My undergraduate journey in Information Science and Engineering has enabled me to contribute to diverse, impactful research particularly in the last two years, spanning the paradigm of Machine Learning (ML) with a recent spotlight on **the evaluation of Large Language Models (LLMs)**. Driven by the principle that the lack of measurability hinders progression, my work has largely focused on a **Comprehensive Safety Benchmark** and **Reliable & Efficient Evaluation Algorithm**. They are now the leverage I lean on to build more powerful, responsible AI systems.

Early Research Experiences At my home university, I had the avenues to build a foundation in the research lifecycle through two holistic projects - a microwave simulation acceleration project with ML, where I was mainly responsible for developing the Neuro-TF model which combines neural networks and pole-residue-based transfer function (Yuheng et al., 2023), and a Federated Learning (FL) project, where I did all the empirical implementation for the NQFL algorithm, which quantize the gradients with Lloyd-Max quantizer to reduce communication costs (Guojun et al., 2023). While these established a strong affinity for research, my pursuit of the same kicked into high gear when I discovered the space of LLM evaluation.

Comprehensive Safety Benchmark At UC Berkeley, I was an audience to Prof. Bo Li, at an AI safety seminar where she discussed the rapid growth of LLMs that present an intrinsic contrast. She covered how the benefits of LLM are at direct loggerheads with significant risks such as generating toxic content and spreading misinformation and that governments and companies have developed comprehensive regulations and policies to address this. However, the popular benchmarks used to evaluate LLM safety were not current and were based on former literature, intuition, and/or common sense, which had not kept up with the times. Inspired, I joined her research group at Virtue AI, where we built AIR-Bench 2024 – the first AI safety benchmark designed to align with emerging government regulations and company policies. Equipped with 5,694 prompts spanning 314 categories with context from 8 government regulations and 16 company policies, we evaluated a minimum of 22 of the leading LLMs today. I had the opportunity to be a primary contributor to this initiative, developing a quarter of all benchmark prompts, as well as the open-source repository to practically evaluate LLM safety per the benchmark. The detailing of this work is currently under review at **ICLR'25** as a manuscript where I am a first co-author (Yi et al., 2024). I now have a street-level view into the safety challenges of real-world models, which has added powerful fuel to my interest in developing novel evaluation methods for broader real-world impact.

Reliable & Efficient Evaluation Algorithm My time at Virtue AI was also a holistic learning opportunity regarding challenges in LLM evaluation. To reduce the computational cost, a common practice is to use the average score on randomly selected small subsets of large benchmarking datasets as an LLM performance measure. A typical subset consists of just 500 prompts, as opposed to a full dataset which ranges anywhere from 1,000 to 200,000 prompts. While this approach does reduce computation, it comes at the cost of variations in evaluation outcomes across different subsets beyond correction with seeds, which creates inconsistency in performance assessments. This gap was my prime motivator to work with Prof. Sanmi Koyejo at STAIR Stanford, where our custom application of Item Response Theory (IRT) improved reliability while maintaining the efficiency of subset evaluation, as demonstrated across 184 LLMs and 25 datasets. A hallmark of this approach was our introduction of amortized calibration, and the subsequent fine-tuning of an LLM to generate questions conditioned on a desired difficulty level for the first time. This work is also under review at **ICLR'25** (Sang et al., 2024), where I am one of the two student authors of the paper. My specific contributions included the empirical implementation of algorithms and large-scale experimental validation. I do believe that the continued presence of novel approaches such as these is pertinent to pushing the boundaries of evaluation methods and paving the way for continued progression in the same.

Going forward The experiential learning from these experiences has motivated me to explore the leverage in current evaluation strategies to build more powerful and responsible AI systems. The reliable and efficient evaluation of these methods across fine-grained categories still begs a question – Given the models' abilities on different task categories, can we classify a new prompt into a specific category, infer its difficulty parameter, and then assign it to an appropriate model based on the computational budget and desired answer quality? Recent research where task-specific small models are outperforming large generic models (Yilun et al., 2024) is a backdrop to this question. Along these lines, I am excited to explore LLM evaluation further to raise the robustness of the evaluation space. Additionally, I am deeply passionate about the intersection of trustworthy AI, cryptography, and security. I have extensively explored both classical AI safety topics, such as adversarial samples and backdoors, as well as emerging issues in LLM safety, including jailbreaks and watermarks. My attempts to implement jailbreak algorithms like GCG (Andy et al., 2024) and Advprompter (Anselm et al., 2024) on audio models have further enriched my understanding of these challenges. Among my academic experiences, the cryptography course I took at UC Berkeley stands out as especially inspiring, igniting my passion for both the theoretical foundations and practical applications of secure systems. More broadly saying, my research philosophy is to draw insights from established theories to advance AI development, while simultaneously revitalizing traditional methodologies within the context of modern AI.

My immediate goal is to deepen my understanding of computer science, with a particular focus on theoretical foundations, through structured learning. In the medium term, I aim to contribute to impactful research, with the possibility of pursuing a PhD to further expand my expertise. Ultimately, my long-term aspiration is to become a research scientist in industry, driving innovation and addressing complex challenges.

References

[1] Sang Truong, **Yuheng Tu**, Percy Liang, Bo Li, Sanmi Koyejo. Reliable and Efficient Amortized Model-based Evaluation. *International Conference on Learning Representations (ICLR; under review)*, 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. *International Conference on Learning Representations (ICLR; under review)*, 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)*, 2023.

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

[5] Du Yilun, and Leslie Kaelbling. Compositional generative modeling: A single model is not all you need. *arXiv preprint arXiv:2402.01103*, 2024.

[6] Zou Andy, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

[7] Paulus Anselm, Arman Zharmagambetov, Chuan Guo, Brandon Amos, and Yuandong Tian. Advprompter: Fast adaptive adversarial prompting for llms.. *arXiv preprint arXiv:2404.16873*, 2024.