Steven Au

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

SUMMARY

Machine Learning Engineer specializing in NLP and multimodal retrieval infrastructure with experience in Python and ML frameworks such as PyTorch, TensorFlow, and scikit-learn. Built client-facing prototypes and automated ETL pipelines on AWS Bedrock and Databricks, reducing manual processing by 40%. Developed RAG pipelines optimizing retrieval relevance and evaluation workflows.

EXPERIENCE

Mount Sinai, Icahn School of Medicine **Sep 2025 - Present**
Volunteer ML Researcher, Bytes of Mind Lab *New York, NY / Remote*
- Built PeReGrINE, a machine learning benchmark for personalized review generation using temporally grounded user-item graphs with multimodal and neighborhood evidence.
- Designed a medication knowledge graph linking patient reviews, side effects, and clinical-trial evidence for cross-source comparison of treatment experiences.

Eternos Inc. / Uare.ai **Apr 2025 - Aug 2025**
Machine Learning Engineer, Forward-Deployed AI Systems *Remote*
- Created a client-specific knowledge base from 250+ long-form interviews, handling diarization, transcript cleaning, gold-label testing, and classification 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% and supporting model evaluation workflows.
- Designed multimodal RAG pipelines on AWS Bedrock for ext, audio, and video retrieval, improving retrieval relevance by 25% through hybrid embedding optimization, prompt tuning, and RAGAS evaluation.

University of California, Santa Cruz **Jun 2024 - Jan 2025**
Web Development & Digital Communications Specialist *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 **May 2024 - Sep 2024**
Machine Learning Engineer Intern *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

- Au, S..MIDI-PHOR: Multi-View Extraction and Captioning for Music Understanding.*NLP4MusA at EACL, 2026*
- Au, S., et al..Personalized Graph-Based Retrieval for LLMs.*PACLIC 39, 2025*
- Wan, N., Au, S., et al..Emotion Discovery and Reasoning its Emotion Flips in Conversation.*SemEval at ACL, 2024*

EDUCATION

University of California, Santa Cruz **Sep 2023 - Dec 2024**
Master of Science, Natural Language Processing *Santa Cruz, CA*

University of California, Santa Cruz **Sep 2015 - Mar 2023**
Bachelor of Science, Computer Science *Santa Cruz, CA*

SKILLS

- **Languages**: Python, SQL, Java, C/C++, R, MATLAB, LaTeX
- **ML / NLP**: Machine Learning, PyTorch, TensorFlow, scikit-learn, Pandas, spaCy, NLTK, LangChain, LLMs, RAG, RAGAS, Classification, Model Evaluation
- **Platforms / Tools**: AWS, AWS Bedrock, S3, Databricks, Docker, Git, Linux, Jupyter, Neo4j, Tableau