Aryan Keluskar

linkedin.com/in/aryankeluskar — github.com/aryankeluskar — (602) 552-6402 — aryan @ aryankeluskar.com

Experience

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

- Cut inference costs by 93% ($36k annually) using efficient machine learning techniques, such as distillation & Flash Attention, for protein language models using **PyTorch** and **Transformers** libraries in Python.
- Optimized GPU usage by implementing distributed training with gradient checkpointing and memory-efficient batching using **JAX** and **Slurm**, cutting training walltime from 25 to 7 hours.
- Developed reproducible benchmarking pipelines in Python using **HuggingFace**, **Hydra**, and **Slurm** to evaluate 6 models across 12,000+ protein domains to quantify inference speedups.
- Presented a talk to an audience of 200+ biology and machine learning researchers hosted by Chan Zuckerberg Biohub.

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 voice agent TTFT from 2.2s to 1.2s.
- Built a voice agent testing and A/B prompting infrastructure using **Gemini, Deepgram,** and **Twilio** for automated testing of our core product directly integrated into **GitHub Actions** and **Posthog**.
- Developed tool calling functionality for web search and RAG for an LLM-based chatbot using **Google Vertex AI**, enabling contextual and scaled it to 345k calls and 8.4 million messages for 40k users.

Student Researcher, DMML Lab @ Arizona State University — August 2024 – December 2025

- Implemented a model inference API using **vLLM** for GPU memory management and **FastAPI** for RESTful endpoints supporting batch processing and concurrent requests, which increased throughput by 12x.
- Published three first-author research papers at IEEE BigData, ICML and COLM conferences, which are available on Google Scholar: https://scholar.google.com/citations?user=wVDeZtcAAAAJ

Projects

AI Dance & Fitness Coach - 2x Winner at TreeHacks 2026 | February 2026 — dub.sh/treehacks

- Built a real-time AI dance & fitness coach by engineering 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.

Polymarket & Canvas MCP Servers | July – November 2025 — dub.sh/canvas-mcp and polymarket-mcp

- Built and deployed two MCP servers to fetch financial information from Polymarket and educational materials from Canvas with scoped access control and OAuth integration. Used **TypeScript**, **Bun** and **AWS Lambda**. Scaled to **51.3k users** with 99.91% uptime.

Education

B.S. Computer Science (Honors), Arizona State University — GPA: 3.99/4.0

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

TypeScript, JavaScript, Python, Java, C++, PostgreSQL, Linux, Git, Cursor, Claude Code, Arduino, Docker, Stripe, React.js, Node.js, Next.js, Amazon Web Services (AWS), Google Cloud (GCP), Cloudflare, SpringBoot, Flask, Tailwind, Google OAuth, PyTorch, Transformers, JAX, vLLM, Slurm