

Diego Sanchez Tirado

Stanford, CA · (984) 895-6386 · dsanchezt22@gmail.com

Interest: computer architecture / memory systems

EDUCATION

Stanford University — *B.S. Electrical Engineering, expected May 2029*

Sept. 2025 – May 2029

Coursework: Intro to Making (EE40M); Vector Calculus for Engineers; Programming Abstractions; Probability; ODEs for Engineers.

RESEARCH

Area-Scalable Memory Analysis for LLM Inference

Jan. 2026 – Present

Researcher, Stanford Engineering Robust Systems Group (Prof. Subhasish Mitra)

- Model LLM-inference throughput, latency, and memory in Python under competing chip memory organizations, relating capacity, bandwidth density, and mesh/NoC topology to batch size and context length.
- Compare mesh-distributed (area-scalable) and HBM backends with tensor- and pipeline-parallel (TP/PP) planning, and report compute, memory, and interconnect breakdowns in bandwidth-bound long-context KV-cache regimes.

EXPERIENCE

Amazon Lab126 — *Software Engineering Intern, RAVEN → Governed Knowledge Engine*

June 2026 – Sept. 2026

- Built a daily idempotent ingest of Outlook, Slack (DMs and channels), and Confluence into a searchable Markdown corpus.
- Implemented an LLM-as-judge classifier with a hybrid preset and auto-discovered taxonomy, typed links, and a conflict queue; reached $\geq 80\%$ accuracy on a labeled eval and indexed 2,000+ items on the first production run.
- Re-architected the pipeline as a Governed Knowledge Engine with MindCraft as the retrieval source of truth and prompt-injection plus PII admission gates before indexing for RAG agents.

Amazon Lab126, Product Integrity / Reliability Lab

June 2026 – Sept. 2026

Humanoid Robotics Apprenticeship, Dexmate “Labot”

- Ran VR-headset teleoperation demos on the Dexmate bimanual humanoid under a Senior Test Automation Engineer and labeled datasets for manipulation-policy training.
- Trained and evaluated ACT and diffusion imitation policies on physical bimanual tasks, then applied residual RL (TD3) beyond demonstrated behavior.

Duna Strategies — *Developer*

April 2026 – Present

- Build FStea AI Python collectors and a Node cockpit; verified Module 7 at 102 launch rows, 48 articles, 8 chains (mean quality 0.90) and fused Modules 7–11 into Module 4 scoring.

Code Four (YC X25) — *Head GTM Intern*

July 2025 – Oct. 2025

- Built GTM automation in n8n, Apollo, and HubSpot with AI agents/scrapers covering 10,000+ police contacts and email/call sequencing.
- Reached 1,100+ departments, yielding 70+ qualified meetings, 10+ MOUs, and 520+ officers.

PROJECTS

Software and Electrical Hardware Development — *Independent*

Aug. 2022 – Present

- Design a typing-force piezoelectric energy harvester with rectification and storage, and instrument voltage and power versus force and geometry.
- Build a split ergonomic keyboard (CAD, 3D-printed enclosure, custom PCB) and a free-air ANC system with a microphone array and LiDAR on embedded hardware.

Cyber Threat Intelligence Research — *Duke Pratt & Digi Americas (Prof. Arturo F. Ehuan)*

Mar. 2024 – July 2025

- Co-authored *LATAM Financial Threat Landscape 2025* with DigiAmericas and Recorded Future (LockBit, MisPadu, CL0P, Blind Eagle, HoraBot) and maintained a Git-tracked IOC/TTP pipeline mapped to MITRE ATT&CK.

TECHNICAL SKILLS

Hardware: Circuit design, PCB, soldering, CNC, 3D printing, laser cutting, mic-array/LiDAR, piezoelectric instrumentation

Architecture & modeling: HBM, NoC, and TP/PP modeling; Python models of capacity, bandwidth density, and KV-cache regimes

Languages: Python, C++, JavaScript, TypeScript, HTML/CSS; Spanish and English (fluent)

ML / systems: LLM-as-judge pipelines, MindCraft RAG, prompt-injection/PII gates, ACT, diffusion, TD3, FastAPI, Next.js/React

CAD / tools: Onshape (certified), SolidWorks (FEA), Git, n8n, Apollo.io, HubSpot; PLTW IED; MIT BSWI Python & GitHub

Selected: 5th/390, NC State Championship (FTC RoboKnights 8569)