

# YUHENG TU

Southeast University, Nanjing, Jiangsu, China, 211189

yuhengtuece@gmail.com

[Google Scholar](#) | [Github](#)

## EDUCATION

### Southeast University (SEU)

Senior, Pursuing B.Eng. in Information Science and Engineering

GPA: 3.81/4.0

Nanjing, China

Sep 2021 - Jun 2025

### University of California, Berkeley (UCB)

Visiting Student, Computer Science

GPA: 3.9/4.0

Berkeley, CA

Jan 2024 - Aug 2024

## PUBLICATIONS

[1] Sang Truong, **Yuheng Tu**, Percy Liang, Bo Li, Sanmi Koyejo. “Reliable and Efficient Amortized Model-based Evaluation.” *Under review at ICLR 2025*.

[2] 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 2024: A Safety Benchmark Based on Risk Categories from Regulations and Policies.” *Under review at ICLR 2025*.

[3] 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)*.

[4] **Yuheng Tu**, Jianan Liu, Tian Qiu, Yunlang Cai, Jianan Zhang, Jianwei You, and Tiejun Cui. “Fast Design of Metasurface-Based Microwave Absorber Using the Neuro-TF Approach.” *Photonics and Electromagnetics Research Symposium (PIERS)*, 2023.

## RESEARCH EXPERIENCE

### Reliable and Efficient Amortized Model-based Evaluation

Palo Alto, CA

Research Assistant, Stanford Trustworthy AI Research (STAIR), Supervisor: Prof. Sanmi Koyejo

Jul 2024 - Oct 2024

- Evaluate 161 LLMs across 24 datasets reliably and efficiently with Item Response Theory (IRT)
- Introduce amortized calibration to reduce the cost with minimal sacrifice of accuracy
- Fine-Tune Llama-3-8B to generate questions conditioned on item parameters

### Safety Benchmark Based on Risk Categories from Regulations and Policies

San Francisco, CA

Research Assistant, Virtue AI, Supervisor: Prof. Bo Li

May 2024 - Jul 2024

- Generate 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

## PROJECTS

### Federated Learning Algorithms pursuing Gradient Compression

Nanjing, China

Project Leader, Provincial-level Undergrad Research Project, Supervisor: Prof. Yinfei Xu

May 2023 - Dec 2023

- Develop the NQFL algorithm which normalizes the gradients and quantizes them with the Lloyd-Max quantizer
- Implement NQFL along with 3 comparative algorithms: QSGD, AdaQuantFL, and SLMQ in FedML framework

### Developing ML Algorithms to Accelerate Microwave Simulation

Nanjing, China

Project Leader, National-level Undergrad Research Project, Supervisor: Prof. Jianan Zhang

Sep 2022 - May 2023

- Extract poles and residues from absorbance curves with the Vector-Fitting algorithm
- Develop the Neuro-TF model: the MLP is used to derive poles and residues from geometric parameters, and the transfer function is used to derive absorbance from poles and residues

## COMPETITIONS

- Rank 2<sup>nd</sup> at the UC Berkeley’s CS189 HW6 Kaggle Competition on CIFAR-10 image classification with CNN