QUẾ XUÂN MẠNH

Hà Nội, Việt Nam
quemanhmcr@gmail.com | (+84) 967 584 126 | github.com/quemanhmcr | quemanhmcr.github.io

TÓM TẮT CHUYÊN MÔN

Kỹ sư AI với nền tảng cơ bản về Generative AI, Hệ thống Gợi ý (RecSys) và MLOps. Có năng lực xây dựng các hệ thống phức tạp từ nguyên lý cơ bản (First Principles), bao gồm các luồng truy vấn dữ liệu lớn và kiến trúc Transformer. Cam kết nghiên cứu chuyên sâu và vận dụng tư duy toán học để phát triển các giải pháp AI có khả năng mở rộng (Scalable) và tính ứng dụng cao.

KỸ NĂNG & CÔNG CỤ

AI & Thuật toán cốt lõi: PyTorch, Transformers Architecture, RecSys (Two-Tower, DCN), FAISS, Optimization.

Generative AI: Fine-tuning LLM (LoRA/QLoRA), RAG Pipelines, Prompt Engineering, Multi-Agent (CrewAI).

MLOps & Hạ tầng: Docker, Kubernetes, Terraform, ArgoCD, CI/CD, Vector Databases.

Ngôn ngữ lập trình: Python (Advanced), SQL, Bash Scripting.

DỰ ÁN TIÊU BIỂU

Hệ thống Gợi ý Âm nhạc Quy mô lớn (Large-Scale Music RecSys) **2025**

Nghiên cứu & Phát triển Độc lập *PyTorch, FAISS, Polars*

- Thiết kế quy trình gợi ý hai giai đoạn (Retrieval & Ranking) xử lý **450 triệu tương tác** trên tài nguyên phần cứng giới hạn. (Colab)
- Triển khai kiến trúc **Two-Tower** cho giai đoạn truy vấn, tích hợp **FAISS** để tìm kiếm vector tốc độ cao.
- Phát triển mạng **Deep & Cross Network (DCN-v2)** cho giai đoạn xếp hạng để tối ưu hóa chỉ số `played_ratio`.
- Tối ưu hóa bộ nhớ bằng **Memory Mapping** và Custom DataLoaders, giảm **90% RAM** tiêu thụ khi huấn luyện.

Tái cài đặt Kiến trúc Llama 3.2 Pre-training **2025**

Nghiên cứu viên Độc lập *Python, PyTorch, Distributed Training*

- Viết lại mã nguồn kiến trúc Llama 3.2 (1B tham số) từ đầu để kiểm soát toàn bộ luồng dữ liệu và cơ chế Attention.
- Tích hợp các kỹ thuật hiện đại: **RMSNorm, RoPE (Rotary Embeddings), GQA (Grouped Query Attention)**.
- Xây dựng pipeline dữ liệu hiệu năng cao sử dụng **MinHash LSH** để lọc trùng (deduplication) dữ liệu đầu vào.

Nền tảng MLOps End-to-End **2024**

Kỹ sư DevOps/MLOps *Terraform, Kubernetes, ArgoCD*

- Thiết lập quy trình **GitOps** tự động hóa triển khai mô hình với ArgoCD trên Kubernetes.
- Xây dựng hạ tầng dưới dạng mã (IaC) với **Terraform** trên AWS, đảm bảo tính nhất quán của môi trường.

GIẢI THƯỞNG & THÀNH TÍCH

Giải Nhì, Cuộc thi Sáng tạo Trẻ tỉnh Nghệ An **2022**

Dự án: Dịch máy Nơ-ron (NMT) cho tiếng dân tộc Thái - Kinh

- Phát triển mô hình Transformer dịch thuật low-resource language (ngôn ngữ ít tài nguyên).
- Tự xây dựng bộ dataset song ngữ và tối ưu hóa chiến lược Tokenization cho tiếng dân tộc.

Top 50 Chung khảo, Cuộc thi Data for Life **2022**

Dự án: Hệ thống phát hiện Deepfake đa phương thức

- Đề xuất giải pháp Multimodal sử dụng LLMs để phát hiện bất thường ngữ nghĩa và sự lệch pha âm thanh-hình ảnh.

NGHIÊN CỨU & ĐÀO TẠO ĐỘC LẬP

Nghiên cứu Độc lập về Khoa học Máy tính & AI **2024 – Hiện tại**

- Thực hiện lộ trình học tập trung vào các nguyên lý Machine Learning và ứng dụng thực tiễn.
- Reproduce (tái hiện) code từ các bài báo khoa học kinh điển: Attention Is All You Need, LoRA, DLRM.