Wesley Hang

wesley.important@gmail.com | 240-286-4393 | wesleyhang.dev | linkedin.com/in/wesley-hang-901a04214/

EXPERIENCE

NVIDIA x UMD | *Undergraduate Researcher - Cloud LLM* **College Park, MD | Aug 2025 - Dec 2025**

Engineered scalable cloud environments to benchmark and optimize LLM inference frameworks using containerization.

- Benchmarked vLLM + SGLang on cloud GPUs, improving throughput and enabling hardware comparisons across 5 models.
- Reduced experiment setup time by 40% using Helm + Kubernetes deployments with automated infrastructure provisioning.
- Automated large-scale Python benchmarking, reducing runtime 50% while tracking latency, throughput, and GPU usage.
- Developed a Prometheus + Grafana monitoring dashboard tracking 12+ metrics to identify performance patterns.

District Department of Transportation | *Software Engineer Intern* **Washington D.C. | Jun 2025 - Dec 2025**

Developed an advanced asset management platform and created multiple automation / monitoring tools to ensure asset availability.

- Developed a FastAPI + React platform to replace legacy Windows Form inventory system supporting 3000+ transportation assets, implementing full CRUD operations to reduce manual database workflows by 50%
- Built a secure 13+ endpoint REST API with JWT-based access control with 4 customizable roles serving 50+ internal users, reducing asset lookup time by 40%.
- Integrated OpenStreetMap with Leaflet to visualize 2,500+ real-time assets, utilizing clustering to maintain 60fps UI.
- Deployed internal ITS inventory system via IIS (Internet Information Services) to allow secure company-wide usage.

PROJECTS

AI Agent Proxy | *Python, FastAPI, SQLite, Javascript* *https://github.com/Wuzu11517/agentic-proxy*

Middleware proxy intercepting Anthropic API traffic to reduce cost and improve observability with a single endpoint change.

- Eliminated redundant API calls using SHA-256 indexed LRU/TTL caching for duplicate requests on repeated workloads.
- Supports real-time streaming and standard requests, compatible with production chatbots and agent frameworks
- Auto-routes requests to cheaper models based on prompt complexity, tracking per-call cost delta between models.
- Aggregates cost savings across sessions in SQLite, attributing reductions to caching vs model routing on a live dashboard

TxnFlow | *Go, PostgreSQL, Docker, JavaScript* *https://github.com/Wuzu11517/TxnFlow/*

Asynchronous ETH transaction pipeline ingesting data via Infura RPC that optimizes queries through strategic indexing.

- Optimized blockchain queries from 200ms to under 5ms through PostgreSQL indexing, achieving 1,200+ req/sec throughput.
- Implemented CORS middleware and state machine tracking transaction lifecycle and audit logging for all transactions
- Built async worker with exponential backoff retry logic and 10 concurrent goroutines, maintaining 6-10 tx/min throughput while eliminating RPC blocking

Loomi | *Python, FastAPI, Docker, Tailwind, Supabase* *https://loomi.life*

Anonymous perspective platform that connects short life perspectives through AI-powered semantic matching

- Built a vector similarity system using OpenAI embeddings + pgvector, tuning cosine similarity thresholds to balance relevance and viewpoint diversity
- Designed a personalized feed by modeling user interest as a time-decayed vector derived from interaction history
- Developed a 2-layer moderation pipeline combining adversarial text normalization with OpenAI's moderation API

SKILLS

Programming Languages: Python, C++, C, MATLAB, SQL, TypeScript, R, Go
Web & Backend Development: React, FastAPI, HTML, REST APIs, Playwright, Vercel, Render, Data Structures & Algorithms
Cloud, DevOps & Tooling: Docker, Kubernetes, Helm, Azure, Git, npm, Locust, Supabase
Data, ML & Analytics: Pandas, NumPy, Scikit-learn, GIS, SAS

EDUCATION

University of Maryland, College Park: *B.S. Computer Science / Minor in Statistics* **Dec 2025**

GPA: 3.71

Relevant Coursework: Algorithms: Deep Learning, Advanced Data Structures, Network Security, Data Science, Cloud Computing, Compilers