

Haoran (Casper) Zhang

(206) 619-4535 | hzhang498@outlook.com | casperzhang.dev | github.com/Wuj1Zhang | [LinkedIn](#)

EDUCATION

Yale University | M.S. in Computational Biology and Biomedical Informatics
University of Washington | B.S. in Computer Science, GPA 3.9

New Haven, CT | Aug. 2024 – May 2026
Seattle, WA | Sep. 2020 – Jun. 2024

EXPERIENCE

NLP Research Intern, NIH-Funded PFAS Water Project [[Demo](#)] [[Code](#)]

Jun. 2025 – Feb. 2026

Yale School of Medicine (Advisor: Kei-Hoi Cheung)

New Haven, CT

- Co-built an LLM pipeline screening **18.7M** PubMed records down to 1,044 PFAS drinking-water studies (**0.95 F1**), extracting 10,996 entities across 8 PECO types at 0.87 macro-F1.
- Built the geographic evidence map end to end: Google Maps geocoding onto an interactive OpenStreetMap frontend with chemical/year filters, citation-weighted markers, and entity-highlighted popups.
- Ran the human-in-the-loop evaluation: iterative prompt refinement, held-out benchmarks (200 screening / 100 extraction abstracts), and error analysis separating boundary mismatches from true failures.

LVLM Research Assistant, LMOD+ Benchmark [[Paper](#)] [[Web](#)]

Dec. 2024 – Jun. 2025

Yale University (Advisor: Qingyu Chen)

New Haven, CT

- Built the unified PyTorch inference framework benchmarking **24 MLLMs** (GPT-4o, InternVL, Qwen, DeepSeek) under identical prompts and configs across 4 task families on 32,633 images.
- Designed the curation pipeline unifying 10 public datasets with incompatible formats into one benchmark (5 modalities, 12 conditions, ~50% expansion); released the dataset, pipeline, and leaderboard.

ML Research Assistant, Edge Vision MoE

Jan. 2023 – Dec. 2023

University of Washington (Advisor: Kurtis Heimerl)

Seattle, WA

- Designed a parallel adapter mixture-of-experts for compact ViTs, replacing softmax routing with input-dependent sigmoid gates for a static execution graph and deterministic latency on ANE / Edge TPU.
- Raised iNat'21 fine-grained accuracy **+8.9 points** (61.5 to 70.4) at ~2% FLOPs overhead using 1x1-conv bottleneck adapter experts; +4.8 on RareSpecies at the same cost.
- Stabilized training with an entropy-regularized gating loss preventing expert collapse, plus two-stage tuning: dense search on a 10% subset, then joint gate-expert optimization via Optuna.

Data Scientist Intern

Jun. 2023 – Sep. 2023

JICHUANG Technology Co., Ltd. (online education platform)

Taiyuan, China

- Shipped a hybrid course recommender combining collaborative filtering with BERT content embeddings, raising top-5 precision about 20% and closing the cold-start gap for new students.
- Trained Random Forest and XGBoost models on student activity data to flag dropout risk at 90% accuracy, 15 points above the existing baseline.

PROJECTS

FleetLab | Accelerator Scheduling Simulator and Control Plane [[GitHub](#)]

Jun. 2026 – Jul. 2026

- Wrote a deterministic C++20 discrete-event scheduler: 100,000 tasks over a 10,000-node TPU-style 3D torus in about 80 s single-process, with four placement policies, verified by **381 tests** and a 10,000-seed invariant sweep.
- Extended it into a control plane of 20 OS processes (gRPC, Docker) with heartbeat failure detection, epoch-fenced reassignment, and dedup under injected duplicates; a scripted split-brain run asserts the eight-event recovery chain.
- Reported where the policy breaks (gang placement with preemption re-runs its candidate search per victim, stalling under contention) instead of only favorable configs; every figure traces to a scenario hash, seed, and artifact.

VibeCourt | Local-First AI Mediation Platform [[GitHub](#)] [[Demo](#)]

Mar. 2026 – Jun. 2026

- Built a crash-safe four-stage LLM pipeline (interview → extraction → statement → ruling) that never trusts model output as state: answers persist append-only before any AI call, so a crash or refresh resumes at the exact same step.
- Enforced AI correctness in code, not prompts: **35 machine-checked rules** across 5 per-stage validators reject bad output — failed calls stay pending for explicit retry, never a silent local fallback; **784 tests** run fully offline.

SKILLS

Languages: Python, C++20, TypeScript, Java, SQL, Bash

ML / LLM: PyTorch, Hugging Face, vLLM, CUDA, FlashAttention, LangChain, RAG, LoRA/PEFT, scikit-learn, pandas

Systems: gRPC, Protocol Buffers, Docker, CMake, Linux, GoogleTest, pytest, AWS, Git

Data & Web: PostgreSQL, Redis, ChromaDB, Neo4j, React, Next.js, FastAPI

PUBLICATIONS

- LMOD+: A Comprehensive Multimodal Dataset and Benchmark for Developing and Evaluating Multimodal Large Language Models in Ophthalmology. *ACM Transactions on Computing for Healthcare*, 2026. [DOI](#)
- A Large Language Model Approach for Automatic Extraction of Metadata Related to PFAS in Drinking Water. Under review, *Nature* collection on AI for Sustainable Clean Water, 2026.