YUHENG TU

UCLA, Los Angeles, California, United States, 90025
yuhengtuece@gmail.com | yuhengtu@ucla.edu
Website | Scholar | Github | LinkedIn

Education

University of California, Los Angeles (UCLA) *M.S. ’27 in Computer Science*

University of California, Berkeley (UCB) *24-Spring Visiting Student in Computer Science* **GPA**: 3.9/4.0

Southeast University (SEU) *B.Eng. ’25 in Electrical & Computer Engineering* **GPA**: 3.81/4.0

Publications

[1] Sang Truong*, **Yuheng Tu***, Rylan Schaeffer, Sanmi Koyejo. “Item Response Scaling Laws: A Measurement Theory Approach to Generalizable Neural Performance Prediction.” *Under Review*.

[2] Sang Truong*, **Yuheng Tu***, Michael Hardy*, Anka Reuel, Zeyu Tang, Jirayu Burapacheep, Jonathan Perera, Chibuike Uwakwe, Benjamin W. Domingue, Nick Haber, Sanmi Koyejo. “Fantastic Bugs and Where to Find Them in AI Benchmarks.” *NeurIPS ’25 D&B*.

[3] Sang Truong, **Yuheng Tu**, Percy Liang, Bo Li, Sanmi Koyejo. “Reliable and Efficient Amortized Model-based Evaluation.” *ICML ’25*.

[4] Yi Zeng*, Yu Yang*, Andy Zhou*, Jeffrey Ziwei Tan*, **Yuheng Tu***, Yifan Mai*, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, Bo Li. “AIR-Bench ’24: A Safety Benchmark Based on Risk Categories from Regulations and Policies.” *ICLR ’25 Spotlight*.

[5] Guojun Chen, Kaixuan Xie, **Yuheng Tu**, Tiecheng Song, Yinfei Xu, Jing Hu, and Lun Xin. “NQFL: Nonuniform Quantization for Communication Efficient Federated Learning.” *IEEE Communications Letters (COMML)*.

Featured Research Experience

Item Response Scaling Laws: A Measurement Theory Approach to Generalizable Neural Performance Prediction **Remote**
Research Assitant, Stanford Trustworthy AI Research (STAIR), Supervisor: Prof. Sanmi Koyejo Mar 2025 - Present

- Derive interpretable and generalizable scaling laws with Item Response Theory (IRT)
- Extend IRT with a Beta loss to model AI-specific empirical probability responses
- Study pre-training downstream scaling on 25 models with up to 359 checkpoints across 15 datasets
- Study test-time scaling on 15 models across 10 datasets with up to 10,000 samples

Fantastic Bugs and Where to Find Them in AI Benchmarks **Remote**
Research Assitant, Stanford Trustworthy AI Research (STAIR), Supervisor: Prof. Sanmi Koyejo Mar 2025 - Present

- Propose a scalable, theory-driven framework for systematic AI benchmark revision using psychometric tools
- Revise 9 AI benchmarks, achieving up to 84% precision in detecting flawed questions

Reliable and Efficient Amortized Model-based Evaluation **Palo Alto, CA**
Research Assitant, Stanford Trustworthy AI Research (STAIR), Supervisor: Prof. Sanmi Koyejo Jul 2024 - Mar 2025

- Evaluate 183 LLMs across 22 datasets reliably and efficiently with Item Response Theory (IRT)
- Integrate Computerized Adaptive Testing (CAT) into `stanford-crfm/helm` ★2.5k
- Propose amortized calibration to predict question difficulty from embedding
- Fine-tune Llama-3-8B to generate question conditioned on difficulty using SFT and PPO

AIR-BENCH 2024: A Safety Benchmark Based on Risk Categories from Regulations and Policies **San Francisco, CA**
Research Assitant, Secure Learning Lab (SL^2), Supervisor: Prof. Bo Li May 2024 - Jul 2024

- Curate 5,694 detailed and diverse instruction prompts across 314 risk categories and 3 language styles
- Evaluate 22 leading LLMs with GPT-4o as a judge and category-specific system prompts

Competition & Service

- Rank 2nd at the UC Berkeley’s CS189 HW6 Kaggle Competition on CIFAR-10 image classification with CNN
- Serve as a reviewer for ICLR ’26 and multiple previous workshops