Dean Shahin

deanshahin@berkeley.edu | linkedin.com/in/deannshahin | github.com/DeanShahin5

EDUCATION

University of California, Berkeley **Graduation: May 2028**

Bachelor of Arts, Computer Science **GPA: 3.91/4.0**

- Relevant Coursework: Data Structures & Algorithms, Computer Architecture, Operating Systems, Computer Security, Discrete Math, Linear Algebra, Probability & Statistics, Database Systems, Machine Learning

EXPERIENCE

Tesla Aug 2026 – Present

Software Engineer Intern

- Engineered Go/Kafka PLC telemetry pipelines for **5,000+** robots, cutting fault detection from **minutes** to **seconds**
- Built Optimus fault diagnosis agents on Google ADK with RAG and tool calling into PLC, Splunk, and Grafana data
- Architected Kubernetes agent services with Kafka worker queues and autoscaling, raising diagnostic throughput ∼**45%**
- Optimized ClickHouse/MySQL queries using partition pruning and materialized views, reducing p95 latency ∼**40%**

Crusoe May 2026 – Aug 2026

Software Engineer Intern

- Rearchitected GPU test cluster provisioning to reuse VMs across **12K+** GPUs, making cluster spin-up **7x** faster
- Built Go/gRPC APIs on Kubernetes (**90%** test coverage), cutting reused-VM reprovisioning from ∼**10 min** to ∼**5 sec**
- Cut cold-start latency for new GPU VMs **34%** by eliminating redundant cloud API calls and parallelizing initialization
- Designed a distributed VM lease protocol with a peer team's cleanup service, preventing teardown of active VMs
- Migrated Terraform state to locked GCS buckets with idempotent cleanup, avoiding ∼**$125K/year** in orphaned GPUs

Visa Jan 2026 – May 2026

Software Engineer (Contract)

- Engineered async Python/LangGraph agents generating financial models and slide decks, cutting prep time **50%+**
- Deployed containerized Temporal workflows to Azure via CI/CD, making agent runs durable with automatic retries
- Designed two-stage LLM evaluation (rules + judge) with PostgreSQL tracing, raising first-pass acceptance to **93%**
- Integrated Microsoft Graph REST APIs to automate Teams-to-SharePoint delivery for **40+** weekly analyst requests

Samsung Electronics Sep 2025 – Dec 2025

Machine Learning Engineer (Contract)

- Built a SmartThings anomaly detection pipeline on Kafka, FastAPI, and Kubernetes handling **1M+** device events
- Scaled Kafka consumers and Kubernetes inference workers for **4x** throughput and **40%** faster incident triage
- Reduced p95 inference latency **77%** by precomputing telemetry features and caching them in Redis
- Trained PyTorch time-series models on rolling device telemetry, improving detection **31%** over a rule-based baseline

PROJECTS

AI Inference Engine | *C++20, CUDA, vLLM, Kubernetes, Redis, Prometheus*

- Built a C++ inference router for game NPCs with KV-cache-aware scheduling, increasing LLM throughput **38%**
- Implemented token-budget admission control and priority queues, sustaining **128** concurrent requests without OOMs
- Engineered Kubernetes GPU autoscaling with Prometheus metrics, reducing idle GPU-hours **29%** under bursty load

Calendera (1st Place, Cal Hacks '25) | *Claude SDK, MCP, Next.js, Gmail & Google Calendar APIs*

- Built an MCP scheduling agent extracting events from Gmail, Slack, and SMS with **94%** accuracy on **500+** messages
- Implemented idempotent Google Calendar sync with conflict detection and retries, reducing duplicate bookings **91%**

SKILLS & INTERESTS

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

Infrastructure: Kubernetes, Docker, Terraform, GCP, Linux, CI/CD, Grafana, PostgreSQL, Redis

Frameworks & Libraries: gRPC/Protobuf, Kafka, Temporal, PyTorch, CUDA, NCCL, vLLM, Google ADK

Interests: Esports, Pokémon TCG, Anime, Football, Volleyball, Gym, Streaming, Trying new cuisines