

Huey Phan

Los Angeles, CA | phan.huey389@gmail.com | [ichomchom.github.io](https://github.com/ichomchom) | linkedin.com/in/huey-phan
github.com/ichomchom

Summary

AI/ML and Software Engineer with 8+ years of experience across software, data, and platform engineering, with hands-on experience building and deploying production ML systems. Currently pursuing an M.S. in Artificial Intelligence at UT Austin. Experienced with Python, PyTorch, scikit-learn, LightGBM, FastAPI, Go, React, Docker, AWS, LLM inference, RAG, and agentic AI systems. Strong production engineering background bridging models with APIs, databases, cloud infrastructure, and Linux systems.

Education

The University of Texas at Austin, MS in Artificial Intelligence – Austin, TX Jan 2025 – May 2027

California State University, San Bernardino, BS in Computer Science – San Bernardino, CA Sept 2014 – Dec 2017

Skills

Machine Learning: PyTorch, scikit-learn, LightGBM, TensorFlow/TFLite, pandas, NumPy

Generative AI: Transformers, Hugging Face, LoRA/PEFT, embeddings, RAG, quantization, llama.cpp, AnythingLLM

AI Systems: AWS Bedrock Agents/Guardrails, MCP tool integrations, SearxNG, Firecrawl, Tavily, Exa

Software: Python, Go, JavaScript/TypeScript, FastAPI, React, Streamlit, REST, GraphQL

Cloud/Data: AWS, Docker, CI/CD, Linux/RHEL, Ansible, Oracle, SQL

Experience

Programmer Analyst (ML / Software / Platform Engineering), Pasadena City College – Sept 2021 – present
Pasadena, CA

- Built and deployed a production machine-learning pipeline using Python, LightGBM, and scikit-learn to forecast course enrollment and cancellation risk from 10 years and approximately 104K Oracle records, reducing section-level prediction error by 17% and improving at-risk detection ROC-AUC from 0.80 to 0.95 versus a historical baseline.
- Engineered temporal feature and validation pipelines using time-based train/test splits to prevent data leakage and produce realistic out-of-sample forecasting performance.
- Designed and deployed ML model-serving architecture using FastAPI behind an existing Go backend, exposing enrollment predictions through REST APIs to a React application used by the scheduling department.
- Built and maintained Go-based application services supporting a student platform serving 40K+ concurrent users, with Docker-based deployment and CI/CD automation.
- Designed and developed a student class-schedule platform using React on AWS Amplify and a Go API on AWS EC2, integrating AWS networking, HTTPS/TLS through an Application Load Balancer, Apache reverse proxying, application caching, and production troubleshooting.
- Redesigned the class-schedule caching architecture from an incremental comparison model to a complete three-minute snapshot refresh, simplifying the system and eliminating timeout and stale-data issues during high-demand registration periods.
- Developed Python and Bash automation for data exchange and scheduled integrations with 5+ external vendors, including file transfer, validation, retries, monitoring, and error handling.
- Optimized application and Oracle query workflows, including GraphQL-backed services, reducing query latency by approximately 60%.

Software Developer, Cloudsquare – Woodland Hills, CA

Jan 2020 – Sept 2021

- Built a real-time edge-AI face-detection pipeline using TensorFlow Lite on Raspberry Pi hardware for client demonstrations and prototyping.
- Developed backend integrations and enterprise applications integrating Microsoft Graph, Plaid, Reuters, and Salesforce APIs.
- Built software across frontend, backend, integration, and deployment layers for enterprise client projects.

Database Administrator, ADR Services, Inc. – Los Angeles, CA

July 2018 – Jan 2020

- Optimized Oracle query execution plans, reducing unnecessary table scans and improving retrieval performance by approximately 60%.
- Developed Python ETL pipelines to automate data transformation, governance, validation, and integrity checks.

Jr. System Administrator, Cal State San Bernardino – San Bernardino, CA

June 2017 – Dec 2017

- Administered Linux/Unix and Windows infrastructure supporting campus computer labs serving 1,000+ students.
- Automated system administration and application deployment using Ansible, Bash, PowerShell, and Docker, reducing repetitive provisioning work from hours to minutes.
- Supported Samba, Active Directory, monitoring, backup, patching, and operational automation while developing documentation and runbooks for support staff.

Projects

Agentic AI & Local LLM Systems Lab

2025-2026

- Built a self-hosted AI environment using llama.cpp with Gemma-family models for private inference, RAG, and agentic workflows.
- Integrated AnythingLLM, Qwen3 embeddings, and agent tooling including Hermes, OpenCode, and Pi Agent with locally hosted model endpoints.
- Connected research/retrieval tools including SearxNG, Firecrawl, Tavily, Brave Search, and Exa through MCP-style tool workflows.

AWS Bedrock Student Pathways Prototype

2026

- Built a hackathon-style generative-AI prototype for helping community-college students explore programs and career pathways using AWS Bedrock Agents and Streamlit.
- Integrated S3-based knowledge sources, Lambda-backed actions, and Bedrock Guardrails for retrieval, deterministic application logic, and safety controls.
- Designed the prototype as an end-to-end AI application spanning retrieval, cloud permissions, agent orchestration, tool execution, and user interaction.

Research

Multi-Label ECG Classification — Deep Learning / Transformers

2026

- Developed and compared CNN, recurrent, hybrid, and Transformer architectures for multi-label ECG classification using PTB-XL.
- Achieved a best macro F1-score of 0.718, micro F1-score of 0.750, and macro ROC-AUC of 0.916.

Autoregressive Image Generation

2025

- Implemented a decoder-only Transformer for autoregressive image generation, representing images as sequences of discrete visual tokens.
- Built BSQ/VQ-based quantization pipelines and trained the model end-to-end for next-token prediction and autoregressive sampling.

Debiasing Natural Language Inference Models

2026

- Investigated shortcut learning in NLI models by studying whether hypothesis-only confidence could reliably identify biased examples.
- Demonstrated limitations of confidence-based debiasing and analyzed model failure cases; published findings as an EngrXiv preprint.

Los Angeles County Housing Prediction

2026

- Built an end-to-end ML workflow for ZIP-level housing-value prediction using public Los Angeles County datasets.
- Performed data preparation, feature engineering, model development, evaluation, and analysis for a real-world regression problem.

Publications

Comparing CNN, Recurrent, Hybrid, and Transformer Models for Multi-Label ECG Classification on PTB-XL

Aug 2026

Phan, H.

[10.31224/7898](https://engrxiv.org/engrxiv/10.31224/7898) (EngrXiv)

The Challenge of Debiasing NLI Models: Why Hypothesis-Only Confidence is Insufficient

Jan 2026

Phan, H.

[10.31224/6210](https://engrxiv.org/engrxiv/10.31224/6210) (EngrXiv)

Predicting ZIP Level Housing Values in Los Angeles County: A Machine Learning Approach Using Public Data

May 2026

Phan, H.

[10.20944/preprints202605.2102.v1](https://arxiv.org/abs/10.20944/preprints202605.2102.v1) (Preprints)

Presentations

UT REAL Health AI Symposium

2026

- Accepted for poster presentation: *Comparing CNN, Recurrent, Hybrid, and Transformer Models for Multi-Label ECG Classification on PTB-XL.*

California Heart Rhythm Symposium

2026

- Accepted for poster presentation: *Comparing CNN, Recurrent, Hybrid, and Transformer Models for Multi-Label ECG Classification on PTB-XL.*