

Jacob Cutter

📍 Bay Area, CA • ✉ jecutter92@gmail.com • ☎ (510) 861-0706 • 🌐 jecutter.github.io • 📺 jacob-cutter

Summary

Experimental physicist turned AI engineer/scientist, specializing in low-latency AI systems and the generative measurement layer – evaluation frameworks, metric design, and experimentation infrastructure – across speech-to-text (STT), voice agent, and content-generation products from prototype to GA.

Skills

Languages/Tools: Python, C++, UNIX/Shell, Git, Docker, Kubernetes, FastAPI, Flask, Streamlit, Splunk

ML/AI: Deep Learning (PyTorch), Transformers/LLMs, Prompt/Context Engineering, RAG/Retrieval Pipelines, Agentic Reasoning, Multi-Agent Orchestration, LlamaIndex, LangChain, Scikit-Learn

Speech/Audio: ASR/STT Modeling and Evaluation, Open-Source STT (Whisper), Conversational AI/Voice Agents, VAD & Signal Processing, Audio Featurization, Synthetic Audio Generation & Augmentation, torchaudio, librosa, pydub

Engineering: Low-Latency System Design, Distributed/Async Workflows (Temporal), Model Serving & Inference Optimization (vLLM, faster-whisper, Inferless), BDD/Integration Testing, Async Messaging (NATS Jetstream, Kafka)

Evaluation & Measurement: Metric Design, LLM-as-Judge & Human Alignment, Competitive Benchmarking (self-hosted models + closed-source APIs), Synthetic Data Generation, Experimental Design, Statistical Inference

Data: (Postgres, My)SQL, MongoDB, Redis, Vector Databases (Weaviate), Pandas, Visualization (Matplotlib, Seaborn)

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

Patent

US Patent 12,380,880 B2: Seagraves et al., *End-to-end automatic speech recognition with transformer*. Filed 2023-04-03, granted 2025-08-05.

Experience

GrowthX, Founding Staff Applied AI Scientist/Engineer – San Francisco, CA Mar 2026 – present

- First AI hire, reporting to the CTO – building the measurement, evaluation, and experimentation practice for a closed-loop platform that generates brand content optimized for AI Visibility (AEO/GEO) and search
- Prototyped new feedback mechanisms to incorporate and enforce customer-level writing specs in the generation pipeline; ran drafter loop experiments that yielded an 11.5-point quality improvement in fewer iterations
- Authored an AI Visibility study (~111,000 page-query pairs) for internal strategy and client-facing use, applying statistical controls to isolate which domain- and page-level features predict citation by generative search, across AI engines, industries, and funnel stages
- Built a Python experimentation playground and UI with injection points into production Temporal workflows, enabling generative debugging against real customer traces with modified prompts, models, and loop parameters
- Reducing costs across the generative stack – content generation workflows, page-quality audits, AI Visibility probes, and pipeline reliability improvements – with combined annual savings projected to be hundreds of thousands of dollars
- Audited evaluations across the platform and developed content quality CI test suites for gating change deployment across the core content generation path

Salesforce, Software Engineering LMTS (Lead Member of Technical Staff) – Palo Alto, CA Jan 2026 – Mar 2026

- Promoted from Senior Member of Technical Staff (SMTS) in recognition of contributions to Agentforce Voice

Salesforce, Software Engineering SMTS (Senior Member of Technical Staff) – Palo Alto, CA Sept 2024 – Jan 2026

- Tech Lead for Flash Planner, a low-latency LLM reasoning engine powering full-duplex conversational Voice Agents in Agentforce Voice; built and shipped it to GA on a compressed timeline
- Served as both technical coordinator and highest-contributing IC for a team of ~6 engineers, achieving sub-second time-to-first-audio (TTFA) through custom routing, task parallelization, connection pooling, and a from-scratch, Redis Cluster-backed state management layer preserving context across turns
- Built an end-to-end Voice Agent evaluation pipeline to settle the program's core architecture decision, by comparing real-time speech-to-speech models and cascade-style pipelines and justifying the latter's complexity tradeoff; the approach seeded the next generation of voice evaluation and extended into Trust & Safety guardrail coverage

- Implemented BDD-style integration tests encoding real-world conversation scenarios, wired into CI workflows to gate production deployments and block regressions in the voice pipeline before impacting customers
- Co-led a study benchmarking third-party ASR/STT providers against a self-hosted, optimized Whisper deployment (faster-whisper) on Lambda GPUs; demonstrated a viable streaming implementation with context injection that brought Whisper accuracy to parity or better on domain-specific audio at marginal latency cost
- Drove integration of Voice with Service Cloud, Data Cloud, and Agentforce Observability via cross-collaboration, and diagnosed production issues across the enterprise stack (including downstream Kafka event consumers) by leveraging Splunk and telemetry traces

Tenyx (acquired by Salesforce), Member of Technical Staff – Los Altos, CA May 2024 – Sept 2024

- Built a configurable synthetic customer and chat generation pipeline to evaluate AI voice agent quality, fine-tune and prompt-steer self-hosted open-source LLM classifiers (served via vLLM) to resolve recurring pipeline failures, and simulate failure modes to minimize customer impact before production releases
- Implemented an auto-evaluation system using LLMs to dynamically score models on generative tasks, including agent behavior, synthetic customer behavior, and slot filling/entity extraction tasks
- Created an audio augmentation pipeline to synthesize realistic phone call audio, enabling robustness evaluation of agent system components across configurable acoustic environments and noise conditions
- Developed data/ETL tools for extracting targeted chat data from GCP Firestore, and analyzed outbound call scheduling to diagnose pipeline failures

Deepgram, Data Scientist – Remote Oct 2020 – May 2024

- Led a Product Development team within Research, rebasing the full catalog of language and use-case models onto next-generation architectures and expanding product coverage across languages and NLU features
- Delivered a proof-of-concept ASR model for an extremely low-resource, primarily oral language with effectively no existing labeled data: iterated on transliteration schemes and multi-stage curricula to train different model variants from < 100 hours of labeled audio, securing continued partnership with a high-value enterprise prospect
- Trained and evaluated dozens of ASR/STT models in PyTorch and shipped them to production; sourced and curated 10,000s of hours of unstructured speech audio by cross-referencing production PostgreSQL databases and disparate customer metadata to build representative training and evaluation sets
- Spearheaded an ASR punctuation/capitalization improvement effort: defined in-house error/recall metrics (with ASR-level corrections to isolate formatting errors from recognition errors), curated ~13M stratified text chunks, and trained transformer-based seq2seq/token-classification models deployed in a Dockerized Flask microservice
- Built automated evaluation flows to stream metrics via Jetstream for real-time competitive benchmarking against SOTA models, applying bootstrapped CIs/hypothesis testing to establish whether new variants produced real gains
- Served as the central technical resource for public-facing benchmarks and pre-launch QA, partnering with Sales and Marketing on product launches and securing enterprise deals

Insight Data Science, Data Science Fellow – Bay Area, CA May 2020 – July 2020

- Built and deployed a Streamlit/AWS EC2 application curating Spotify playlists by mood, combining audio features with lyric sentiment against mined, crowdsourced Last.fm labels to reach >71% classification accuracy

UC Davis Physics Department, Graduate Student Researcher & Teaching Assistant – Davis, CA Sept 2014 – Aug 2020

- Served as Data Processing Manager for an international, multimillion-dollar dark matter particle search
- Led local R&D particle detection experiments end to end – ideation, design, assembly, data acquisition, and analysis
- Developed C++/Python signal processing pipelines reducing terabytes of noisy, unstructured waveform data into physical measurements, and used calibration data for event reconstruction and position correction algorithms
- Produced Monte Carlo simulations from empirical models to predict detector response and iterate against experiments

Education

PhD University of California, Davis, PhD in Experimental Particle Physics – Davis, CA Sept 2014 – Aug 2020

- Designated Emphasis in Nuclear Science

BS University of California, Davis, BS in Particle Physics – Davis, CA Sept 2010 – June 2014