

YISI (AMY) LYU

805-618-0358 | amylv777@hotmail.com | Stanford, CA | www.linkedin.com/in/yisilyu0323

EDUCATION

Stanford University, *MS in Data Science, Statistics*

Stanford, CA (June 2026)

University of California, Santa Barbara, *BS in Statistics and Data Science*

Santa Barbara, CA (June 2024)

TECHNICAL SKILLS

- **Programming:** Python (pandas, NumPy, scikit-learn, TensorFlow, PyTorch), R, SQL, Bash, HTML, LaTeX, SAS
- **ML/DL:** Regression (Linear, Ridge, Lasso), Classification (Logistic, RF, XGBoost), Time Series (ARIMA, LSTM), Computer Vision (CNN, YOLOv4), OCR
- **Generative AI:** Transformers, LangChain, fine-tuning (LoRA/ReFT), embeddings, vector search, RAG, ChromaDB/FAISS
- **Data Engineering & Modeling:** ETL, data modeling, schema validation, PostgreSQL, Azure SQL Database
- **Data Analysis:** Bayesian Analysis, Statistical Inference, Hypothesis Testing (T-test, ANOVA, Chi-square tests, (non)parametric tests, Z-test), descriptive statistics, correlation, security/privacy compliance (HIPAA, SOC 2)
- **Tools:** Power BI, Tableau, Excel, Jupyter Notebook, GitHub, Docker; Azure ML, Azure Fabric/Data Warehouse, GCP Batch/dsub, Snowflake
- **Finance:** Monte Carlo simulation, portfolio optimization, A/B testing, online experiments, market/campaign/pricing analysis

EXPERIENCE

Penumbra, Alameda, CA - *Data Science Engineering Intern*

07/2025-09/2025

- Developed a 95% field-level accurate OCR + Azure AI pipeline for a 600GB multi-template IQC/LHR PDF and spreadsheet backlog, extracting supplier, part, and inspection data while reducing manual transcription by ~80% and errors by 76%
- Designed Azure Fabric PostgreSQL/pgVector schemas to integrate structured master data into a low-latency RAG retrieval layer, enabling ~150ms retrieval for supplier, part, and inspection records
- Built Power BI dashboards tracking inspection throughput, rejection codes, and vendor performance, accelerating root-cause analysis and cutting turnaround from around 3 days to less than 1 day
- Automated QMS-compliant Python validation/schema checks across 30K+ values, 305 fields, and 10+ templates
- Designed grounded RAG prompt templates and retrieval constraints, cutting hallucinations from 28.4% and raising answer precision to 90%+ for audit-ready outputs

Data Application Lab, Los Angeles, CA - *Data Analyst Intern*

07/2023-10/2024

- Architected an end-to-end medical imaging ETL pipeline using Python, TensorFlow, and OpenCV to preprocess and augment 38K+ raw MRI, X-ray, and CT images into expanded training samples
- Fine-tuned vision models to 92% accuracy on edge-case classification, including fractures and tumors; coordinated delivery across a 6-person team
- Developed a BERT/GPT radiology report-generation system mapping symptoms and imaging findings to diagnostic labels, cutting report time 30.2% and saving 15 minutes/patient
- Designed a smart-money trading strategy using logistic regression, time-series analysis, and crypto news ranking with scikit-learn and PyTorch, improving decision accuracy by 30.5% and modeled annual returns by \$1.2M

TESTIFI, Munich, Germany - *Data Scientist Intern*

07/2022-10/2022

- Increased testing efficiency by 20.4% and saved €150K annually by building predictive algorithms for auto-APL test generators using Python, logistic regression, and decision trees
- Developed a Flask and Dash web app to visualize test case performance and implemented a YOLOv4 model to automate UI test element identification, reducing manual workload by 40.3% and saving 25 developer hours weekly
- Centralized multi-source QA data into ETL, optimized SQL, and presented testing insights to improve decision quality/speed

DATA SCIENCE RESEARCH AND PROJECTS

Reference-Guided Artistic Style Transfer with Stable Diffusion + IP-Adapter

03/2026-06/2026

- Built a Stable Diffusion v1.5 + IP-Adapter style-transfer pipeline using WikiArt/COCO pairs, evaluating content preservation and style alignment with LPIPS and CLIP similarity across 1,000 WikiArt/COCO evaluation pairs
- Implemented training-free semantic attention and DDIM latent energy guidance with frozen VGG/CLIP gradients; improved LPIPS content preservation from 0.4517 to 0.2573 and CLIP style similarity by 5.5%

Robustness of AIPW and Causal Forest Estimators to Propensity Score Misspecification (IHDP)

09/2025-12/2025

- Built a semi-synthetic IHDP benchmark (747 obs., 25 covariates) with known ATE/CATE, varying nonlinearity and overlap
- Benchmarked AIPW vs. CausalForestDML (50×/setting) on ATE bias/RMSE/coverage + CATE recovery; added “unsafe” flags (ESS/n<0.2, >5% $\hat{p}$ outside [0.05,0.95], max weighted |SMD|>0.2) to guide estimator choice

Curriculum + Reward-Guided RL Fine-Tuning for Aligned LLMs

03/2025-06/2025

- Fine-tuned Qwen-2.5 0.5B with SFT, curriculum, and DPO, adding Nemotron-70B reward-guided reranking; raised win rate 50% to 95%, with curriculum cutting verbosity and reranking boosting precision for reliable instruction following
- Engineered a production-ready 1-GPU fine-tuning pipeline with mixed precision, gradient checkpointing, and reproducible Nemotron evaluation; led inference/curriculum modules to stabilize training and deliver deployable alignment gains