

# Aryan Keluskar

[linkedin.com/in/aryankeluskar](https://www.linkedin.com/in/aryankeluskar) — [github.com/aryankeluskar](https://github.com/aryankeluskar) — (602) 552-6402 — [aryan@aryankeluskar.com](mailto:aryan@aryankeluskar.com)

## EXPERIENCE

---

**Software Engineer (ML Research)**, The Biodesign Institute May 2024 – Present

- Optimized GPU inference, implemented distributed computing with **Slurm** with gradient checkpointing and memory-efficient batching using **PyTorch**, cutting compute time by 30% and throughput time by 25%.
- Cut GPU costs by 75% (approximating **\$3,850**) by implementing efficient Machine Learning techniques, such as Model Distillation & Flash Attention, for Protein Language Models using **PyTorch** and **Transformers**.
- Achieved a speedup of **250x** and reduced parameter count by **75%** while maintaining **99%** accuracy to state-of-the-art protein language models. Co-authored a **NeurIPS** Workshop Paper.

**Machine Learning Researcher**, Data Mining & Machine Learning Lab August 2024 – Present

- Implemented a model inference API using **vLLM** for GPU memory management and **FastAPI** for RESTful endpoints supporting batch processing and concurrent requests, which lowered inference latency by 30%.
- Built classifiers for text datasets over 2TB in size and fine-tuned LLMs using **LoRA** and **Python** on **AWS Sagemaker** to improve accuracy in ambiguous human text.
- First-author on IEEE BigData research paper and co-author on ACL 2026 Workshop paper.

**Software Engineer Intern**, HealthGC (VC Backed Startup) May 2025 – December 2025

- Engineered a sub-200ms (P90) retrieval augmented generation (RAG) system with HIPAA compliance using **PostgreSQL**, **Google Cloud**, **WebAssembly** and **JavaScript** for retrieving multimodal data.
- Built a fault-tolerant webhook infrastructure using **Twilio**, **Deepgram**, **Express.js**, and **Docker**, achieving real-time cellular communication with 99.9% uptime and cutting message delivery latency by 35%.
- Developed tool calling functionality for web search and RAG for an LLM-based chatbot using **Google Vertex AI**, enabling contextual and up-to-date responses for 1000+ messages daily.

**Software Engineer Intern**, RCV Innovations June 2023 – July 2023

- Built a cloud-based backend server using **JavaScript**, **Node.js** and **Express.js** to efficiently deliver digital twins & 3D assets, cutting website loading times by 46% and increasing user session durations by 17%.

## PROJECTS

---

**Jiggle Wiggle - 2x Track Winner at TreeHacks 2026** — Python, Computer Vision [dub.sh/treehacks](https://dub.sh/treehacks)

- Jiggle Wiggle is a real-time AI dance & fitness coach accepting webcam, YouTube, and Zoom input. We placed first out of 30 teams for prize of \$2000 and runner-up out of 75 teams winning a prize of \$5000.
- Engineered an end-to-end pose detection pipeline with **MediaPipe** running entirely in-browser via **WASM** for client-side tracking, paired with **SAM2** segmentation on Modal serverless GPUs.

**Canvas MCP Server** — TypeScript, Bun, OAuth 2.0, Model Context Protocol [dub.sh/canvas-mcp](https://dub.sh/canvas-mcp)

- Built an MCP server to mitigate hallucinations by letting AI agents access educational data on Canvas LMS using **TypeScript**, **Bun** and **AWS Lambda**. Scaled to 1200+ monthly active users with 100% uptime.

## EDUCATION

---

**B.S. Computer Science (Honors)**, Arizona State University

## HACKATHONS & AWARDS

---

**TreeHacks 2026** at Stanford University: Zoom Challenge Winner and AI Inference Track Runner-up

**HackMIT 2024**: Modal's Sponsor Award Winner; **SFHacks 2024**: 'Best Use of AI' Track Winner

## SKILLS & TECHNOLOGIES

---

**Languages:** TypeScript, JavaScript, Python, Java, C++, SQL

**Technologies:** Linux, Git, Cursor, Claude Code, Raspberry Pi, Arduino, Docker, DigitalOcean VPS, Stripe API

**Full-Stack Development:** React.js, Node.js, Next.js, TypeScript, HTML, CSS, JavaScript, PostgreSQL, Amazon Web Services (AWS), Google Cloud (GCP), Google OAuth, SpringBoot, Python, Flask, Tailwind