BENJAMIN WONG

me@benjiwong.com • benjiwong.com • linkedin.com/in/benjibenji/ • github.com/chiyeon

EDUCATION

University of California Irvine Sep 2021 – Jun 2025

- **Computer Science (Intelligent Systems), B.S. – 3.9 GPA** *Irvine, CA*
- **Relevant Coursework:** Machine Learning & Artificial Intelligence, Compilers & Interpreters, Computer Vision, Operating Systems & Computer Architecture, Search Engines, Quantum Computing, Software Design & Development, Data Structures & Algorithms, Linear Algebra, Statistics

EXPERIENCE

Persimmons.ai Jun 2024 – Dec 2024
Software Engineer Intern *San Jose, CA*

- Engineered and optimized a **scalable distributed inference system** handling 40+ concurrent clients utilizing **70B+ parameter LLMs** via vLLM & llama.cpp across multiple CUDA devices
- Implemented API-driven tool usage within LLMs, extending decision-making capabilities and facilitating flexible integration into AI assistant workflows
- Developed robust **pytest suite** and detailed documentation to ensure system stability and expansion

Recogni Jun 2022 – Sep 2022
Software Engineer Intern *San Jose, CA*

- Built a Python-based **synthetic image dataset generator**, producing multiple 3000+ photorealistic image datasets rendered in Blender to enhance 3**D object detection** for autonomous vehicles
- Improved experimental **PyTorch** model accuracy by 20% through training and iterative tuning
- Composed detailed, extensible documentation for streamlined onboarding and future development

Mechanical Keyboard Club at UC Irvine Apr 2022 – Present
Webmaster *Irvine, CA*

- Partnered with designer to craft a responsive **Vue**-based website, emphasizing interactivity and efficiency
- Enhanced and optimized user experience through intuitive interface design and efficient data loading
- Diagnosed and resolved production issues, deployed fixes, and consistently updated content to ensure reliability and relevance

PROJECTS

Crux Language Compiler – *Java, Maven, Antlr4* Jan 2025 - Present

- Built a compiler for the Crux language, employing graph- and tree-based lowering to optimize syntax validation, semantic analysis, and efficient x86 machine code generation

TMF: Music Sharing Platform – *Vue, Node.js, Firebase, Google Cloud Platform, ffmpeg* May 2024 - Present

- Developed a full stack application enabling users to upload, stream, organize, and share music
- Deployed secure, load-balanced API on **Google Cloud Platform**, ensuring high availability and optimized response times for a seamless user experience
- Implemented JWT-based user authentication, with **granular permission control** and action validation
- Tuned database caching strategies to **cut load times** by 75%, significantly improving responsiveness

Woodstock Chess Engine – *C++, Emscripten, HTML/Javascript* Jun 2023 – Nov 2023

- Engineered a powerful C++ based chess engine using iterative deepening and alpha-beta pruning capable of **deep positional analysis** and leveraging **object oriented design** for a modular yet efficient approach
- Constructed a comprehensive testing and evaluation framework to identify performance bottlenecks and ensure accurate move evaluation

SKILLS

Languages: Python, C/C++, Java, Javascript, Typescript, C#, SQL, HTML/CSS, Bash

Frameworks: Node.js, Vue, React, PyTorch, llama.cpp, Express, FastAPI, Unity, Godot, scikit-learn, pandas

Tools & Platforms: Git, Docker, Google Cloud, Firebase, AWS, MongoDB, Cmake

Additional: GNU/Linux (Debian, Arch, Ubuntu), UI/UX, Agile, CI/CD, Game Development (Unity, Godot)

Interests: Photography, Music Production, Video Production, Powerlifting, Mechanical Keyboards