

JOHNNY HUANG

San Mateo, CA, 94403 | (650)-278-6570 | h.johnny@wustl.edu | [Portfolio](#) | [Linkedin](#) | [GitHub](#)

EDUCATION

St. Louis, MO

Washington University in St. Louis

August 2022- May 2026

- B. S: Computer Science + Mathematics

GPA: 3.9

- Honors and Activities: Chancellor's Fellow, Taylor Scholar, **3x** WashU Hackathon (2023 Co-organizer), Computational Imaging Group, **Head TA** for "CSE 412: Intro to AP", **VP** of WashU Robotics, **VP** of First-Generation Investors Club
- Relevant Coursework: Machine learning, Bayesian ML, RL, AI, LLM, DSA, Data Mining, Computer Engineering, Convex Optimization

PROFESSIONAL EXPERIENCES & INVOLVEMENT

Large Language Models Intern

May - July 2024

Rad AI

San Francisco, CA

- Led the development of a full-stack RAG chain system using **Langchain** and **Pinecone** vector database for optimized search versatility; established access endpoints with **FastAPI** and deployed through **Docker** on **AWS Sagemaker**.
- Fine-tuned Gemma2-7b for function-calling leveraging HF's **PEFT** and **LoRA** finetuning which increased training speed by **600%**; optimizing model performance through sharding; utilized the **Pydantic** framework for data validation.

Machine Learning Intern

Jan - May 2024

Mallinckrodt Institute of Radiology

St. Louis, MO

- Developed a modified U-Net using **PyTorch**, **CUDA** and **Caffe** for fMRI segmentation; applied YOLO for anomaly detection, integrating a PnP-FISTA pipeline for improving runtime speed; algorithm successfully predicts **93%** of critical regions.
- Constructed a deep cGAN to streamline more robust synthetic CGM vs Cognitive function data generation using **TensorFlow**.

Software Engineering Lead

September 2023 - Feb 2024

WashU Robotics, MATE ROV

St. Louis, MO

- Spearheaded robot sensory systems processing efficiency speed through integrating the AutoViz and Gazebo **API** in **C++**; optimizing scripts compilation time in poolside controller by **150%** using **ROS** and **C** on **Linux**.

Data Science Intern

May - July 2023

Couch Biomedical Science

St. Louis, MO

- Denoised and refined collected data by implementing a single-celled Deep-Count Autoencoder using **PyTorch**; updating databases with **PySpark** and **MySQL** on **Kubernetes** clusters for parallelization, increasing processing speeds by **500%**.
- Clustered and classified DNA sequence into gene segments by developing SVM and Agglomerative models using **Sci-Kit**, **NumPy**, and **Pandas**; processed models on **Azure**; visualized results and presented to wet lab.

PERSONAL PROJECTS (See More on my [Portfolio!](#))

- **HalluAgent**: Developed a framework for utilizing LoRA tuned SLMs to detect and correct hallucination patterns in GPT-3.5; leveraged SLMs as agents to evaluate LLM response w/ confidence scores by calling various custom-deployed functional **API**'s; retrained GPT-3.5 using **HF's Trainer** library; generated **2000** robust trajectory training data with GPT-4 for agent tuning.
- **Diary App**: Created a full stack website for users to take notes/write diaries utilizing a **ASP.NET** MVC framework in **C#**; stores user information dynamically utilizing a document-based **NoSQL** database.
- **Petrichor**: A mental health website aiming to match users with their perfect therapist; implementing user login, menus, calendars; data stored with **MySQL** on **Linux**; hosted on **AWS**; coded mainly using **HTML/CSS**, **PHP**, and **JS**.

SKILLS

- **ML**: PyTorch, TensorFlow, HuggingFace, Pydantic, Langchain, PySpark, Azure, SageMaker, Pinecone, CUDA
- **Backend**: Java, C++, Scala, Python, C, NodeJS, ExpressJS, FastAPI, MySQL, NoSQL, .NET
- **Frontend**: PHP, React, JavaScript, HTML, CSS, Swift, Apache
- **Others**: Jupyter, AWS, MongoDB, Docker, Kubernetes, Git, Bash, PowerShell, R, Matlab