

NGUYỄN NGỌC QUÝ

☎ +84 918 211 296 ✉ pipyakas@gmail.com 🌐 <https://quynn.dev> 📍 Hà Nội, Việt Nam

SUMMARY

Kỹ sư Phần mềm với hơn 5 năm kinh nghiệm về hệ thống C++ hiệu năng cao, công cụ Backend và hạ tầng AI Local. Chuyên sâu về triển khai vLLM Inference dạng container, xây dựng Custom Agentic Gateway, quy trình Tool-calling và hệ thống kiểm thử chịu tải quy mô lớn. Tối ưu hiệu quả độ trễ, khả năng xử lý đồng thời và hiệu suất lập trình trong môi trường làm việc phân tán.

EDUCATION

Đại học Bách khoa Hà Nội

09/2014 – 10/2021

Cử nhân in Global ICT (Chương trình Kỹ sư CNTT-TT Quốc tế / Việt - Nhật)

EXPERIENCE

Gameloft

03/2021 – Hiện tại

Lập trình viên Game (Hệ thống C++ & Hạ tầng AI)

- Hạ tầng LLM Local & Model Serving (06/2026 – Hiện tại): Triển khai hệ thống LLM Inference dạng container trên cụm multi-GPU; tối ưu vLLM (phân bổ bộ nhớ, KV-cache, độ dài context, xử lý đồng thời) và Speculative Decoding dựa trên bộ benchmark stress-test & tool-use tự động.
- LLM Gateway & Định tuyến Agentic: Xây dựng AI Gateway chuẩn giao thức Anthropic/OpenAI bằng TypeScript để định tuyến các công cụ lập trình AI (Claude Code, OpenCode) qua mô hình vLLM local và cloud; tích hợp chuyển đổi streaming protocol, xác thực client và đo lường tài nguyên sử dụng trên PostgreSQL.
- Tự động hóa phân tích Crash Log: Thiết kế nền tảng chẩn đoán Linux core-dump tự động (kết nối S3 đến WSL/GDB với hệ thống lưu trữ symbol); truy xuất stack trace theo ngữ cảnh, gom nhóm lỗi liên quan và ứng dụng LLM để tự động phân tích nguyên nhân gốc (root-cause) & đề xuất phương án sửa lỗi qua dashboard PostgreSQL.
- Hệ thống Review Code SVN tự động: Phát triển nền tảng review code linh hoạt (model-agnostic) cho SVN; tự động đánh giá mã nguồn định kỳ bằng LLM, phân loại phát hiện, quản lý trạng thái xử lý và gửi thông báo/hiển thị dashboard trên nền PostgreSQL.
- Kiểm thử chịu tải Backend: Thiết kế nền tảng giả lập hàng nghìn người chơi kết nối đồng thời để stress-test hệ thống game server và API sự kiện trước thời điểm phát hành.
- Lập trình C++ Client-Server & Game Systems: Phát triển các tính năng và hệ thống gameplay C++ hiệu năng cao trên cả Client và Server, đồng thời tham gia tối ưu và nâng cấp bản port trên Android.
- Phát triển Đa nền tảng & Tích hợp SDK: Đảm nhận triển khai tính năng trên codebase ActionScript legacy và dự án chuyển đổi sang Unreal Engine 5; tích hợp thành công các SDK backend lớn bao gồm Netflix SDK và SDK nhà phát hành khu vực.
- Hạ tầng CI/CD: Thiết kế và duy trì pipeline build & deploy tự động trên Jenkins cho dự án Sniper Fury, nâng cao độ ổn định và hiệu quả vận hành.
- Dẫn dắt Kỹ thuật & Mentoring: Chủ trì các buổi workshop kỹ thuật, tổ chức cuộc thi lập trình; trực tiếp hướng dẫn kỹ sư junior và chuẩn hóa tài liệu kỹ thuật bằng tiếng Anh thương mại.

TECHNICAL SKILLS

Hệ thống Cốt lõi & Kỹ thuật Backend: C++, Python, Node.js, REST APIs, PostgreSQL, Phát triển Nền tảng Android, Linux, GDB, Docker, Podman

Kỹ thuật AI & Hạ tầng LLM: Local LLM Inference, vLLM, Tối ưu hóa Model Serving, LLM Gateways, Tool Calling & Function Calling, Agentic Workflows, Claude Code, OpenCode, Claude Agent SDK

DevOps & Công cụ Chuyên nghiệp: Java, Unreal Engine 5, Jenkins CI/CD, Git / GitFlow, SVN, Jira