# Statement of Purpose

**Research Interests:** The advent of foundational models trained through (semi-)supervised pretraining—or imitation learning from human data—has undeniably pushed the boundaries of AI capabilities. This focus, however, has caused us to overlook critical advanced topics essential for scalable oversight, genuine reasoning, and surpassing human-level performance. For instance, RL continues to evolve and has often been rediscovered in foundation models' contexts, e.g. RLHF, where challenges like reward overoptimization have re-emerged. To thrive in the era of foundational models, we need to employ more principled methods rather than relying on blind reinvention or trial and error; we shall reclaim and build upon the "forgotten wisdom" of our field. To advance beyond these limitations and move toward the next generation of AI systems—including scalable oversight in AI alignment and superhuman performance in tasks like reasoning—we must transcend simple pretrain scaling. My research focuses on directions such as:

- **Integrating Advanced Theoretical Frameworks:** How can we incorporate principles from fields like **game theory** and **information theory** to gain deeper insights into foundational models? *e.g., investigating game equilibria in LLM settings; understanding model uncertainty; analyzing information flow in multimodal systems.*
- **Developing Principled Advanced Post-Training Methods:** How can we leverage principles in fields like reinforcement learning and multi-agent systems to improve foundational models beyond their pretraining limitations? *e.g., post-training with convergence guarantees under equilibria to surpass human experts; uncertainty-guided information selection in large model systems.*

My enthusiasm for understanding and improving foundational models with theoretical insights stems from my belief that grounding AI research in strong theoretical foundations is essential. In academia, limited computational resources often restrict our ability to compete with industry's large-scale experiments. By focusing on theoretical principles, we can bridge this gap and contribute meaningfully to the field without relying solely on massive computational power. Moreover, a solid theoretical grounding provides me with confidence and comfort amidst the complexity and unpredictability inherent in deep learning, guiding my research aesthetic toward elegant and principled solutions.

**Core Research Experience:** My journey into combining foundational models with game theory and information theory began during my sophomore year research on LLM alignment algorithms under the guidance of Prof. Yaodong Yang at Peking University. This project aimed to address the issue of safety alignment for LLMs from a novel perspective. Instead of conventional methods like prompt tuning or heuristic attack designs, we proposed a unified framework leveraging game theory for LLM applications. This approach seeks not only to develop safer models but also to create attack models that surpass human experts by moving beyond imitation learning. In our approach, we implemented an adversarial setting where LLMs engaged in red-team/blue-team interactions to improve alignment in safety testing. We modeled this setup as a two-sided zero-sum game, using a population-based multi-agent algorithm with convergence guarantees to reach equilibrium. I implemented the proposed formulation and developed all the coding and experiments, making the theoretical approach practically work. I built an LLM-based multi-agent training pipeline, which was trained on an 8×A100 cluster, providing me with extensive engineering experience in managing large-scale deep learning systems. This work, combining LLM alignment and game theory, is currently under review at T-PAMI and have received a revision decision, where I am a co-first author.

Throughout this project, I encountered a significant challenge: ensuring generation diversity in multi-agent scenarios using LLMs. As training progressed, mode collapse became an issue; an imperfect reward model could lead to reward hacking, where policies exploit loopholes and degrade generation quality. To address this challenge, I investigated techniques to maintain generation diversity and prevent mode collapse. I refined the reward model and introduced regularization methods, which successfully mitigated reward hacking and improved the quality of generations. This experience prompted me to explore the relationship between generation diversity, uncertainty, and model performance.

While working with Prof. Zhiting Hu at the University of California, San Diego, I led a project investigating uncertainty in large models through information-theoretic approaches. I started from first

principles, seeking to understand the fundamental sources of uncertainty. After thoroughly reviewing the existing literature on uncertainty in machine learning, I focused on uncertainty decomposition, as I realized the traditional epistemic-aleatoric dualism might not be applicable or sufficient in the context of foundational models. As a result, we proposed a new uncertainty decomposition framework that quantifies and decouples prediction uncertainty influenced by various factors, including context information, prompting techniques, and even model architecture, examining its relationship with model performance. This work, on which I am the first author has been released as a preprint.

Building upon experience and insights into model uncertainty, I co-led a project to develop a novel decoding strategy for vision-language models (VLM) grounded in information-theoretic principles to mitigate hallucinations. Recognizing that misinterpretations in VLMs often stem from uncertain visual inputs, I proposed leveraging uncertainty quantification and decomposition during inference as a signal for selective information flow. We introduced an innovative method to quantify and decompose inference-time uncertainty into aleatoric and epistemic components without additional supervision. By focusing on epistemic uncertainty, which reflects the model's lack of knowledge, we selectively suppressed high-uncertainty visual tokens during decoding. This approach enabled the model to concentrate on high-confidence information, enhancing the reliability and accuracy of outputs. Our method demonstrated the practical effectiveness of integrating information-theoretic principles into VLM reliable decoding.

**Additional Research Explorations:** In addition to my work on LLM alignment with game theory, I co-authored two papers published at NeurIPS that further reflect my commitment to advancing AI alignment by integrating previously overlooked theoretical principles. The first paper explores multi-objective learning for LoRA-based fine-tuning of large language models, aiming to achieve Pareto optimality across multiple dimensions of human preferences through RLHF. By incorporating concepts like Pareto optimality from multi-objective learning, we move beyond naive PPO to better steer LLMs for personalization and alignment. The second paper centers on multimodal alignment, introducing a dataset and benchmark that emphasize the harmless-helpful trade-off for multimodal models, specifically video generation models. This work addresses the challenge of aligning multimodal large models with human preferences and intentions.

**Conclusion:** After my undergraduate research journey, I am eager to continue this research path in a Ph.D. program. I firmly believe that game theory, information theory, and reinforcement learning provide essential frameworks for developing more powerful and steerable models that extend beyond supervised learning. Reinforcement learning and game theory offer structured methodologies to guide models toward higher states of performance, like achieving a higher ELO score in a generalized game. By employing information-theoretic metrics, we can gain valuable insights into how these models process information, make decisions, and how various inputs affect their outputs. My vision is to combine theoretical foundations with practical applications to create work that not only advances the field but also benefits the community and brings me personal fulfillment. I aspired to establish a universal framework that transcends previous assumptions and is applicable across diverse AI systems. Although this journey is challenging, I am fully committed to pursuing my goals.