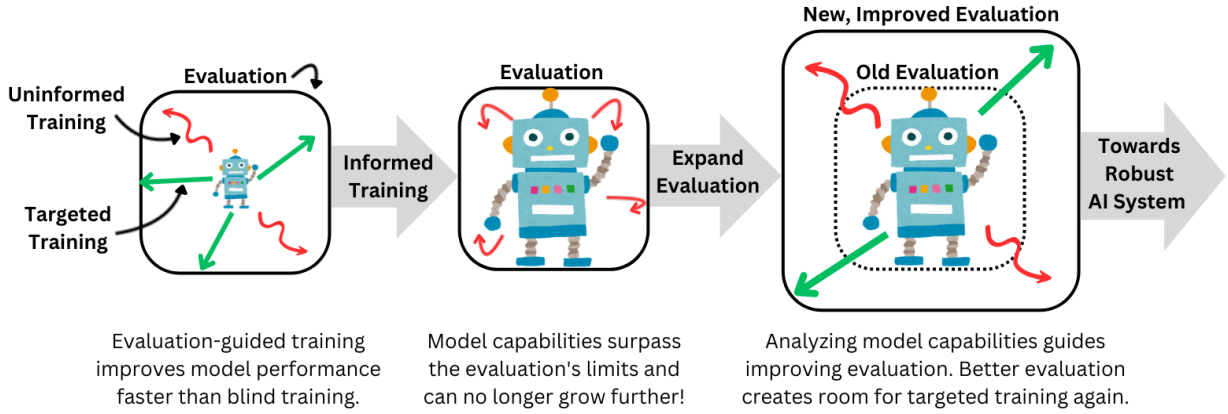
The future of AI hinges on the development of robust and trustworthy models, much like the societal adaptation of automobiles depended on the reliability of their engines. As artificial intelligence (AI) is becoming an integral part of modern systems, its reliability is crucial. Creating such AI systems requires an iterative process: improving evaluation methods drive model enhancements, while deeper insights into training processes refine evaluation techniques. Rapid and actionable feedback is essential for efficient model enhancement, and closing the gap between evaluation and training could revolutionize AI.



For the past one and a half years, I have focused on the intersection of **Large Language Model (LLM) evaluation** and **post-training**, working to integrate these domains for more capable AI systems. Working with Professors Ion Stoica, Joseph E. Gonzalez, and Jiantao Jiao, my research spans large-scale crowdsourced LLM evaluation, automated benchmarking, and post-training via Reinforcement Learning from Human Feedback (RLHF). I am part of Berkeley Sky Computing Lab, Large Model System Organization (LMSys), and Nexusflow. My research has been published as ICLR 2024 Spotlight and in ICML 2024, EMNLP 2024, and was recently submitted to ICLR 2025.

Crowdsourced Model Evaluation

Since July 2023, I have been a core researcher and open source contributor to Chatbot Arena (LMSys), with Professors Ion Stoica and Joseph E. Gonzalez, as well as Wei-Lin Chiang, Anastasios Angelopoulos, Lianmin Zheng, Ying Sheng, and Lisa Dunlap. Chatbot arena is one of the largest open-source evaluation platforms for LLMs and VLMs, crowdsourcing over 2 million pairwise preference votes from users to create leaderboards. Through Chatbot Arena, we seek to test models against human preferences on real-world, open-ended user queries, and share our data and analysis with the broader community. We published Chatbot Arena (ICML 2024 Poster), where I contributed to developing methods for analyzing and validating large-scale crowdsourced data. Some of my works included leading efforts to build statistical frameworks and categories for Chatbot Arena, such as Style Control and Hard Prompt, offering deeper insights into more granular model performances. Our platform has emerged as one of the most referenced leaderboards for LLM evaluation, gaining recognition from and fostering close collaborations with many academic research groups and frontier AI labs, including OpenAI, Anthropic, Google, and Meta. Building upon Chatbot Arena's success, we released and published LMSYS-Chat-1M (ICLR 2024 Spotlight), a comprehensive dataset of over 1 million real-world conversations between users and LLMs, where I contributed to developing use cases for the dataset, including exploring applications for LLM safety research and analyzing real-world jailbreak attempts.

While developing crowdsourced evaluation systems, I observed several limitations. One of the primary challenges is that *gathering real-time human feedback is both expensive and time-consuming*, which makes it impractical for frequent model evaluations required during the training. Meanwhile, automated benchmarks often fall short in aligning with real-world user preferences. This realization led me to work

on bridging the gap between the "gold standard" of human preference data and faster, more cost-effective automated evaluation methods.

Benchmark Automation
To make LLM benchmarking scalable and automatic, I led the development of Arena-Hard-Auto (under review ICLR 2025), an novel end-to-end pipeline that fully automates both the data curation process and replaces human voters with an LLM-as-a-Judge approach. Recognizing the absence of established metrics to assess benchmark quality, my team introduced innovative statistical measures to evaluate alignment with human preferences and the model's ability to distinguish performance levels. Arena-Hard-Auto outperformed existing benchmarks, setting a new standard for scalable, automated benchmark creation. It has since been adopted by numerous LLM developers and serves as an official evaluation metric for major AI organizations, including Nvidia, Google, Mistral AI, Alibaba, Deepseek AI, and Microsoft.

Working on automated evaluation deepened my understanding of model behavior, including their strengths and limitations. These insights revealed that many model limitations could be addressed through targeted post-training techniques. Motivated by these observations, I was eager to expand my horizon to model training, specifically the potential of RLHF and fine-tuning to enhance model capabilities.

LLM Post-training
Working with Professor Jiantao Jiao, I co-led the training of Athene-70B, post-trained from Llama-3-70B via RLHF. Leveraging my background in model evaluation to identify and understand key weaknesses in the base model, I explore various techniques on curating a targeted AI preference dataset, which we then used to train a reward model and refine Athene-70B through Proximal Policy Optimization (PPO). We assessed various RLHF strategies and found our novel approach of utilizing a seven-wise comparison instead of the traditional pairwise method to be more effective. Finally, Athene-70B's performance not only surpassed Llama-3.1-70B-Instruct but also rivals GPT-4-Turbo, making it one of the most robust models available in open-weight.

Additionally, during our work on RLHF training, we identified a major bottleneck: assessing downstream LLM performance typically requires running a full training pipeline, which is costly and time-intensive. Since an RLHF-ed model's performance hinges on the quality of its reward model, we prioritized evaluating the reward model's effectiveness before training. To address this, working with Professors Ion Stoica, Joseph E. Gonzalez and Jiantao Jiao, we published How to Evaluate Reward Models for RLHF (under review ICLR 2025), a benchmark that measures reward model reliability for RLHF, and validated its correlation to downstream LLM performances. I contributed by developing methods to assess reward model alignment with human preferences and extending these evaluations to LLM-as-a-Judge scenarios.

Future
In my undergraduate years, I pushed the boundaries of model evaluation and trained state-of-the-art LLMs, equipping me with tools to tackle open challenges, and more importantly, giving me a sense for identifying key problems in the current AI landscape. Moreover, my experiences in model evaluation and post-training complement each other, as researching in both areas constantly inspires me with new ideas on how to enhance the other, creating a loop for iterative improvements. If granted the opportunity to pursue a Ph.D., I aim to explore foundational research that bridges the gap between evaluation and training, deepening our understanding of large models. I envision a future where the entire lifecycle—from collecting feedback to identify model limitations, to designing training datasets and algorithms, to deploying the next iteration of models—is a seamless and efficient process. With an exceptional PhD program, industrial connections, and dedication to high-impact research, I am excited to contribute to shaping the next generation of reliable, intelligent systems.