Luis Sanchez

(916) 633-9090 luisanchez@berkeley.edu linkedin.com/in/suislanchez/ github.com/suislanchez x.com/iamsuislanchez

Portfolio: suislanchez.com

Education

University of California, Berkeley **Expected Graduation: Dec 2025**

B.S. Computer Science **GPA: 3.9**

Relevant Coursework: Artificial Intelligence**,** Machine Learning**,** Operating Systems**,** Data Structures/Algorithms**,** Great Ideas of Computer Architecture**,** UX/UI, Computer Security**,** Internet Architecture**,** Efficient Algorithms & Intractable Problems**,** Database Systems**,** Computer Vision, Deep Neural Networks

Activities: Cal Nu Jazz Ensemble Pianist, Berkeley CalHacks Mentor, CS61A & CS61B Academic Intern

Experience

Robo Labs - Forward Deployed Software Engineer **Sep 2025 - Present**

- Built MemoryAid, a low-latency RAG inference system serving 200+ users in real-time; optimized retrieval pipeline with caching and batched embedding lookups, deployed via containerized microservices in partnership with 3 neurologists

Palo Alto Labs - Founding Engineer **2025 - Present**

- Architected multi-stage ML inference pipeline for real-time basketball video analysis serving NBA coaches; implemented request batching and GPU-optimized model serving, reducing inference latency 40%; secured SAFE investment from UC Investments

Adobe - Software Developer Intern **June 2024 - August 2024**

- Built API middleware aggregating telemetry from Creative Cloud services into automated stability metrics and dashboards; reduced incident detection time by eliminating manual log analysis for engineering teams

Duolingo - Software Engineer Intern **May 2023 - July 2023**

- Deployed embeddings-based relevance model serving millions of inference requests daily; optimized throughput via batching strategies and model caching, validated 12% engagement lift through A/B experiment

Projects

Zigsaw (Y Combinator 25' Dedalus Hackathon 1st Place Prize Winner) **zigsaw.dev**

- Built LoRA fine-tuning and inference pipeline for video generation; optimized GPU memory utilization and implemented streaming inference for real-time ad generation, winning 1st place at Y Combinator Hackathon

LegalGPT Github Repo

- Graph-augmented legal outcome prediction system (EMNLP 2026 submission) combining GraphSAGE citation network embeddings with QLoRA fine-tuned Mistral-7B, achieving 0.80 AUROC on Supreme Court case prediction.

LuisOS Github Repo

- Original: Built operating system from scratch with Pintos framework, including kernel threading, virtual memory, and file system reliability enabling seamless multi-process execution and safe concurrent access

TransferAgent - *Contract (2025)* degreesight.com

- Built distributed data pipeline on AWS processing 3K+ university policies; implemented async job orchestration with containerized workers, achieving 84.8% automation rate at scale

Skills

- **Programming:** Python, TypeScript/JavaScript, Go, Rust, C++, SQL, C, Ruby, MySQL, Java, Kotlin, Scala, C#, Bash, NoSQL
- **Agentic AI & Automation:** Multi-Agent Systems, Agent Orchestration, AI Workflow Automation, Autonomous Browser Agents, Event-Driven Agents, Memory Systems, Tool-Calling, Long-Context Reasoning, MCP (Model Context Protocol), Bedrock Agents, LLM-Powered Planning, Observability for Agents
- **AI/ML Stack:** RAG Pipelines, Vector Databases (Pinecone, Milvus, Weaviate), LangChain, LlamaIndex, HuggingFace, Bedrock (Claude / Llama / Stable Models), Multimodal Models (VLMs), Embeddings, Retrieval Optimization, Semantic Search, Structured Outputs, Knowledge Graph Augmentation, Lightweight Fine-Tuning (LoRA/QLoRA) Pandas, NumPy
- **Full-Stack & APIs:** Next.js, React, Node.js, gRPC, GraphQL, WebSockets, Supabase, Firebase, OAuth2, WebAuthn, Edge Functions, Agile, Serverless Architectures, End-to-End, AngularJS, Flask, Django, Laravel, VueJS, Spring Blockchain: web3, ethers.js, smart contracts
- **Systems & Infra:** Linux, Docker, Kubernetes, Terraform, NGINX, GCP Event-Streaming, Redis, Postgres, Kafka (basic), AWS (EC2, S3, Lambda, Bedrock), GCP (Cloud Run), Cloudflare Workers, Maven, Azure, Nexus, IaaS, Oracle, Prometheus, Kibana, Grafana, Ansible, Packer,
- **Mobile & Client:** React Native, Expo, Android (Kotlin), iOS (Swift), Native Device APIs, Mobile App
- **Product & Tooling:** Vercel, Figma, Postman, Cloudflare, Airflow, dbt, Playwright, Browser Automation, Web Scraping, Telemetry/Tracing for Agents, Greenfield