

SAKCHAI SAEHOEI

Bangkok | yangshiyuang@gmail.com | +66900761399 | [LinkedIn: Sakchai Saehoei](#) | [GitHub: SakchaiSaehoei](#)

PROFESSIONAL SUMMARY

AI/ML Engineer and AI Research Engineer building LLM routing frameworks, AI platforms, and observability systems. Experience spans model evaluation and quantization, HPC inference, NLP/translation research, and agentic business intelligence with secure data infrastructure.

EXPERIENCE

AI Research Engineer | Core Artificial Intelligence Research Team

Pathum Thani, Thailand

NECTEC (Thailand National Electronic and Computer Technology Center)

Aug 2024 - Present

LLM Routing & AI Platform Engineering

- Developed a Python LLM routing framework with LangGraph, LiteLLM, and Pydantic, implementing 10 procurement-agent policies across three model tiers with capability checks, context limits, structured-output validation, and bounded escalation.
- Deployed a centralized LiteLLM gateway using Docker Compose, PostgreSQL, and Caddy HTTPS; smoke-tested seven model endpoints covering text generation, Thai-language embeddings, and image-based OCR.
- Implemented Langfuse and OpenTelemetry tracing for workflow ancestry, token usage, latency, and cost; verified inference continued during a controlled 30-second tracing-service outage.
- Built an automated quality gate using pytest, Ruff, mypy, and uv, with 173 passing tests recorded at the integration checkpoint covering routing, output validation, failure handling, and telemetry privacy.
- Provisioned PostgreSQL, Redis, Neo4j, and Qdrant with individual developer credentials; migrated application and database access behind self-hosted NetBird with private DNS, trusted HTTPS, and verified administrator/developer access boundaries.
- Deployed Prometheus and Grafana monitoring with two dashboards and 26 panels covering server resources, gateway activity, and service health.

Research Engineer - LLM Infrastructure & HPC Systems (LANTA Supercomputer)

Aug 2024 - Present

- Built reproducible LLM fine-tuning and inference pipelines using **Apptainer**, **DeepSpeed**, **SLURM**, **vLLM**, and **Ollama** for 4B-70B+ models on Thailand's LANTA HPC cluster.
- Benchmarked **Sarvam-30B** baseline and AWQ 4-bit variants across 4,683 MATH-500 and MedMCQA test cases, reducing evaluation time from 145 to 116 minutes (1.25x) while quantifying 4.4 and 18.9 percentage-point accuracy trade-offs.
- Created modular SLURM templates and FastAPI/WebSocket inference APIs, reducing researcher onboarding from days to hours and enabling consistent experiments across multi-node clusters.

AI & Data Engineering Consultant - Agentic AI for Business Intelligence

Jan 2026 - Jun 2026

- Integrated four 1C ERP data sources and remote Microsoft SQL Server systems into an on-premises analytics platform using **Airflow**, **PostgreSQL**, **Docker Compose**, **Metabase**, and **Power BI**, supporting automated reporting and natural-language BI.
- Developed resilient, idempotent ETL workflows with incremental ingestion, retry handling, and **SQL Server MERGE** upserts for production, sales, inventory, and employee-performance data.
- Built local LLM/FastAPI and **LangGraph**, **LangChain**, and **MCP** Q&A workflows to deliver business answers through dashboards, email, and messaging workflows.

PROJECTS

Selected NLP & Machine Translation Systems

- Built a Thai-Chinese lexical-constraint pipeline using Thai segmentation, Jieba, BPE/subword-nmt, fast_align, and grow-diag-final-and for low-frequency, POS-filtered terminology; evaluated offline Qwen-14B/72B GGUF models through llama.cpp in restricted HPC environments.
- Designed a multi-adapter DIP translation system with **Lorax**, routing 45 classes into 7 LoRA adapters on google/gemma-3-4b-it; achieved CER 0.3740 and BLEU 0.4247 with deterministic domain routing.

Constraint-Based NMT for WIPO Product and Service Descriptions

iSAI-NLP 2026 Research Paper

- Designed a six-stage evaluation of terminology-constrained English-to-Thai WIPO translation over **2,785 descriptions across 45 Nice Classification classes**, comparing direct prompting, class context, fixed class-conditioned examples, FTS5/BM25, MiniLM/FAISS, and BGE-M3/FAISS retrieval.
- Evaluated **Gemma-3-1B-IT, Qwen3-1.7B, and Mistral-7B-Instruct** with vLLM on one NVIDIA A100 GPU; dense retrieval reached 0.528-0.549 micro BLEU versus 0.046-0.201 static settings and reduced mean CER to 0.265-0.301 under failure-aware scoring.

EARLIER EXPERIENCE

Software Engineer Intern, SmartCare

Bangkok, Thailand

Huawei Technologies (Thailand) Co., Ltd.

Jan 2024 - May 2024

- Built Flask/Plotly dashboards and IQR anomaly detection for 10+ application traffic streams; automated report generation, improving marketing reporting efficiency by 17x.

Enterprise Storage Engineer Assistant Intern

Bangkok, Thailand

Huawei Technologies (Thailand) Co., Ltd.

May-Jul 2023

- Supported enterprise-storage infrastructure involving RAID, NAS, SAN, and storage migration.

PUBLICATIONS

Constraint-Based Neural Machine Translation for WIPO Product and Service Descriptions. Manuscript under review, iSAI-NLP 2026.

SKILLS

AI/LLMs: PyTorch, Hugging Face, DeepSpeed, LoRA, vLLM, Ollama, llama.cpp, Lorax, AWQ, GGUF, LiteLLM, LLM evaluation

Agentic AI & Data: LangGraph, LangChain, MCP, Pydantic, Airflow, PostgreSQL, SQL Server, Redis, Neo4j, Qdrant, ETL, Medallion Architecture

Infrastructure: Docker Compose, Apptainer, Linux, Bash, SLURM, Kubernetes, ArgoCD, GitHub Actions, Caddy, NetBird, FastAPI

Programming & Quality: Python, SQL, JavaScript/TypeScript, C, Node.js, React, Next.js, pytest, Ruff, mypy, uv

NLP & Analytics: NMT, lexical constraints, BPE, fast_align, RAG, FAISS, BGE-M3, BM25, BLEU/CER, Power BI, Metabase, Plotly

Observability: Langfuse, OpenTelemetry, Prometheus, Grafana

LANGUAGES

Thai - Native; **Chinese** - Professional proficiency (HSK 6); **English** - Professional working proficiency (IELTS Academic 7.0); **Burmese** - Basic

CERTIFICATIONS & AWARDS

Certifications: DeepLearning.AI Fast & Efficient LLM Inference with vLLM (Sep 2026); CompTIA Security+ (2022); EC-Council Security Specialist (2022); Oracle Cloud Associate (2021); IBM AI & Development (2023)

Awards: Top 10% Super AI Thailand (2024); Winner - Berkeley SkyDeck Hackathon (2023); Thai Science Ambassador (2023); Runner-Up Microsoft RE:THINK (2022); 2nd Place THNG Camp #10 (2020)

EDUCATION

Bangkok University | BSc in Computer Science

Bangkok, Thailand

First Class Honors - GPA: 3.8/4.0

Aug 2020 - May 2024