

David Cho

Fremont, CA | (510) 789-3806 | david12cho@gmail.com | www.linkedin.com/in/dcho12

SUMMARY

2025 CS grad from Cal Poly SLO who builds AI applications for the operations teams that have to use them daily. Experienced sitting with non-technical stakeholders to map manual workflows, prototyping fast, and shipping automations that get adopted rather than abandoned. Fremont resident and available onsite immediately.

EDUCATION

California Polytechnic State University, San Luis Obispo
Bachelor of Science in Computer Science

September 2021 - September 2025

WORK EXPERIENCE

Software Engineering Intern at Liontalk - Remote

May 2026 - July 2026

- Built and deployed an AI translation platform used daily by a non-technical operations team, designing the review interface around how they actually worked and writing the documentation they used to run it without engineering support
- Integrated multiple enterprise AI platforms behind a provider-agnostic abstraction, evaluating OpenAI and DeepSeek on cost and output quality and cutting inference spend roughly 90% at comparable results
- Built an evaluation framework combining structural metrics with an LLM judge to measure accuracy and reliability before release, validating 6 of 8 configurations for production and gating regressions across 125 tests in CI

Software Engineering Intern at 1-Decision - Austin, TX

February 2025 - May 2025

- Partnered directly with early users of an AI consumer product to map how they were actually using it versus how it was designed, translated the gap into a prioritized backlog, and communicated tradeoffs clearly to non-technical stakeholders
- Built a production Python REST backend integrating multiple third-party AI APIs with resilient error handling, rate-limit management, and retry logic, then designed a semantic caching layer handling 10K daily requests at a 70% hit rate that cut average response time from 1.8s to 1.0s and tripled throughput
- Owned the Stripe subscription integration end to end for 30 active users, implementing idempotent webhook processing and failure recovery with full test coverage, and shipped it with zero post-launch payment breaks

PROJECTS

Automated Financial Operations Dispatch System | *Python, TypeScript, Azure*

November 2025 - January 2026

- Embedded with a business's operations team to map a fully manual billing workflow, defined success metrics with them before writing any code, and shipped a 14-stage GPT-4o automation to production that cut manual work 80%
- Automated document parsing and intelligent routing across email and PDF intake, extracting structured data from unstructured client correspondence and integrating with the systems the team already used
- Caught an adoption problem where the AI was confidently wrong in ways that eroded trust, diagnosed it through direct conversations with the team rather than from logs, and redesigned it to keep a human check on outgoing work

Onboard: AI Data-Cleaning Agent | *Python, Anthropic Claude API, Pandas, SQLite*

March 2026 - May 2026

- Built an AI agent that turns messy business spreadsheets into schema-validated data by inspecting the input, writing its own pandas transformations, and executing them, cutting an inventory cleanup task from roughly 5 hours to 8 minutes
- Designed the agent to escalate ambiguous records like conflicting date formats and impossible values to human review rather than guess, surfacing 170 would-be errors across a 12-file benchmark, and kept cost flat at \$0.40 per file regardless of size
- Made every run auditable by storing each transformation as a content-addressed pure function in a provenance ledger, so any row can be re-verified byte-for-byte months later when someone asks where a number came from

Course Companion; RAG Assistant | *Python, LangChain, Pinecone, FastAPI, React*

January 2025 - March 2025

- Built a retrieval-augmented assistant over an institutional document corpus using LangChain and a Pinecone vector database, letting users ask questions in plain English and get answers grounded in real source data rather than hallucinated ones
- Engineered the ingestion pipeline end to end, parsing unstructured PDFs into structured records, generating embeddings, and indexing them for semantic search, then tuned retrieval parameters against real query patterns to improve relevance
- Demonstrated the assistant live to non-technical user groups, adapting the explanation to the audience and using their reactions to identify which parts of the experience needed rework before the next release

TECHNICAL SKILLS

Languages & Frameworks: Python, TypeScript/JavaScript, SQL, React, Node.js, FastAPI, Flask, REST APIs, full-stack dev

AI Systems: LLM integration (Claude, GPT, DeepSeek), agentic workflows and tool use, RAG pipelines, vector databases, MCP

Infrastructure: AWS, Azure, GCP, Docker, Kubernetes, PostgreSQL, Redis, CI/CD, GitHub Actions, Grafana, Prometheus