Peiyu Lin

(206) 666-7021 | poppy.linpy@gmail.com | LinkedIn | Github | Willing to Relocate

Summary

AI Application–focused Software Engineer experienced in building LLM/RAG/agent systems end-to-end: data pipelines, evaluation, low-latency inference (vLLM/ONNX), and production APIs (FastAPI, Docker, AWS/GCP)

Education

- **Northeastern University** Sep 2023 – Apr 2026
  Master of Science in Computer Software Engineering GPA: **4.0/4.0**
  - Data Structures and Algorithms, Object-Oriented Design with C++, Operating Systems Data Management and Database Design, Web Design and User Experience Engineering

Experience

- **KLA** Jun 2025 – Dec 2025
  Software Engineer Intern | Python, LLMs, On-Device AI, Edge Computing Milpitas, CA
  - Shipped Fab Assistant, a customer-facing **LangGraph** multi-agent system deployed **on-premise** for real-time semiconductor troubleshooting, resulting in 25% reduction in troubleshooting time, enabling offline operation for sensitive manufacturing data
  - Collaborated with fab engineers to capture real troubleshooting workflows; implemented **model compression** and **edge-optimized embeddings** for deployment on resource-constrained fab floor terminals
- **Northeastern University** Aug 2024 – May 2025
  Graduate Teaching Assistant | Object-Oriented Design with C++ Seattle, WA
  - Led weekly 6-hour TA sessions mentoring 30 students; debugged C++ projects and toolchains, reviewed code and assessments, and authored 25 pages of documentation (examples, FAQs, setup guides) to standardize workflows

Hackathons

- **Salesforce TDX 2025 Agentforce Hackathon** | *AI Agents for Healthcare Institutions*
  - Prototyped a medical scheduling assistant on **Agentforce** with multi-turn flows for intent capture, patient verification, and slot booking
  - Implemented integrations using **Apex**, **Flow**, and **Salesforce Einstein** to exchange data with patient portals and hospital EMR systems in real time
- **Qualcomm & Microsoft 2025 On-Device AI Hackathon** | *On-Device Medical Interpreter*
  - Engineered a full-stack solution for processing medical record images with automated recognition, interpretation, and translation; emphasized **privacy by design** through local inference
  - Delivered a **FastAPI** backend and **React** frontend, integrated local **LLM APIs**, and **containerized** the system with Docker for portable demos

Projects

- **Bird Recognition Site** | *TensorFlow, CNN, EfficientNet*
  - Developed and deployed a **Streamlit** web app for bird species recognition; integrated **EfficientNetB4** transfer learning and an inference pipeline with preprocessing and postprocessing
  - Applied advanced data augmentation to improve generalization and connected the **Wikipedia API** to surface species facts and images for end users
- **Open Deep Research** | *LangGraph, LangChain, FastAPI, BGE, MiniLM*
  - Benchmarked two **LangGraph**-based pipelines, **LangChain** structured chains and a custom **GMIModel** API with JSON-schema I/O; authored a decision report to select the default architecture

Skills

Programming Languages	Python, C++, Java, JavaScript, SQL
AI / ML	TensorFlow, Scikit-learn, CNN, Transformers, OpenCV
Backend & Web	FastAPI, Node.js, React, Django, Spring Boot, REST APIs
Data & Infra	MySQL, MongoDB, Git, Docker, Agile, High Performance Cluster