Jimmy P. Le

(316) 518-1839 | lejp3@vcu.edu | Richmond, VA
linkedin.com/in/jimmy-le-vcu | github.com/jimmyphuocle | jimmyphuocle.com

Summary

ML/NLP engineer and PhD researcher specializing in Retrieval-Augmented Generation, Large Language Models, and biomedical knowledge graphs. 1+ year of hands-on LLM application development on open-source models with local inference, vector databases, and HPC. Designs RAG pipelines, diagnoses retrieval failures through rigorous ablation, and manages concurrent projects in higher education.

Education

Virginia Commonwealth University — Richmond, VA
Ph.D. in Computer Science — Natural Language Processing — *Jan 2026 – Present*
- **Advisor:** Bridget McInnes, PhD. **Focus:** explainable retrieval-augmented generation, literature-based discovery, knowledge graphs, biomedical question answering.

Virginia Commonwealth University — Richmond, VA
M.Sc. in Computer Science — *Dec 2025*
- **Coursework:** Advanced NLP, Machine Learning, Neural Networks & Deep Learning, Knowledge Discovery & Data Mining, High-Performance Distributed Systems.

Wichita State University — Wichita, KS
B.Sc. in Computer Science — *May 2023*

Research & Applied ML Engineering

Literature-Based Discovery for Explainable GraphRAG *(manuscript in preparation)* — Sept 2025 – Present
- Designed and deployed an end-to-end Retrieval-Augmented Generation (RAG) pipeline integrating open-source LLMs (Qwen3-14B, Llama 3.1-8B served via Ollama on local HPC nodes) with a biomedical knowledge graph; all inference runs on institutional infrastructure with no data egress to public APIs, preserving institutional data sovereignty.
- Built and managed a Neo4j vector database containing 452K+ nodes and 3.18M+ relationships, with triple-level embeddings generated via Clinical ModernBERT and native vector indexes enabling high-speed semantic retrieval across structured clinical knowledge.
- Engineered a parallelized triple filtration system (Linking Term Count, 32-worker) reducing pipeline runtime from 72+ hours to under 2 seconds on SLURM-managed GPU-H100 nodes, with checkpoint/resume support for fault-tolerant jobs.
- Discovered a systematic "context distraction" effect where retrieved triples degraded LLM accuracy below zero-shot (0.6165 vs. 0.597); isolated cosine-similarity misalignment as the root cause via controlled ablations and LLM-as-Judge evaluation across 4,183 MedMCQA questions.
- Applied PEFT via LoRA fine-tuning (Qwen3-14B, LitGPT, multi-GPU) and authored reproducibility tooling (environment isolation, run tagging, pipeline READMEs) enabling a 4-person research team to collaborate on shared HPC infrastructure.

Projects

Biomedical NLP API | *FastAPI, Hugging Face Transformers, PyTorch, Docker* — 2026
- Designing and building a REST API exposing biomedical NLP inference (NER, relation extraction, embeddings) via authenticated endpoints, with OpenAPI documentation, containerized deployment, and prototype-grade throughput for demonstrating AI tools to non-technical stakeholders.

Experience

Network Operations Center Analyst — Oct 2023 – Present
Infrastructure Services, Virginia Commonwealth University — *Richmond, VA*
- Monitor and troubleshoot 1,400+ production servers, databases, and enterprise applications across Linux and Windows environments supporting university-wide academic, research, and clinical workloads.
- Authored the team's on-call operations guide (escalation ladder, triage workflow, shift handoff), translating technical procedures into clear, actionable documentation for analysts at all experience levels.

IT Team Lead — Jan 2022 – Aug 2023
Industry and Defense Programs, Wichita State University — *Wichita, KS*
- Led technical operations team supporting 2,000+ endpoints and 1,300+ users in a secure research environment, administering identity, endpoint, and configuration-management systems.
- Authored PowerShell and Bash automation reducing manual deployment and compliance-reporting effort; mentored 12+ interns, translating complex operational procedures into training materials and documentation for non-technical audiences.

Technical Skills

LLMs & RAG: Retrieval-Augmented Generation (RAG), LLM application development (Qwen3, Llama 3.1 via Ollama & vLLM), embeddings (Clinical ModernBERT), semantic retrieval, LangChain, LLM-as-Judge evaluation, PEFT / LoRA fine-tuning, prompt engineering
ML Frameworks: PyTorch, Hugging Face Transformers, LitGPT, Named Entity Recognition, Relation Extraction
Vector & Graph Databases: Neo4j (native vector indexes, Cypher), Qdrant, Chroma, SNOMED CT, UMLS, MeSH
Languages & Infrastructure: Python (pandas, NumPy), SQL, Bash/Shell, LaTeX; FastAPI, REST APIs, Docker, Git/GitHub, SLURM, HPC (multi-GPU H100), Linux (Rocky, Ubuntu)