

Jiawei Tang

☎ (510) 292-1396 | ✉ jiaweitang@berkeley.edu | [in linkedin.com/in/jiaweitang22](https://www.linkedin.com/in/jiaweitang22)

EDUCATION

University of California at Berkeley

Dec 2026

B.S. in Electrical Engineering and Computer Sciences (EECS)

GPA: 4.0/4.0

- **Relevant Courses:** Data Structures and Algorithms, Discrete Mathematics and Probability, Computer Architecture, Efficient Algorithms, Computer Security, Machine Learning, Optimization Models, Databases, Networking

EXPERIENCE

SWE Intern, Samsara | San Francisco, CA

May–Aug 2026

- Architected a DINOv3 computer-vision classifier that replaced a production vision-LLM for cargo-theft detection, raising accuracy from 84% to 98% and cutting missed thefts by ~90%.
- Built a reusable ML framework for computer-vision apps, unifying multi-GPU training (Ray Train) and autoscaling, request-batched serving (Ray Serve); reduced latency ~10x from 7.6s to sub-second and cut per-prediction cost ~99% by moving from paid API calls to local inference.
- Shipped three production vision models—door state, fill level, and cargo shift—enabling theft detection, load optimization, and damage/insurance workflows on a low-cost camera platform positioned against premium edge-AI hardware.

SWE Intern, TikTok | San Jose, CA

Jan-May 2026

- Developed backend infrastructure for AI-powered creative agents in TikTok Symphony’s Creative Studio, enabling advertisers to auto-generate video concepts and drafts and supporting MVP launch.
- Architected MCP-based backend services for multi-step video generation, handling 100 requests/second with high success rate by introducing asynchronous workflows, fault-tolerant retries, and secure cross-service integrations.

SDE Intern, Amazon Web Services | Seattle, WA

May-Aug 2025

- Implemented automated monitoring dashboard for model availability, usage metrics, and deployment tracking across Bedrock Marketplace and SageMaker JumpStart, supporting 1,000+ models across 37 AWS commercial regions
- Designed real-time ETL pipeline with 100% data capture rate for business intelligence dashboards, saving Product Managers and Solutions Architects 20+ hours weekly by eliminating manual tracking

MLE Intern, Viridien | London, UK

Jun-Sep 2024

- Built automated table extraction system combining LLMs with fine-tuned object detection models on curated datasets, achieving 60% improvement in extraction accuracy and 75% reduction in processing time
- Designed and deployed agentic AI system utilizing Chain-of-Thought prompting to parse tabular data, reducing model hallucinations and improving robustness across document types

Machine Learning Developer, Launchpad | Berkeley, CA

Sep 2023-present

- Engineered a legal database query system using LLM agents, reducing query resolution time by 90% and increasing client satisfaction by 25% for small-to-medium law firms
- Implemented Retrieval Augmented Generation with vector databases, achieving a 10x improvement in query response time and generating \$22,000 in revenue

RESEARCH

Research Assistant, MIT Database Group | Remote

Jun 2025-Present

- Developed benchmark framework (KramaBench) for evaluating LLM agents on end-to-end data science pipelines and analytical workflows across diverse domains
- Designed LLM agents and implemented systematic validation pipelines to ensure benchmark accuracy across multiple difficulty levels

Research Intern, MIT CSAIL | Cambridge, MA

Jun-Aug 2022

- Designed and trained generic Entity Resolution model for matching and deduplicating data across diverse domains, achieving comparable performance to specialized models with 30% less training data and 25% less storage
- Published as first author: “Generic Entity Resolution Models” at Table Representation Learning Workshop @ NeurIPS 2022, introducing a scalable approach for multi-domain data integration

Research Intern, Qatar Computing Research Institute | Doha, Qatar

Jun-Aug 2021

- Developed Sevi, an end-to-end Speech2Visualization system using Transformer encoder-decoder architecture, enabling natural language-driven data visualization with 85% accuracy
- Published as first author: “Sevi: Speech-to-Visualization through Neural Machine Translation” at SIGMOD 2022 (demo)

TECHNICAL SKILLS

Programming Languages: Python, Java, C/C++, Go, SQL, JavaScript/TypeScript, Swift, React

Machine Learning: PyTorch, TensorFlow, Scikit-learn, Hugging Face Transformers, LangChain, OpenCV, Ray

Data Science & Cloud: Pandas, NumPy, Matplotlib, Jupyter, AWS (EC2, S3, Lambda, Athena, QuickSight, Bedrock, SageMaker), Docker, Kubernetes, Git, CI/CD, Terraform, Redis, PostgreSQL