YUHENG TU

UCLA, Los Angeles, California, United States, 90025
+1 (510) 809-5148 | yuhengtu@ucla.edu
Website | Scholar | Github | LinkedIn

EDUCATION

University of California, Los Angeles (UCLA) *M.S. '27 in Computer Science* **GPA**: 3.77/4.0

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

FEATURED ACADEMIA OR INDUSTRY EXPERIENCE

TikTok Inc. Trust and Safe Foundation Model Group — San Jose, CA
Research Scientist Intern, Advisor: Dr. Yi Zeng — Jun 26 - Sep 26

[Preprint] Pan-Safety Core: Multidimensional and Reliable Evaluation of Language Model Safety
Yuheng Tu, *Weiyu Sun, Zhengyue Zhao, Xin Liu, Andrew Estornell, Chaowei Xiao, Weiguo Feng, Dylan Hadfield Menell, Yi Zeng*

- Build a unified evaluation framework spanning chatbot/agent safety, guardrails, and safety reward modeling
- Select high-quality data sources, then filter flawed questions with detection algorithms and LM semantic checks
- Design deployment-matched evaluation procedure and validate LM judges against human agreement
- Evaluate 63 LMs, produce leaderboards with statistical uncertainty and multiple-choice bias analysis

Stanford University Stanford Trustworthy AI Research (STAIR) — Palo Alto, CA
Research Assitant, Advisor: Prof. Sanmi Koyejo — Aug 24 - Jun 26

[ICML '26] Item Response Scaling Laws: A Measurement Theory Approach for Efficient and Generalizable Neural Scaling Estimation
Sang Truong, **Yuheng Tu***, Rylan Schaeffer, Sanmi Koyejo*

- Derive question difficulty-aware scaling laws that generalize across benchmarks with varying difficulties
- Extend Item Response Theory to model continuous LM token probabilities and empirical pass@1 rate
- Study pre-training downstream scaling on 6,612 LM checkpoints and 37,682 questions from 10 benchmarks
- Study test-time scaling on 12 LMs and 120 questions from 4 benchmarks with 2,500 samples per question

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

- Propose scalable, measurement theory-driven algorithms for systematic AI benchmark revision
- Revise 9 AI benchmarks, detecting flawed questions via response patterns with up to 84% precision

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

- Evaluate 183 LMs across 22 benchmarks reliably and efficiently with Item Response Theory
- Integrate Adaptive Testing into `stanford-crfm/helm` ★2.5k, cutting 50% evaluation queries
- Propose amortized calibration to predict question difficulty from text embedding
- Fine-tune Llama-3-8B to generate questions conditioned on difficulty using SFT and PPO

University of Illinois Urbana-Champaign Secure Learning Lab (SL^2) — San Francisco, CA
Research Assitant, Advisor: Prof. Bo Li — May 24 - Jul 24

[ICLR '25 Spotlight] AIR-BENCH 2024: A Safety Benchmark Based on Risk Categories from Regulations and Policies
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*

- Curate 5,694 prompts across 314 risk categories derived from 9 company policies and 3 regional regulations
- Evaluate 22 LMs with GPT-4o as judge and category-specific judge prompts, achieving high human agreement

COMPETITION & SERVICES

- Rank 2nd at the UC Berkeley's CS189 HW6 Kaggle Competition on CIFAR-10 image classification with CNN
- Serve as a reviewer for ICLR, ICML, NeurIPS; Serve as a reader for CS163 and CS260 at UCLA
- Lead SEU Flybook 2025, a publicly available guide for SEU students applying to MS/PhD programs abroad