SANJAY MANIVASAGAM

San Francisco, CA | (530) 979-6675 | sanmanivas@ucdavis.edu
github.com/Sanjaayyy7 | linkedin.com/in/sanjaymanivasagam

AI / ML Engineer — LLM agent systems, evaluation infrastructure, RAG pipelines | B.S. Data Science, UC Davis 2026

EXPERIENCE

AI Engineer Sep 2025 to Present
ClearAgent *San Francisco, CA (Remote)*

- Built a **LangGraph** multi-agent orchestration service on a planner-executor pattern, coordinating **3** specialist subagents (verification, audit, policy enforcement) for autonomous **tool-using AI agents** across **2 LLM providers** (**OpenAI**, **Anthropic**).
- Engineered a semantic **retrieval** pipeline pairing **vector embedding search** with **cross-encoder reranking** over the EU AI Act corpus, returning the single governing clause for an agent action instead of a ranked candidate list.
- Designed an **event-sourced audit ledger** with cryptographic chain-of-custody capturing token usage, **tool-call traces**, and policy-violation flags on every agent invocation, turning opaque agent behavior into queryable **observability** data.
- Built a model-agnostic **evaluation harness** emitting **structured outputs** for policy adherence, audit completeness, and decision provenance, replacing manual review with deterministic, reproducible scoring.

AI/ML Engineering Intern Jun 2025 to Sep 2025
DataCorp Traffic Private Limited *Remote, UK*

- Shipped an **LLM report-automation pipeline** (Python, OpenAI API, Docker) running in **production** across **200+ monthly reports** for a multi-tenant SaaS platform, cutting delivery from 3-5 days to **4 hours (85%)** and scaling **15 to 30 enterprise accounts** without added headcount.
- Built a schema-aware, tenant-scoped **SQL generation** layer with validation and **ETL** workflows across **50K+ queries**, eliminating cross-tenant data access and cutting runtime errors **80%** and support tickets **40%**.
- Owned the **full-stack** build as **1 of 2 engineers**: template validation through **JWT-authenticated REST APIs** with role-based access control and audit logging.

Software Engineering Intern, Agents & Frontend Jun 2024 to Sep 2024
Headstarter *San Francisco, CA*

- Extended a **code-generation AI agent** to reason across multi-file GitHub repositories and propose cross-file refactors, cutting time from agent run to first useful suggestion **20%+**.
- Shipped a **React**/**TypeScript** agent review dashboard consolidating PR review into a single workflow, cutting **clicks-to-accept 50%** with adoption by most active users inside **2 weeks**.
- Shipped features to **production** on **AWS** via **Docker**, **Kubernetes**, and **Kafka**-backed pipelines with Git-based **CI/CD**, in **cross-functional collaboration** with senior engineers and the founder.

TECHNICAL PROJECTS

Groundtruth, Agent Red-Teaming & Evaluation Harness | *Python, LLM Evaluation, Detector Calibration, Reproducible ML, CI/CD*

- Built an offline, deterministic **red-teaming** and **evaluation harness** for **tool-using LLM agents**, then measured its own detectors against **68 hand-labeled traces**: micro **F1 0.923** (**precision 0.95** / **recall 0.89**) across **6** failure categories including secret exfiltration and instruction hijacking.
- Traced a false safety result to the harness: at temperature 0 a stateless adapter rebuilt byte-identical prompts, guaranteeing stalls at any step budget (**23 identical tool calls in one 24-step run**). A stateful rebuild across a **5-cell matrix** took stalls **9/9 to 0/9**, exposing **2** models that exfiltrated a key or transferred funds, **inverting the leaderboard**.
- Benchmarked rules against **LLM-as-a-judge**: a local **8B** judge scored **0.36 precision** against the detectors' **0.95** on identical items. Pre-registered predictions (one falsified); pinned every known miss to a test so **regressions** fail **CI** — **170 automated tests**, deterministic artifacts.

Chain-of-Thought Prompting vs. Fine-Tuning for Table Reasoning | *Python, PyTorch, Hugging Face, FLAN-T5*

- Benchmarked **11 conditions** spanning zero-shot, plain and structured **CoT prompting**, and answer-only versus rationale-trace **fine-tuning** on **FLAN-T5 transformer** models (**250M**/**780M**) over WikiTableQuestions with dual seeds, isolating a defensible negative result: zero-shot (**0.241 EM**) beat every intervention.
- Confirmed it with a **TabFact cross-dataset generalization study** and **McNemar's significance testing**, showing **fine-tuning** overfit to task format and fell below the **0.551** majority-class floor, with error attribution across lookup, aggregation, and multi-hop reasoning.

Interrogating Agents, Multi-Agent LLM Evaluation Harness | *Python, Ollama, ChromaDB, RAG, LLM-as-a-Judge*

- Built a two-channel **RAG** harness (**ChromaDB**, MiniLM **embeddings**, dialogue-phase filters) over **18** technique cards, scoring matched control/treatment legs in **1 batched blinded-judge call** returning **structured JSON**, cancelling judge calibration bias.
- Owned reproducibility for a **4-person** team: offline **Ollama** (**Llama 3.1 8B**) **inference** at zero API cost, one-command setup, and a **GitHub Actions CI** matrix across **3 OSes** and **2 Python versions**.

L-Store, HTAP Database Engine Built from Scratch | *Python, Columnar Storage, Bufferpool, Concurrency*

- Built an **HTAP** database engine in Python (**5-person** team): lineage-based columnar storage, an **LRU bufferpool** with dirty-page write-back, background merge, and **concurrent multithreaded transactions** under record-level **two-phase locking**, validated across **3** autograded milestones.

EDUCATION

University of California, Davis Jun 2026
Bachelor of Science in Data Science | GPA: 3.60/4.0 | 3x Dean's Honors List
Relevant Coursework: Artificial Intelligence, Algorithms, Data Processing Pipelines, Databases, Big Data & High Performance Computing

TECHNICAL SKILLS

Languages: Python, SQL, TypeScript, JavaScript, C++, R
ML & AI: PyTorch, Hugging Face Transformers, LangGraph, LangChain, RAG, Vector Embeddings, Cross-Encoder Reranking, Fine-Tuning, LLM-as-a-Judge, Multi-Agent Orchestration, Agentic Systems, Prompt Engineering, Model Evaluation, Model Context Protocol (MCP), LLM APIs (OpenAI, Anthropic), TensorFlow, scikit-learn
Data & Experimentation: Pandas, NumPy, ETL Pipelines, Experiment Design, Statistical Significance Testing, A/B Testing, Feature Engineering
Infrastructure: Docker, Kubernetes, AWS (EC2, S3, Lambda, RDS), PostgreSQL, ChromaDB, Kafka, Redis, FastAPI, REST APIs, GitHub Actions, CI/CD, Distributed Systems