

# BENJAMIN WONG

[me@benjiwong.com](mailto:me@benjiwong.com) • [benjiwong.com](http://benjiwong.com) • Irvine, CA • [linkedin.com/in/benjibenji/](https://linkedin.com/in/benjibenji/)

---

## EDUCATION

University of California Irvine – Computer Science, B.S.

September 2021 - June 2025

- **GPA:** 3.9
  - **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
- 

## WORK EXPERIENCE

Persimmons.ai – Software Engineer Intern, San Jose, CA

June 2024 – December 2024

*Python, vLLM, HuggingFace transformers, llama.cpp, FastAPI, Pydantic, pytest, Docker*

- Developed and optimized a scalable API server handling 40+ concurrent clients performing inference on LLMs (up to 70b parameters) across multiple CUDA devices
- Implemented tool usage to extend LLM decision-making capabilities, enabling seamless API tool calls, improving integration flexibility for AI-powered applications like personal assistants
- Maintained comprehensive documentation and testing suite to ensure stable development and usage

Recogni – Software Engineer Intern, San Jose, CA

June 2022 – September 2022

*Python, Blender, PyTorch*

- Designed & implemented a synthetic image dataset generator in Python, generating multiple 3000+ image datasets for refining 3D object detection models used in self-driving automobiles
  - Refined algorithms based on model results and supervisor feedback to improve detection accuracy
  - Authored detailed documentation, covering the project's usage and extensibility
- 

## PROJECTS

Crux Language Compiler – Basic language compiler for x86

January 2025 - Present

*Java, Maven, Antlr4*

- Developed a compiler for the Crux language, generating optimized x86 machine code
- Implemented graph and tree-based data structures for efficient syntax checking and compilation

TMF Music Website – Full stack music sharing platform

May 2024 - Present

*Vue, Node.js, Firebase, Cloud, ffmpeg*

- Developed authoritative RESTful API deployed to Google Cloud with load balancing and SSL certificates to ensure security and responsiveness
- Secure user authentication (server authoritative, salting & hashing) with user permissions and customizable user profiles
- Optimized caching mechanisms for fast and efficient database retrieval, reducing load times by 75%

Woodstock – Efficient Chess engine and AI opponent

June 2023 – November 2023

*C++, Emscripten, HTML/Javascript*

- Implemented algorithms iterative deepening, alpha-beta pruning for a powerful, efficient engine in C++
  - Constructed AI opponent that chooses the “best” moves in variable scenarios, looking 7+ steps ahead
  - Developed full testing suite to uncover performance faults, engine inaccuracies, and bottlenecks to ensure a fully accurate, fast move generation and evaluation
- 

## CAMPUS INVOLVEMENT

Webmaster – Mechanical Keyboard Club, Irvine, CA

April 2022 – Present

- Collaborated with designer to build website with Vue, prioritizing interactivity, efficiency, and aesthetics
  - Optimized for a responsive, user-friendly interface through intuitive design and efficient data loading
  - Troubleshooted and maintained stable public release, fixing reported issues and updating events to ensure user satisfaction and information accuracy
- 

## SKILLS

**Programming:** Bash, C, C++, C#, Java, Javascript, Typescript, Node, Python, R, Rust, SQL, HTML & CSS

**Technologies:** Google Cloud, Firebase, Vue, React, vLLM, PyTorch, Transformers, llama.cpp, Matlab

**Misc:** GNU/Linux (Debian, Arch, Ubuntu), Docker, Git, Cmake, Unity, Godot, Blender, UI/UX, Agile