

TRUONG BINH THUAN

Ho Chi Minh City, Vietnam

Web: thuanbt.id.vn

Email: sunbv56@gmail.com

GitHub: github.com/sunbv56

Personal Profile

AI and full-stack engineer promoted from AI Intern to IT Engineer at Tiximax after two months, building production LLM applications from model fine-tuning (LoRA, RLHF) to a multi-tenant SaaS on Next.js, TypeScript and PostgreSQL. Extended Claude and Gemini content pipelines, implemented a multilingual SEO scoring engine, and built internal AI agents on Model Context Protocol (MCP) servers. Seeking to build reliable, well-tested AI products end to end.

Professional Experience

Tiximax JSC | Jun 2026 – Present

IT Engineer | Aug 2026 – Present

- Core engineer on **TixiGen**, a multi-tenant AI SEO content CMS built on Next.js, TypeScript and PostgreSQL; second-largest contributor, with **850+** commits, **41** tickets and **50+** merged MRs in 2.5 months.
- Led workspace-isolation hardening in the SaaS conversion: audited **445** API routes and closed **43** cross-workspace read/write isolation gaps, backed by **368** test cases (341 automated); shipped multi-workspace accounts, per-workspace subdomains, invites and public sign-up.
- Built a customer-facing content API after benchmarking 10 headless CMSs against the **OWASP API Top 10**: scoped, expiring per-workspace API keys with rotation and a 3-day grace period; designed subscription tiers priced from measured LLM cost.
- Extended the LLM content pipeline: made the **Claude API** the primary provider, generated **Gemini** response schemas from Zod, and added embedding-based context retrieval with LLM re-ranking; grounded generation and scoring in a verified-facts set to block fabricated statistics.
- Implemented a **60+-criterion** SEO scoring rubric with a live in-editor scorecard and language-aware rules (Vietnamese, Japanese, Indonesian, Tagalog); its AI auto-fix cannot break criteria that already pass.
- Re-architected the **PostgreSQL** schema into a trigger-enforced content tree (149 → 137 tables) across **120+** migrations and **130+** test files; replaced string-interpolated SQL with parameter binding, closing SQL-injection vectors.
- Localized the entire UI into Vietnamese, English and Japanese (**5,804** keys per language, DB-backed, editable without redeploy) and built an interactive onboarding tour; cut editor SEO-analysis time from **53.6 ms to 22.4 ms**.

AI Intern | Jun 2026 – Aug 2026

- Developed an internal AI assistant platform on **Claude Code**: department skills over an Obsidian knowledge base plus **5 Python MCP servers** (50 tools: GA4, Search Console, revenue, mail), with role-based access per department; lead developer (72% of commits).
- Built **AutoDub Studio** solo, an automated video dubbing and localization pipeline (Whisper → Gemini translation → Edge-TTS → FFmpeg, optional Wav2Lip lip-sync) supporting **12+** languages, with desktop and FastAPI web UIs.
- Co-developed an AI video studio: shipped a multi-provider TTS voice studio (Edge-TTS, Piper, ElevenLabs), a usage-quota dashboard, and a GPU crash fix.
- Prototyped an intent classifier for the company landing-page chatbot using multilingual-e5 sentence embeddings served through Transformers.js.

Amazing Total Solutions JSC (Amazing Tech) | Apr 2026 – Jun 2026

AI Engineer Intern

[\[GitHub Repository\]](#)

Project: Mamba-Forecast-AD, anomaly detection via time-series forecasting

- Developed a self-supervised anomaly detection system using a **Hybrid Mamba** (Selective State Space Model) to achieve $O(N)$ linear complexity, achieving a POT F1-score of **75.65%** and a **14.1×** inference speedup over LSTM.
- Integrated an EMA-based **Series Decomposition** layer to decouple signals into low-pass *Trend* (via a Linear layer) and high-pass *Seasonal* (via the Mamba backbone) components.
- Engineered a **Physical Stats Head** incorporating 8 time-domain vibration metrics, reducing validation forecasting MSE by **5.56%** with minimal VRAM overhead.
- Deployed an Extreme Value Theory-based **Peak-Over-Threshold (POT)** calibration, achieving a False Alarm

Rate of **0.18%** and a calibration latency of **12.12 ms**.

Education

Sep 2021 – Sep 2025: Bachelor of Science in Computer Science

Ton Duc Thang University, Vietnam

GPA: 8.09/10

Relevant Coursework: Artificial Intelligence, Machine Learning, Deep Learning, Data Science, Data Structures & Algorithms, Probability & Statistics.

Key Research Interests: Machine Learning, Deep Learning, Time-Series Forecasting & Anomaly Detection, Speech Processing.

Projects & Research

Qwen3-ASR Fine-Tuning for Vietnamese Music | Mar 2026

[\[Hugging Face Space\]](#)

- Fine-tuned Qwen3-ASR (0.6B & 1.7B parameter models) using Parameter-Efficient Fine-Tuning (**PEFT/LoRA**) on a curated dataset of 50,000 Vietnamese song samples.
- Designed and optimized a high-throughput audio chunking pipeline to handle and train models on long audio sequences (3–10 minutes).
- Achieved a **12.57% absolute reduction in Word Error Rate (WER)** and a **3.88% reduction in Character Error Rate (CER)** on the 0.6B model compared to baseline performance.

AI Chatbot & Workflow Automation | Dec 2025 – Present

- Developed and deployed conversational chatbots using LLM APIs and RAG to automate consulting and sales customer support.
- Designed and built automated marketing workflows using n8n and Make.com to automatically generate marketing content, schedule postings, and handle customer comments/inquiries.
- Integrated multiple APIs and webhooks to connect chatbots with social media platforms (Facebook, Telegram) and databases for user engagement tracking.

ViLawT5-RL: RLHF for Vietnamese Legal QA | Nov 2024 – Feb 2025

[\[GitHub Repository\]](#)

- Developed a Vietnamese legal question-answering assistant by fine-tuning a ViT5-base model using Reinforcement Learning from Human Feedback (**RLHF**) and Proximal Policy Optimization (**PPO**).
- Improved chatbot response coherence, alignment, and contextual relevance, reducing model hallucination in domain-specific legal QA, achieving a BLEU score of **38.41%** and a ROUGE-L of **64.75%**.
- Optimized reward shaping and policy updates to stabilize PPO training on low-resource Vietnamese text.

Skills

Technical Focus: LLM Application Engineering, Agentic AI & Model Context Protocol (MCP), RAG, Prompt Engineering, Machine Learning, Deep Learning, NLP, Speech Recognition (ASR), Time-Series Forecasting & Anomaly Detection, Full-stack Web Development, Multi-tenant SaaS.

AI & ML: Python, PyTorch, Hugging Face (Transformers, PEFT/LoRA, Accelerate), Claude API, Claude Code (Agent Skills), Google Vertex AI (Gemini, Imagen), LangChain, Sentence Transformers, Vector Databases (Chroma/Pinecone), Scikit-learn, Pandas, NumPy.

Web & Data: TypeScript, JavaScript, Next.js, React, Tailwind CSS, Node.js, FastAPI, PostgreSQL, Drizzle ORM, Zod, SQL, MinIO/S3, REST & OpenAPI, i18n (next-intl).

Tools: Git, Vitest, Pytest, Playwright (browser automation), Linux, n8n, Make.com, LaTeX.

Languages: English (B1 Intermediate), Vietnamese (Native).