Sonnet Xu

248-834-3142 | sonnet@stanford.edu | linkedin.com/in/sonnetxu | github.com/sonnetx

EDUCATION

Stanford University, BS, MS 2023 - 2027 — Stanford, CA
Computer Science (AI Track) and Math — *GPA: 4.0/4.0*
Relevant Coursework: Cryptography, Deep Reinforcement Learning, Computer Vision, Artificial Intelligence, Blockchains
Research Interests: Representation learning, multimodal foundation models (VLMs), interpretability, cryptography/privacy.

PROFESSIONAL EXPERIENCE

Incoming Machine Learning Engineering Intern | *DatologyAI* — 01/2026-03/2026

LLM Guardrails SDE Intern | *Amazon — Rufus (LLM Shopping Assistant)* — 06/2025 – 09/2025
- Unified safety-rule ingestion pipeline (AWS Lambda, DynamoDB, S3); built unit/integration tests with 90%+ coverage.
- Integrated classifier-based and rule-based LLM guardrails for Rufus, spanning model-output filtering and input handling.

Zero Knowledge Research Engineering | *Nethermind* — 06/2025 – 09/2025
- Refactored cryptographic aggregation logic for post-quantum signature schemes; redesigned type-safe interfaces.

AI/LLM Automation Intern | *Harvest Ventures* — 03/2025 – 08/2025
- Built agent-based LLM tools for founder/source discovery; implemented data-collection pipelines across niche verticals.

Technology Policy Student Research Fellow | *Hoover Institution* — 09/2024 – 06/2025
ML systems research applied to large-scale policy and patent datasets.
- Developed RAG system for analyzing policy documents using ChromaDB and LangChain; evaluated retrieval w/ cosine similarity.
- Migrated patent analysis to HPC clusters, enabling inference for 1.5M+ documents, improving throughput by >20×.
- Analyzed U.S.–China technology domains using learned vector similarity metrics and embedding-based clustering.

Research Assistant | Daneshjou Lab at *Stanford Medicine* — 10/2023 - Present
Working on interpretability and evaluation of vision-language foundation models for fairness.
- Investigating concept activation vectors (CAVs) and automated concept discovery for VLM interpretability; developed pipelines for generating visual concepts across multiple architectures (CLIP, SigLIP, LLaVA, Flamingo).
- Studied information degradation in compressed clinical imagery, benchmarking ViT, DINOv2, SimCLR, and VLM encoders under JPEG latency constraints; performed finetuning + linear probing experiments.
- Led experimental workflows for two published papers (MICCAI 2025, NeurIPS 2024), including dataset curation, embedding generation, evaluation metrics, and statistical analysis.
- Built reproducible PyTorch codebases for multimodal interpretability evaluation and scalable embedding extraction.

SELECTED PUBLICATIONS

Xu S, Janizek J, Jiang Y, Daneshjou R. BiasICL: In-Context Learning and Demographic Biases of Vision-Language Models. MICCAI 2025.
Xu S, Gui H, Rotemberg V, Wang T, Chen YT, Daneshjou R. Evaluating the Efficacy of Foundation Embedding Models in Healthcare. medRxiv 2024.
Mello MM, Char D, **Xu SH**. Ethical Obligations to Inform Patients About AI Tool Use. JAMA 2025.
Sagers LW, Shah AP, **Xu S**, Daneshjou R, Manrai AK. Directing Generalist VLMs to Interpret Medical Images Across Populations. NeurIPS GenAI for Health Workshop 2024.

SELECTED PROJECTS

Representation Learning & Vision (CS231N)
- Fine-tuned **ViT, DINOv2, SimCLR** under varying image compression constraints for dermatology generalization.
- Conducted ablations on representation collapse, linear separability, and robustness across training regimes.

Reasoning-Focused VLMs for Biochemistry (CS224R)
- Developed a dataset and trained SFT/GRPO-based VLMs for biochemistry reasoning tasks (protein-ligand binding).

CAV-Based Interpretability Toolkit (CS221)
- Implemented pipelines for computing CAVs on open-source VLMs; automated concept extraction and sensitivity scoring.

TreeTrash (TreeHacks 2025 – OpenAI Winner)
- Built visual-retrieval augmented generation using ColPali embeddings + Vespa AI vector search

Generative AI for Urban Design — Google Summer of Code, City of Boston
- Built collaborative generative-design web app using a diffusion-based backend; integrated constraints into prompt conditioning.

SKILLS

PyTorch, HuggingFace, DSPy, HPC cluster workflows, AWS (Lambda, S3, DynamoDB), Linux, Python, C/C++.