

Haoran (Casper) Zhang

(206) 619-4535 | hzzhang498@outlook.com | casperzhang.dev | [GitHub](#) | [LinkedIn](#)

EDUCATION

University of Washington, Seattle, WA

B.S. in Computer Science | GPA: 3.9

Sep. 2020 – Jun. 2024

Yale University, New Haven, CT

M.S. in Computational Biology and Biomedical Informatics

Aug. 2024 – May 2026

EXPERIENCE

NLP Research Intern: NIH-Funded Water Contamination Project

Jun. 2025 – Feb. 2026

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

New Haven, CT

- LLM-mined PubMed at scale: screened **18.7M** PubMed records, filtered to **3,611** PFAS water studies (GPT-4.1-mini), and extracted **39,061** entities across **7** types (GPT-4.1, JSON schema).
- Prompt engineering + QA: developed few-shot, schema-constrained extraction prompts and an error-driven tuning loop; delivered **0.94 F1 (98% P / 90% R)**; wrote guidelines and managed 6 annotators (IAA=**0.75**).
- Productized exploration: MedViz semantic maps (bge-small-en-v1.5 + t-SNE) + Neo4j/Bloom knowledge-graph visualizations for interactive filtering and discovery.

Multimodal LLM Research Assistant — LMOD+ Benchmark

Dec. 2024 – Jun. 2025

Yale University (Advisor: Qingyu Chen)

New Haven, CT

- Contributed novel eye-related multimodal datasets and developed scalable ETL pipelines for dataset preprocessing and augmentation, expanding annotated data volume by **32%** and improving rare condition coverage by **40%**
- Developed a comprehensive benchmark framework for **5** SOTA LVLMs on ophthalmic imaging, automating evaluation across **16+** diverse clinical tasks and accelerating benchmarking speed by **3x** via FlashAttention integration
- Integrated vLLM and Tensor Parallelism into benchmarking framework for large-scale LVLMs (100B+) on 4×A100 GPUs, delivering **8.2×** speedup and **65%** lower VRAM usage

ML Research Assistant: Low-Resource Environmental Monitoring

Jan. 2023 – Dec. 2023

University of Washington (Advisor: Kurtis Heimerl)

Seattle, WA

- Developed hybrid CNN-Transformer MoE by integrating sparse expert layers with MobileViTV2, achieving **4x** parameter reduction (**4.4M** → **1.08M**) while maintaining competitive accuracy on birds classification tasks
- Developed novel clustering-based approach using K-means on patch embeddings, solving training instability through cosine similarity-based expert assignment without auxiliary losses
- Optimized transformer attention with patch-level routing decisions and multi-scale training with top-k strategy, cutting FLOPs by **8%** with **<1%** accuracy drop on fine-grained bird classification

Data Scientist Intern

Jun. 2023 – Sep. 2023

JICHUANG Technology Co., Ltd.

Taiyuan, Shanxi

- Engineered a hybrid course recommender system using collaborative filtering combined with BERT-based content embeddings, boosting **top-5** recommendation precision by **~20%** and alleviating cold-start issues for new students
- Developed and optimized machine learning models (Random Forest, XGBoost) to predict student performance and dropout risk, achieving **90%** accuracy (**15%** improvement over baseline) and enabling early intervention strategies
- Fine-tuned Transformer NLP models on educational text (student feedback & course content) to perform sentiment analysis and topic extraction, attaining **82%** classification accuracy and uncovering actionable insights for curriculum improvement

PROJECTS

FitNutri AI - Personal Health Analytics with RAG

Aug. 2024 – Dec. 2024

- Built a RAG-based health analytics web app with LangChain, Llama 3.1, and ChromaDB over a corpus of **5K+ research abstracts**; improved answer relevance by **73%** on a manually curated evaluation set.
- Developed a full-stack Django/PostgreSQL backend with Redis caching (**82%** hit rate), cutting API latency by **60%** and ensuring **94%** data accuracy on USDA nutrition queries
- Built a React chat interface with a **2.8s mean response time** and received **85% positive feedback** in a class user study (**n=32**), with the application containerized using Docker.

SKILLS & CERTIFICATIONS

Programming Languages: Python, Java, SQL, C++, Bash/Shell scripting

ML/LLM Frameworks & Models: PyTorch, TensorFlow, Hugging Face, OpenAI, Llama, Qwen VL, vLLM, LangChain

Cloud & Development Tools: AWS, Azure, Docker, Git, TensorBoard, Weights & Biases, ChromaDB, Redis, PostgreSQL

Libraries & Frameworks: NumPy, pandas, scikit-learn, React, Next.js, Django, FastAPI, NLTK, matplotlib

Certifications: AWS Certified Cloud Practitioner

PUBLICATIONS

- LMOD+: A Comprehensive Multimodal Dataset and Benchmark for Developing and Evaluating Multimodal Large Language Models in Ophthalmology. Published in *ACM Transactions on Computing for Healthcare*, 2026. DOI
- A Large Language Model Approach for Automatic Extraction of Metadata Related to PFAS in Drinking Water. Submitted to a Nature Collection on *AI for Sustainable Clean Water*, 2026.