

Steven Au

☎ (831) 332-5989 | ✉ sttau@ucsc.edu | 🔗 linkedin.com/in/steven-au | 📄 github.com/Prowo

SUMMARY

AI / ML Engineer focused on forward-deployed AI systems, NLP, multimodal RAG, and retrieval infrastructure. Experienced building client-facing prototypes, ETL pipelines, and evaluation workflows across AWS Bedrock and Databricks.

EXPERIENCE

Mount Sinai, Icahn School of Medicine

Volunteer ML Researcher, Bytes of Mind Lab

Sep. 2025 – Present

New York, NY / Remote

- Built **PeReGrINE**, a personalized LLM review-generation benchmark using temporally grounded user-item knowledge graphs, multimodal product evidence, and neighborhood retrieval for product intelligence.
- Designed a healthcare NLP knowledge graph linking patient medication reviews, adverse effects, and clinical-trial evidence, using information extraction and entity normalization for cross-source treatment comparison

Eternos Inc. / Uare.ai

Machine Learning Engineer, Forward-Deployed AI Systems

Apr. 2025 – Aug. 2025

Remote

- Created a client-specific knowledge base from 250+ long-form interviews, handling diarization, transcript cleaning, gold-label testing, and retrieval evaluation for digital twin workflows.
- Engineered automated ETL and orchestration pipelines in Python across AWS S3 and Databricks Lakehouse environments, reducing manual processing time by **40%**.
- Designed multimodal RAG pipelines on AWS Bedrock for text, audio, and video retrieval, improving retrieval relevance by **25%** through hybrid embedding optimization, prompt tuning, and RAGAS evaluation.

University of California, Santa Cruz

Web Development & Digital Communications Specialist

Jun. 2024 – Jan. 2025

Santa Cruz, CA

- Led front-end development for CruzChat, an AI tutoring application built with React and Next.js, integrating assistant-based RAG workflows and improving student engagement by **20%** for an engineering course.

Intel Labs

Machine Learning Engineer Intern

May 2024 – Sep. 2024

Santa Clara, CA

- First-authored research on personalized product review generation using query-based retrieval over user-item graphs, ratings, and top- k neighbor evidence for sparse user profiles.
- Tested BM25, dense retrieval, user-only, neighbor-only, and graph-based retrieval pipelines, improving ROUGE by **12%** over prior baselines.

PUBLICATIONS

Blind Deference: A Diagnostic Benchmark for Authority-Driven Sycophancy in Multi-Turn LLM Conversations

Noronha, S., Au, S.

COLM, 2026 (in acceptance)

Personalized Graph-Based Retrieval for LLMs

Au, S., et al.

PACLIC 39, 2025

EDUCATION

University of California, Santa Cruz

Master of Science in Natural Language Processing

Sep. 2023 – Dec. 2024

Santa Cruz, CA

University of California, Santa Cruz

Bachelor of Science in Computer Science

Sep. 2015 – Mar. 2023

Santa Cruz, CA

SKILLS

Languages: Python, SQL, Java, C/C++, R, MATLAB, LaTeX

ML / NLP: PyTorch, TensorFlow, scikit-learn, Pandas, spaCy, NLTK, LangChain, LLMs, RAG, RAGAS

Platforms / Tools: AWS, AWS Bedrock, S3, Databricks, Docker, Git, Linux, Jupyter, Neo4j, Tableau