

CHRISTINE STRAUB

ML/AI Engineer — AI Systems Architecture — MLOps — LLMOps

☎ (949) 527-5247 ✉ christinemstraub@gmail.com [in straubchristine](https://www.linkedin.com/in/straubchristine) [github christinestraub](https://github.com/christinestraub) [kaggle christinemstraub](https://www.kaggle.com/christinemstraub)
San Clemente, CA

Summary

Senior AI/ML Engineer with 8+ years building production AI systems across **defense, healthcare, finance, and other enterprise domains**, spanning from classified DoD intelligence systems, HIPAA-compliant healthcare applications, and enterprise-scale document processing. Specializes in **applied agentic workflows, LLM/VLM productization, and MLOps/LLMOps infrastructure**, with deep expertise designing **advanced RAG pipelines, distributed edge AI architectures, and multi-agent orchestration systems**.

Education

University of California, Berkeley

Bachelor of Arts in Computer Science

Berkeley, CA

December 2017

University of California, Berkeley

Bachelor of Arts in Cognitive Science

Berkeley, CA

December 2017

Work Experience

Technical Lead, Machine Learning

Bespoke Labs AI

February 2026 – Present

Mountain View, CA

Applied AI research lab pioneering data and RL environment curation for training and evaluating agents.

- **RL Environment Design & Task Authoring:** Lead creation of high-quality RL training and evaluation environments for multi-turn tool-calling agents, authoring task specifications, **verifier functions**, and reward signals, developing repeatable recipes that scaled validated environment production while maintaining rigorous quality bars.
- **Agent Failure Analysis & Benchmark Development:** Analyzed **agent rollouts** across diverse task types to diagnose **reward hacking** and surface failure modes, translating insights into externally released **benchmark suites** packaged for the company's evaluation platform and external-facing dashboards.
- **RL Training & Evaluation Infrastructure:** Built reproducible **PyTorch/Hugging Face TRL** training and evaluation pipelines on **GCP/AWS** for small-scale agent verification runs, establishing automated evaluation protocols and **HITL** review workflows hardening environments against reward exploitation.
- **Core Technologies:** Python, PyTorch, JAX, Hugging Face (Transformers, Datasets, TRL), vLLM, Verifiers, Bespoke Curator, Docker, GCP.

Technical Lead, Machine Learning

Medici Land Governance

July 2025 – Present

Washington, DC

Develops blockchain-based land administration and titling systems to modernize property ownership records.

- **Gemini Document Intelligence Pipeline:** Designed and productionized an end-to-end **Gemini** (Vertex AI) document processing pipeline for legal land records—deeds, liens, mortgages, and court dockets—combining batch and realtime inference with tiered escalation (Flash/Lite → Pro) to maximize throughput while controlling cost.
- **Open-Source VLM Experimentation (Qwen):** Fine-tuned and evaluated open-source vision-language models (**Qwen3-VL** via PEFT/LoRA) against proprietary APIs; integrated self-hosted inference into a hybrid OCR → VLM → frontier-model cascade to identify the most cost-effective path per document difficulty.
- **Cost-Optimized Multi-Model Routing:** Built a dynamic routing layer across **Gemini, Claude**, and open-source VLMs that routes clean pages to cheaper models and escalates only on handwriting/degraded scans—cutting inference costs by ~70% while holding **95%+** field extraction accuracy on complex legal documents.
- **Core Technologies:** Gemini (Vertex AI Batch/Realtime), Qwen Models, Anthropic Models, GCP (Cloud Run, BigQuery, Vertex AI, Document AI, GKE), PaddleOCR, Kedro, vLLM, Kafka, Elasticsearch, MLflow.

Senior Machine Learning Engineer (CV/Edge AI)

RIOS Intelligent Machines

November 2023 – July 2025

Palo Alto, CA

Builds computer vision and ML infrastructure for robotic process automation in manufacturing and industrial environments.

- **End-to-End MLOps Pipeline:** Architected **Metaflow** workflow reducing deployment time by **70%**, enabling **60 FPS** real-time inference via **Kubernetes** auto-scaling for industrial video anomaly detection processing **10M+ daily images**.

- **Active Learning Loop:** Integrated **FiftyOne** and **Encord** to automate data selection, reducing manual annotation time by **60%** via uncertainty sampling and bidirectional annotation-dataset pipeline.
- **Edge AI Optimization:** Designed custom quantized computer vision operators (**YOLO v8/v9**), achieving **5x inference speedup** on edge devices (Jetson/Orin) through **ONNX/TensorRT** optimization.
- **GPU Training Infrastructure:** Implemented GPU-optimized mixed-precision training (**FP16/BF16**) in **PyTorch**, increasing throughput by **40%**.
- **Core Technologies:** **YOLO v8/v9**, **Metaflow**, **FiftyOne**, **Encord**, **PyTorch**, **Kubernetes**, **Docker**, **Weights & Biases**, **ONNX**, **TensorRT**.

Senior Applied AI Engineer

Unstructured IO

May 2023 – April 2025

San Francisco, CA

Develops open-source and API-based tools that transform unstructured documents into structured data.

- **Multi-Agent AI System:** Architected production **Pydantic AI + MCP** (Model Context Protocol) orchestration system, creating custom MCP servers interfacing with unstructured APIs for intelligent document processing workflows.
- **VLM Benchmarking & Integration:** Evaluated **10+ VLM providers** (**Claude 3.5/3.7**, **GPT-4o**, **Gemini 1.5/2.0**), engineered prompt optimization techniques dramatically improving table structure recognition and image text extraction accuracy.
- **Layout Detection Fine-Tuning:** Fine-tuned layout models (**YOLOX**, **YOLO-NAS**, **Detectron2**) on **11,000+ technical PDFs**, achieving **10%** accuracy increase and **13%** reduction in missing text detection.
- **OCR Pipeline Optimization:** Led OCR enhancement across **Tesseract/PaddleOCR**, implementing preprocessing (scaling, contrast, PSM optimization) improving accuracy by **15%** and reducing missing text by **10%**.
- **Document Layout Analysis:** Engineered sophisticated element extraction pipeline separating images, figures, and tables as standalone base64-encoded blocks, implementing xy-cut bounding-box sorting algorithm for improved element ordering and document reconstruction.
- **Production Impact:** Resolved **300+ bugs**, reviewed **500+ PRs**, implemented memory optimizations (chunked PDF processing), intelligent encoding detection for enterprise-scale processing.
- **Core Technologies:** **YOLOX/YOLO-NAS/Detectron2**, **Tesseract/PaddleOCR**, **OpenCV**, **LangChain**, **LlamaIndex**, **Claude/GPT/Gemini APIs**, **Pydantic AI**.

Lead Software Engineer – AI/ML (DoD)

Sapient Logic

Jan 2021 – Feb 2024

San Diego, CA

AI-enabled intelligence systems for the Department of Defense across multiple classified programs.

- **Mobile Intelligence Translation App:** Architected mobile OCR platform enabling real-time document processing in air-gapped field environments using **Google ML Kit + Tesseract4Android** for **French/Arabic/Chinese** materials during sensitive site exploitation operations.
- **AI-Powered Intelligence Management Platform:** Led development of microservices-based intelligence requirements system, engineering semi-automated validation workflows and ML-powered classification that streamlined intelligence processing across security classifications, integrating with **GCGS-J**, **ICSF**, and **Tactical Awareness Kit**.
- **Cybersecurity Threat Intelligence App:** Developed **BERT/SBERT**-based semantic similarity engine automating **MITRE ATT&CK** framework mapping to defense mechanisms, reducing manual security gap analysis time by **94%** through multi-vector NLP pipeline (**Word2Vec**, **GloVe**, **Transformers**).
- **Secure Cross-Domain Architecture:** Designed one-way secure transfer mechanisms for classified environments, implemented military-grade encryption protocols, and established DevSecOps CI/CD pipelines with security documentation for ATO compliance.
- **Technical Leadership:** Directed cross-functional teams through full SDLC for classified systems, created architecture documentation for defense applications in network-denied tactical environments, and established test management frameworks for multiple DoD programs.
- **Core Technologies:** **Google ML Kit**, **Tesseract OCR**, **BERT/SBERT**, **PyTorch**, **Word2Vec/GloVe**, **Microservices**, **Docker**, **React**, **MITRE ATT&CK**, **Air-gapped systems**, **Cross-domain solutions**.

Lead Software Engineer – AI/ML Infrastructure

Sapient Logic

April 2021 – Feb 2023

San Diego, CA

HIPAA-compliant Electronic Health Record system for hospital patient management.

- **HIPAA-Compliant Cloud Architecture:** Architected complete EHR solution on **AWS** (**EC2**, **RDS**, **VPC**, **ECS**, **ECR**) with comprehensive security controls including **KMS** encryption, **MFA** (**Vonage/SendGrid**), role-based access, audit logging via **CloudWatch**.
- **Network Security & Infrastructure:** Deployed **Sonicwall Firewall** with secure VPC architecture, SSL certificates for encrypted transfers, automated alarm systems for **RDS/EC2** monitoring across **Nginx/Django** stack.

- **Patient Data Management:** Engineered comprehensive platform tracking medical histories, demographics, documents from registration through discharge, with real-time ER analytics dashboards for patient flow optimization.
- **Healthcare Interoperability:** Integrated **CollaborateMD**, **eClaimStatus**, **epowerdoc** using **HL7/FHIR** protocols for seamless patient management and claims processing.
- **Technical Leadership & Team Management:** Led engineering team implementing physician-requested improvements, conducted technical interviews and hired full-stack, DevOps, and QA engineers, created comprehensive product requirement documentation, and established **GitHub Actions** CI/CD pipeline with DevOps collaboration.
- **Core Technologies:** **AWS** (EC2, RDS, VPC, ECS, ECR, CloudWatch), **Django**, **Vue.js**, **React**, **PostgreSQL**, **HL7/FHIR**, **Docker**, **GitHub Actions**.

AI Software Architect

June 2022 – June 2023

Speechlab AI

San Francisco, CA

AI-powered audio/video localization platform enabling speech-to-speech translation and synthetic dubbing.

- **Cloud-Native Localization Architecture:** Designed end-to-end **AWS** serverless infrastructure (**Lambda**, **EventBridge**, **ECS**, **App Runner**, **S3**) for event-driven media processing, reducing dubbing costs by **80%** and accelerating timelines from weeks to hours.
- **ML API Gateway:** Built scalable API layer integrating speech recognition, machine translation, and text-to-speech ML models with performance optimization and model monitoring for reliable multilingual workflows.
- **Core Technologies:** **AWS** (Lambda, EventBridge, S3, ECS, App Runner, Cognito, CloudWatch, SES), **Node.js**, **Express.js**, **React**, **TypeScript**, **MongoDB**.

Machine Learning Engineer

May 2021 – May 2022

Collegis Education

Chicago, IL

Data analytics solutions optimizing enrollment growth and student success for higher education institutions.

- **Event-Driven ETL Pipeline:** Architected near-real-time pipelines using **Google Cloud Functions/Cloud Run** ingesting student interaction data from Phoneburner, Five9, LMS Canvas, reducing data latency by **85%**.
- **Modern Data Stack:** Orchestrated multi-source integration via **Fivetran** and **DBT** for **BigQuery** modeling, implementing automated schema evolution and 100% data lineage for governance.
- **Speech Analytics:** Engineered Call Center Intelligence System using **Cloud Speech-to-Text** and **Natural Language APIs** for automated transcription and sentiment analysis.
- **Core Technologies:** **GCP** (Cloud Functions, Run, BigQuery, Speech-to-Text, Natural Language API), **DBT**, **Fivetran**, **Python**, **ThoughtSpot**.

Software Engineer

September 2017 – April 2021

Moody's Analytics

Silicon Valley, CA

Catastrophe risk modeling and location intelligence platforms for insurance and financial services clients.

- **Software Engineering:** Contributed to **RMS(one)**, **RiskLink**, and **RiskBrowser** platforms, building client-facing features for catastrophe risk assessment and portfolio analysis enabling insurers to evaluate multi-billion dollar portfolios.
- **Core Technologies:** **Java**, **TypeScript**, **Python**, **Geospatial visualization**, **Data pipelines**, **Risk modeling**.

Technical Skills

Agentic AI & GenAI Expertise: Multi-Agent Systems (MAS), Agentic Workflows, RAG (GraphRAG/HyDE), Vision-Language Models (VLMs), ReAct, Chain-of-Thought (CoT), Prompt Optimization (DSPy), Model Context Protocol (MCP), Autonomous Agents

ML/AI/NLP Expertise: Computer Vision (Object Detection/Segmentation), Natural Language Processing (NLP), Transformer Architectures (BERT/ViT), Reinforcement Learning (RL), Model Fine-tuning (LoRA/QLoRA), Quantization (GGUF/AWQ), OCR, Edge AI, Distributed Training

Agentic Frameworks & LLMs: LangGraph, CrewAI, Microsoft AutoGen, Pydantic AI, LangChain, LlamaIndex, OpenAI Swarm, Claude 4 (Sonnet/Opus), GPT-4o/o1/o3, Gemini 2.0 (Flash/Pro), Llama 3.x, Ollama, vLLM

LLM Evaluation & Red-Teaming: Ragas (RAG Assessment), LangSmith, Arize Phoenix, DeepEval, Human-in-the-loop (HITL) Evaluation, Hallucination Detection, Bias & Safety Testing

Deep Learning & MLOps: PyTorch, TensorFlow, Keras, Hugging Face Transformers, ONNX, TensorRT, Metaflow, Weights & Biases, MLflow, Ray Serve

Computer Vision: YOLO (v8/v9/NAS), Detectron2, Tesseract, PaddleOCR, OpenCV, FiftyOne, Encord, Google ML Kit, SAM (Segment Anything Model)

Cloud & Infrastructure: AWS (Lambda, SageMaker, Bedrock, ECS, EventBridge), GCP (Vertex AI, Cloud Run, BigQuery, Document AI), Kubernetes (K8s), Docker, Terraform

Fullstack & Databases: Python, Node.js, TypeScript, Rust, PostgreSQL, MongoDB, Pinecone/Weaviate/ChromaDB (Vector Search), Elasticsearch, Kafka

Certifications

Deep Learning Specialization (DeepLearning.AI, Stanford) | Machine Learning Specialization (DeepLearning.AI, Stanford)
MLOps Specialization (DeepLearning.AI) | AWS Cloud Practitioner (AWS) | IBM Data Science (IBM)
Google Data Analytics (Google) | Business Intelligence (Google)