12.62%, and saved ≈300 man-hours/mo via Golden Path optimization & LIME.
- Cut defect RMSE by 20% and saved ≈90 man-hours/mo via PyTorch LSTM & custom trend-consistency loss (patent).

Research Intern Feb. 2019 – Jul. 2019
Academia Sinica - GIS Center *Taipei, Taiwan*

- Built a web-based Taiwan address normalization workflow; awarded Best Student Paper at TGIS Conference.

Projects

Work-Hour Review & Insights Platform | *FastAPI, NL2SQL, Data Security, Web UI* 2026 – Present

- Built an offline manager review web app and NL2SQL Q&A over encrypted legacy HR data with historical org tracking, restricting users to read-only queries within their own department scope with audit logs.

Multimodal Technical Document Review | *OpenAI-compatible API, VLM, Prompt Eng., MCP, FastAPI* 2026

- Reviewed semiconductor datasheets (text, register tables, diagrams) via two-pass prompt engineering, few-shot exclusion rules, prompt/model A/B evaluation, and Web, CLI, and MCP interfaces.

LLM Gateway & Operations Platform | *LLMOps, Model Routing, Metering, Docker, PostgreSQL* 2026

- Unified OpenAI-compatible model access, API key governance, routing, and usage metering with read-only GPU/usage telemetry dashboards and automated reports.

Evidence-Grounded Technical Knowledge | *VLM, RAG, ChromaDB, Pydantic* 2026

- Designed a VLM-assisted, human-reviewed RAG workflow with Pydantic schema validation, grounding checks, ChromaDB vector indexing, and exact page-level citation traceability.

GenAI Enablement & LLM Optimization | *vLLM, Prefix Caching, Prompt Eng., OpenCode* 2026 – Present

- Turned LLM telemetry and 20+ 1-on-1 consultations into company-announced knowledge-transfer guides on Prompt Engineering best practices, OpenCode, context compression, and KV-cache / TTFT optimization.

Technical Skills

GenAI & LLM Systems: OpenAI / Azure OpenAI & OpenAI-compatible APIs, On-Prem LLM Serving (vLLM), Prompt & Context Engineering, RAG, Vector DBs (ChromaDB), Agents (LangGraph, MCP), Multimodal / VLM, OCR / ASR
AI Architecture, LLMOps & Security: Enterprise AI Architecture, LLMOps / MLOps Deployment, AI Security & Data Privacy (Offline Deployment, Access Control, Audit Logs), Feasibility & Adoption Strategy, Project Leadership, Knowledge Transfer
Backend & Data Engineering: Python, FastAPI, REST APIs, SQL, PostgreSQL, Docker, ETL, Databricks, AI-Assisted Web UI Delivery (React / TypeScript)
Model Optimization & Applied ML: LLM Inference Optimization (KV Cache / Prefix Caching / TTFT), PyTorch, Custom Loss, LSTM, LIME

Education

National Yunlin University of Science and Technology Yunlin, Taiwan
Master of Science in Industrial Engineering and Management *2019 – 2021*

National Yunlin University of Science and Technology Yunlin, Taiwan
Bachelor of Science in Industrial Engineering and Management *2015 – 2019*