

ANH NGUYEN HOANG

+84 398 272 620 ◊ Hanoi, Vietnam (open to remote)

anhnh.4work@gmail.com ◊ [Homepage](#) ◊ [GitHub](#)

SUMMARY

LLM application engineer with two years of hands-on experience shipping LLM-powered products and autonomous agents, from a consumer AI companion serving 300,000+ users and roughly 100,000 messages per day to an open-source repository-documentation agent with 1,500+ GitHub stars. Strong background in agent design (planning, memory, tool calling, multi-agent orchestration), evaluation frameworks, and agent reliability, backed by publications at ACL, NAACL, and EMNLP. Comfortable owning a problem end to end, from ambiguous business need to measured, production-ready AI system.

SKILLS

LLM & Agents	Agent architectures (planning, long-term memory, tool calling, multi-agent orchestration), prompt engineering and optimization, RAG and vector search, Model Context Protocol (MCP) tooling and security, LLM fine-tuning (LoRA/full), evaluation harness and benchmark design, LLM-as-judge
Engineering	Python, PyTorch, Hugging Face Transformers, LLM APIs (OpenAI, Anthropic, open-weight models), REST APIs, SQL/NoSQL databases, vector stores, Docker, Git, Linux, CI and observability for AI systems
Domains	Conversational AI, AI for software engineering (code agents, program repair, code migration, vulnerability detection), continual learning, text-to-speech adaptation

EXPERIENCE

FPT Software AI Center	May 2025 – Present
<i>AI Research Resident (LLM agents for software engineering)</i>	<i>Hanoi, Vietnam</i>

- Designed and built **CodeWiki**, an LLM agent framework that generates holistic, repository-level documentation for large codebases across seven programming languages; open-sourced to 1,500+ GitHub stars and published at Findings of ACL 2026
- Built the agent's planning, hierarchical decomposition, and tool-calling pipeline for repository-scale context, plus an evaluation benchmark and LLM-as-judge framework to measure documentation quality
- Co-developed **RustPrint**, a documentation-guided agentic system that migrates C codebases to Rust, coordinating multiple specialized agents with automated build-and-test feedback loops
- Leading **CHAIR**, a controller that automatically improves agent harnesses through hypothesis generation, information-gain experiment selection, and controlled single-factor evaluation
- Collaborating with UIUC and Mila on agentic vulnerability detection and on hardening coding agents against MCP tool-poisoning attacks

Oraichain Labs	May 2024 – April 2025
<i>AI Engineer</i>	<i>Hanoi, Vietnam</i>

- Led development of **Cupiee**, an emotional AI companion platform that grew to 300,000+ users and approximately 100,000 messages per day
- Designed the conversational agent stack: persona and prompt design, long-term user memory, retrieval over conversation history, and safety and behavior guardrails; iterated on prompts and model choices against user-engagement and quality metrics
- Built autonomous multi-agent systems for social media engagement and crypto market analysis, integrating LLMs with external APIs, scheduled workflows, and data pipelines
- Owned production concerns including latency, cost per message, monitoring, and rapid experiment-deploy cycles

- Researched few-shot continual learning for language models, designing data-augmentation and generalization-preserving training methods; equal-contribution first author at EMNLP 2024 and NAACL 2025

SELECTED PROJECTS AND OPEN SOURCE

CodeWiki

1,500+ GitHub stars

Agentic framework for holistic documentation of large codebases; supports seven languages and ships with an evaluation benchmark for repository-level understanding

XTTSv2 Fine-Tuning for New Languages

200+ GitHub stars

End-to-end toolkit for adapting the XTTSv2 speech-synthesis model to new languages, covering data prep, tokenizer extension, training, and inference

AI for Real Action 2025, Champion (Team Converged)

January 2026

First prize among nearly 400 teams in Vietnam’s first national AI competition, delivering a production-style AI solution under competition time constraints

PUBLICATIONS

[1] **CodeWiki: Evaluating AI’s Ability to Generate Holistic Documentation for Large-Scale Codebases.** A. Nguyen Hoang, M. Le-Anh, B. Le, N. D. Q. Bui. *Findings of ACL 2026*. [Paper]

[2] **Beyond Repository Boundaries: Cross-Repository Graph Retrieval for Code Generation.** M. Le-Anh, N. Le Hai, Q. Tran, A. Nguyen Hoang, L. Ngo Van, B. Le, N. D. Q. Bui. *Findings of EMNLP 2026*.

[3] **RustPrint: Documentation-Guided Agentic Codebase Migration from C to Rust.** M. Le-Anh, A. Nguyen Hoang, B. Le, N. D. Q. Bui. *Preprint, 2026*. [Paper]

[4] **Mutual-Pairing Data Augmentation for Few-Shot Continual Relation Extraction.** N. H. Anh*, Q. Tran*, T. X. Nguyen*, et al. *NAACL 2025*. [Paper]

[5] **Preserving Generalization of Language Models in Few-Shot Continual Relation Extraction.** Q. Tran*, N. X. Thanh*, N. H. Anh*, et al. *EMNLP 2024*. [Paper]

* equal contribution.

EDUCATION

Hanoi University of Science and Technology

2020 – 2024

Bachelor of Science in Computer Science, GPA 3.75/4.0

Ranked 15th of 296 (Top 5%)

Techcombank Overseas PhD Study Sponsorship (up to USD 333,000)

April 2026

LANGUAGES

Vietnamese (native), English (professional working proficiency, IELTS 7.0)