

Jacob Cutter

Bay Area, CA | 510-861-0706 | jecutter92@gmail.com | [LinkedIn](#) | [Website](#)

SKILLS

Software/Apps: Python, C++, UNIX/Shell, Git, Flask, FastAPI, Kubernetes, Docker, Splunk, Cursor

Data: (Postgres, My)SQL, MongoDB, Redis, Pandas, Matplotlib, Seaborn, Statistical Analysis/Modeling

ML/AI: Deep Learning (PyTorch), Transformers/LLMs, Prompt Engineering, LlamaIndex, Scikit-Learn

Computing: AWS (S3, EC2, Athena), GCP/Firestore, Microsoft Azure, Lambda, SGE/SLURM Clusters

Specialties: Generative AI, NL(P/U/G), Agentic Reasoning, Speech AI (ASR/STT), Audio/Signal Processing

Soft Skills: Scrappy, Agile, Product-Oriented, Cross-Functional, Collaborative

EXPERIENCE

Salesforce

Software Engineering LMTS (Lead Member of Technical Staff)

Jan 2026 - Present

Software Engineering SMTS (Senior Member of Technical Staff)

Sep 2024 - Jan 2026

- Served as Tech Lead to build a low-latency, LLM-based reasoning engine for full duplex conversational Voice Agents, powering next-gen chat and telephony applications in Agentforce Voice
- Significantly influenced Voice Agent design choices and led execution from prototyping to product GA
- Carried out substantial integration work and collaborated with partner teams to enable features across the Salesforce stack, including Service Cloud, Data Cloud, and Agentforce Observability
- Debugged customer issues across production services via Splunk logs and telemetry dashboards
- Developed an end-to-end Voice Agent evaluation pipeline to bolster Trust & Safety guardrails
- Created an audio augmentation pipeline for configurable synthesis of realistic phone call audio, allowing for robustness evaluations and parameter tuning
- Performed extensive benchmarking of open- and closed-source Speech-To-Text (STT) systems, including self-deployment and inference optimization on Lambda and AWS GPU instances

Tenyx (acquired by Salesforce, September 2024), Member of Technical Staff

May 2024 - Sep 2024

- Built a configurable synthetic customer and chat generation pipeline to evaluate the quality of AI agent systems, fine-tune LLM responses, and simulate failure modes to improve customer experience
- Implemented an auto-evaluation system that leverages LLMs to dynamically score AI models on generative tasks (e.g. agent behavior, synthetic customer behavior, and slot filling/entity extraction)
- Developed data/ETL tools for extracting targeted chat data from GCP Firestore for analysis
- Performed analyses to optimize call scheduling and diagnose pipeline failures for outbound use cases

Deepgram, Data Scientist

Oct 2020 - May 2024

- Cross-functional researcher collaborating with Engineering, DataOps, and Product teams to deliver cutting edge Speech AI products including Speech-To-Text (STT) and NLU/intelligence features
- Trained and evaluated dozens of STT models using PyTorch and shipped them to production, expanding our initial product offerings and rebasing our core models on next-gen architectures
- Developed novel end-to-end model architectures and processes for speech recognition:
 - **Patent:** Seagraves et al. End-to-end automatic speech recognition with transformer. U.S. Patent No. 12,380,880 B2. Patent Filed 2023-04-03, Granted 2025-08-05.
- Built automated model evaluation flows that stream metrics for real-time competitive benchmarking against SOTA models, allowing us to rapidly iterate to market-leading accuracy and performance
- Worked closely with Sales and Marketing on product announcements to boost customer acquisition and upsell opportunities, and provided analytics support to prevent customer churn
- Sourced and curated 10,000s of hours of high-value unstructured speech audio data for Research
- Mined production PostgreSQL databases and leveraged customer metadata from disparate sources to construct representative training/evaluation datasets and facilitate targeted product R&D
- Spearheaded efforts to improve STT punctuation by training experimental transformer-based language models with PyTorch and building necessary data pre-processing/text cleaning pipelines

- Helped build early NLU features (e.g. text summarization) and deploy as Dockerized Flask applications

Insight Data Science, Data Science Fellow

May 2020 - July 2020

- Created a ML-driven application which curates Spotify playlists by emotion/mood
- Derived crowdsourced emotional labels from Last.fm SQLite databases and combined audio features with song lyric sentiment (via NLP libraries) to build song emotion classifiers with > 71% accuracy
- Deployed the classification models on AWS EC2 in an interactive Streamlit web application

UC Davis Physics Department, Graduate Student Researcher and T.A.

Sept 2014 - August 2020

- Served as Data Processing Manager for an international, multimillion dollar dark matter particle search
- Led the development and operation of local R&D particle detection experiments from ideation to design, assembly, end-to-end data acquisition, and statistical analyses
- Developed C++/Python pipelines for signal processing to reduce TBs of unstructured, noisy waveform data into physical measurements and statistically model important nuclear processes
- Used calibration data to perform event reconstruction and develop position correction algorithms
- Produced Monte Carlo simulations using empirical models to predict particle detector response
- Used MySQL database replication and Flask to remotely monitor lab devices via web interface

EDUCATION

University of California, Davis

August 2020

Ph.D. in Experimental Particle Physics, Designated Emphasis in Nuclear Science

University of California, Davis

June 2014

B.S. in Particle Physics