

Tiantong Lu

Berkeley, CA 661-625-8365 lutt@berkeley.edu tiantonglu.pages.dev GitHub LinkedIn

EDUCATION

University of California, Berkeley

B.A. in Data Science; Minor in Computer Science

GPA: 3.83/4.00

Expected May 2027

Domain Emphasis: Applied Mathematics and Modeling

Data Science Honors Program (2026–2027)

Selected courses: Machine Learning (CS 189), Data Engineering (Data C101), Data Mining (Data 144), Probability (Stat 134), Artificial Intelligence (CS 188), Computer Networks (CS 168), Optimization (EECS 127), Numerical Analysis (Math 128A).

College of the Canyons, Santa Clarita (Valedictorian)

GPA: 4.00/4.00

A.S. degrees in Computer Science, Mathematics, and Physics; A.A. in Economics

Aug 2023 – May 2025

RESEARCH EXPERIENCE

TokenWorks LLC — Cuneiform Recognition and Tablet Transcription

Jun 2026 – Present

Machine Learning and Computer Vision Researcher

Mentor: Dr. Adam Anderson

- Designed compositional pre-training from sign crops and text (387,315 pairs); at 100 adaptation tablets, cut fixed-vocabulary, oracle-localized line sign error by 24–52 percentage points versus omitting this stage in CTC/VLMs.
- Engineered modular tablet transcription with detection, recognition, language decoding and CTC verification; achieved 38.05% mean whole-side sign error versus a 68.92% geometric baseline on 198 reused in-period sides (exploratory).

Ziad Obermeyer Lab, UC Berkeley School of Public Health

Sep 2026 – Present

Student Researcher — Clinical Machine Learning

Mentors: Prof. Ziad Obermeyer; Dr. Aamir Javaid

- Scaled cardiac-risk model inference to 800,035 MIMIC-IV ECG records on Savio; linked predictions to longitudinal clinical data and evaluated one-year ventricular tachycardia/fibrillation outcomes.
- Designed longitudinal ECG–echo analyses of newly documented ejection fraction of 35% or less over 1–12 months; accounted for prior dysfunction and follow-up censoring, treating death as a competing risk.
- Developed local phenotyping of discharge summaries with two open-weight LLMs to distinguish sustained VT/VF from nonsustained or historical diagnoses; tested how outcome definitions change cardiac-risk model discrimination.

UC Berkeley DNA Sequencing Facility

Sep 2026 – Present

Researcher — Sanger Trace Analysis and Basecaller Evaluation

Mentor: Dr. Scott Geller

- Independently lead Sanger reprocessing research; built Python tools to inspect AB1 signal channels, peaks, basecalls and quality scores, and score reference-based sequence errors, ambiguity and called length.
- Compiled and tested the open-source TraceTuner basecaller, and designed matched-input, reference-based comparisons with PeakTrace to assess basecalling accuracy, trimming and reliable read length.

TokenWorks LLC — Multilingual Book OCR

Jan 2026 – Jun 2026

Research Intern (UC Berkeley Data Discovery)

Mentor: Dr. Adam Anderson

- Led evaluation of OCR pipelines for noisy, multilingual book scans; compared character error rates, reading order, diacritic preservation and throughput to select recognition systems and guide preprocessing and text-extraction workflows.

UC Berkeley Haas School of Business — AI and Labor Markets

Jan 2026 – May 2026

Data Science Researcher

Mentor: Debbie Liang

- Automated RentAHuman.ai monitoring with six-hourly API ingestion, PostgreSQL storage and deduplication.
- Applied positive–unlabeled learning to study AI-generated postings; analyzed platform visibility and task turnover.

MANUSCRIPTS

From Signs to Sequences: Compositional Synthesis for Cuneiform Recognition Across Periods and Collections.

Co-first author. ICLR 2027 (submitted; awaiting peer review). Token-matched single-/multi-sign CTC experiments.

TabletVLM: End-to-End Cuneiform Tablet Transcription with a Spatially Grounded Vision Language Model. **Co-first author.**

ICLR 2027 (submitted; awaiting peer review). Contributed synthetic-line supervision and comparative model evaluation.

Beyond String Agreement: Source Consensus, Sign Identity, and Provenance in Cuneiform Unicode Conversion.

Co-first author. Manuscript complete (not submitted). Evaluated six resources on 565,542 reference-encoding tokens.

TEACHING EXPERIENCE

Math and Science Tutor, The Learning Center, College of the Canyons

Jun 2024 – Jun 2025

Led one-on-one tutoring and small-group exam workshops; curated practice problems in advanced math and science.

TECHNICAL SKILLS

Languages: Python, SQL, Java, C/C++, MATLAB, Bash, JavaScript, HTML/CSS.

Libraries: PyTorch, Hugging Face Transformers, LightGBM, PaddleOCR, scikit-learn, NumPy, pandas, SciPy, OpenCV.

Methods: LoRA fine-tuning, mixed-precision training, CTC decoding, bootstrap resampling, competing-risk analysis.

Platforms & tools: Linux, Slurm (Savio), CUDA, PostgreSQL, Docker, Git, Node.js, Weights & Biases, LaTeX.