Introduction

Personal Background. My academic journey, driven by a profound passion for research, began during my undergraduate studies in Computer Science at East China University of Science and Technology. I was captivated by the power and versatility of artificial intelligence (AI) in the world today. As I excelled academically, earning various accolades, including the prestigious National Scholarship, I became increasingly committed to research. This passion led me to internships at renowned institutions such as the **Shanghai Artificial Intelligence Laboratory**, **Tsinghua University**, and **Microsoft Research Asia**, where I contributed to several significant projects, including the development of the open-source LLM **InternLM2**. These hands-on experiences continuously improved my research skills and programming expertise. To deepen my research, I was later fortunate to work as a research assistant advised by **Prof. Huaxiu Yao** in the **CS department of University of North Carolina at Chapel Hill**, where I led multiple research projects and successfully contributed papers to top-tier AI conferences. These experiences have solidified my research capabilities and strengthened my resolve to pursue an academic career.

Research Interest. Recent advancements in foundational models have led me to deeply reflect on several pressing challenges. The prevailing approach to improving model performance heavily relies on expanding the scale of training data. However, this strategy faces an inherent bottleneck: while Internet data is abundant, it is ultimately finite, inevitably limiting the potential for further model scalability. Moreover, Whether the Internet-Level dataset is safety for foundational models need to be explored. At the same time, I find myself questioning whether the current training paradigm—comprising autoregressive pretraining, supervised fine-tuning, and RLHF-based post-training—truly represents the most optimal and sustainable path for future development.

Given my understanding of the current challenges in the field of AI , my primary research interest lies in the two fields: (1)**efficient training of foundational models** and (2)**trustworthy machine learning**. Specifically, I focus on three key subfields: **(1) data-efficient training, involving data selection and prioritization, (2) developing more effective alignment and post-training methodologies, and (3) AI Safety in LLM**.

Research Experience

Research in Data-efficient Learning. Driven by my passion for improving data-efficient training methods, I joined the Shanghai Artificial Intelligence Laboratory as a research intern in Fall 2023 under the mentorship of Dr. Yining Li. My research aimed to enhance data selection efficiency to improve training outcomes. Through rigorous theoretical analysis and practical experimentation, I proposed a novel data selection method that improves upon an existing in-context learning-based approach. This new method overcomes the issues of excessive computational resource consumption and the possibility to select low-quality data, which are common in the original method. I independently implemented the code and conducted the related selection and evaluation experiments. The results showed that my approach improved model performance using only 30% of the supervised fine-tuning (SFT) data, while reducing the selection time to one-quarter of that required by previous methods. This efficient data selection framework was subsequently applied to train the open-source large language model, InternLM2. This experience deepened my interest in data-efficient training and further fueled my exploration of this field.

In Summer 2024, I worked as a research assistant at Tsinghua University under the guidance of Professor Xianyuan Zhan. I proposed a novel method that leverages the self-critic capabilities of large language models (LLMs) to assess data quality without introducing external models. This method dynamically adjusts the learning rate for each data point based on its quality. After independently implementing the code and conducting subsequent experiments, the method led to efficient improvements in the Llama model's performance, without introducing additional computational burden.

Later, during my internship at Microsoft Research Asia's Machine Learning Group under Dr. Zhong Li, I conducted research on data-efficient training methods. I identified a common challenge in this area: How to model the training dataset in a way that enables dynamic sampling during training. With my solid mathematical background, I recognized that the training dataset could be effectively modeled as a Gaussian Mixture Model (GMM). The statistical parameters derived from the GMM, such as mean and covariance, could be used to dynamically adjust the training process. Building on this theory, I implemented the code and combined loss parameters with statistical parameters to optimize batch sampling during training. Unlike traditional random sampling approaches, this strategy enabled the model to learn more effectively from existing datasets. I am preparing to submit my findings to *ICML* in the coming months.

Research in developing more effective alignment and post-training methodologies. While these experiences have provided me with a solid foundation in data-efficient training, I have come to realize that my research so far has been confined to the limits of existing training paradigms. Are autoregressive pretraining, supervised fine-tuning, and RLHF the most optimal strategies? To explore alternative approaches, I collaborated with Professor Huaxiu Yao at the University of North Carolina at Chapel Hill to investigate efficient alignment strategies in robotic learning. Specifically, for vision-language-action (VLA) models, existing training methods rely on supervised fine-tuning, where models learn task execution from human demonstration trajectories. However, these methods often fail to teach robots how to perform tasks safely and efficiently. To address this challenge, I proposed trajectory-wise preference optimization, a novel approach that enables VLA models to learn task execution based on human preferences for trajectory-level outcomes. This strategy achieved state-of-the-art performance in a range of simulated and real-world experiments. I led the project, developed the code, designed rigorous experiments and produced two papers, which are currently under review for *ICLR 2025* and *CVPR 2025*(**First Author**).

Research in AI Safety in LLM. In Shanghai Artificial Intelligence Lab, as a Research Intern in InternLM2 team, I conducted AI Safety research in InternLM2 under the mentorship of Dr. Yining Li. My research aimed to identify and defend against the "Indentity Attack" when users interact with InternLM2. I successfully proposed a novel idea of using GPT models to generate adversarial training datasets to SFT internLM2 and developed a benchmark to evaluate the model's defense capabilities. This fresh and comprehensive approach successfully enhanced InternLM2's robustness against identity attacks.

Future Goals

The questions that have arisen throughout my research have inspired me to pursue a Ph.D. in Computer Science. I am eager to develop training paradigms that transcend scaling laws and contribute to the sustainable progress of artificial general intelligence (AGI) and trustworthy machine learning. My goal is to explore not only data-efficient training, but also alternative pre-training and alignment strategies that challenge conventional methods. By doing so, I hope to contribute to the development of foundational models that are not only more powerful but also more resource-efficient and aligned with human values.