

Aryan Harkawat

US Citizen | Austin, TX | (512) 888-2423 | harkawat.aryan@gmail.com | [Github](#) | [LinkedIn](#)

EDUCATION

University of Texas at Austin

Austin, TX

Bachelor of Science in Computer Science

May 2028

- **Coursework:** Operating Systems, Computer Architecture, High Performance Computing, Algorithms & Complexity, Data Structures, Linear Algebra, Statistics and Probability, Principles of Machine Learning

WORK EXPERIENCE

Stripe

Jan 2027 – May 2027

Incoming Software Engineer Intern

Stripe

May 2026 – Aug 2026

Software Engineer Intern, Core Infrastructure

- Rebuilt load balancing in Stripe's service mesh around power-of-two-choices weighted by CPU utilization and queue depth, cutting overload errors 40% and saving ~\$200K/yr across 50 services serving 400M+ requests/day
- Added zone-aware routing that spills traffic to neighboring zones when local utilization passes 80%, reducing p99 latency 23% on average (300ms → 230ms)
- Shipped to production with zero incidents through feature-flagged canary ramps, gated on error and latency rates

Trend Micro

Aug 2025 – May 2026

Software Automation Intern

- Built a scheduled Python ETL pipeline for payroll reconciliation, featuring schema validation, fuzzy record matching across Excel and shared-drive sources, and automated exception reporting for audits
- Designed a network log pipeline using Power Automate and Python regex parsing to sessionize raw logs into SQL, powering Power BI dashboards with incremental refresh and anomaly flagging

Mercor

Feb 2025 – Apr 2025

Machine Learning Intern

- Built a reusable evaluation script with Hugging Face Transformers to benchmark 6 fine-tuned BERT variants on precision, recall, and F1, replacing manual runs and cutting comparison time to under an hour
- Exported the best BERT variant to ONNX with dynamic int8 quantization, cutting CPU inference latency 3× and batch-scoring 1M+ tickets in under 2 hours
- Built Pandas checks for duplicate, missing and conflicting labels on 1M+ tickets, raising baseline F1 by 2 points

PERSONAL PROJECTS

Concertify | *Go, gRPC, Python, PyTorch, Node.js, Redis, Prometheus, Docker* | [Demo](#) · [Source](#)

- Built a concert recommender (500+ users): two-stage retrieval and ranking over Spotify listening history and Ticketmaster events with cosine genre similarity, MMR diversity re-ranking and explainable scores
- Served ranking from a Go gRPC microservice (~1 ms/request) with per-user weights trained online in PyTorch
- Cut cold-start latency 9s to 2.5s under a 5 req/s quota with Redis single-flight caching and priority rate limiting

Herd | *Go, Envoy, gRPC, Kubernetes, Prometheus, Docker* | [Source](#)

- Built an OpenAI-compatible LLM inference gateway routing across self-hosted model servers by live KV-cache utilization and prefix-cache affinity, cutting p99 time-to-first-token 55% vs. round-robin
- Added criticality-based load shedding, weighted fair queuing and budgeted cloud spillover; with two degraded replicas held p99 TTFT at 12.6s where round-robin reached 292s
- Shipped as an Envoy ext_proc endpoint picker with Kubernetes manifests, a load generator and live dashboard

Volumetric Cloud Renderer | *C++, OpenGL, GLSL, CMake* | [Source](#)

- Built a real-time volumetric cloud renderer: ray-marched GPU-generated 3D Perlin-Worley noise through a spherical cloud layer with multi-octave multiple-scattering lighting and Henyey-Greenstein phase
- Cut cloud pass cost 3.5× with half-resolution rendering and temporal reprojection (Halton jitter, neighborhood-clamped history), with headless EGL rendering and GPU timer benchmarks run in CI

TECHNICAL SKILLS

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

Technologies: gRPC, Protobuf, Kubernetes, Redis, Docker, Prometheus, PyTorch, React, Node.js, AWS