Personal Statement

Changshuo Shen

Email: stephen_shen@mail.ustc.edu.cn

School of Artificial Inteligence and Data Science, University of Science and Technology of China

Introduction

As an undergraduate student at the School of Artificial Intelligence and Data Science at the University of Science and Technology of China (USTC), majoring in Data Science, I have developed a strong interest in AI safety and explainability. Throughout my academic journey, research experiences, and extracurricular activities, I have solidified my aspiration to contribute to the field of AI safety and explainability.

Academic Background

My academic journey has been marked by significant progress and growth. Starting in the 60% of my class during my freshman year, I have steadily advanced to the top 10% (6th out of 51) in my major, with a current GPA of 3.91/4.3 (90.53). This progress reflects my dedication to self-improvement and my resilience in overcoming challenges.

Additionally, I have completed courses highly relevant to artificial intelligence, including Machine Learning, Deep Learning and Probability Theory and Mathematical Statistics. These courses have provided me with a solid theoretical foundation and practical skills, preparing me to contribute to cutting-edge research in AI and related fields.

Research Experience

I am currently collaborating on a research project under the guidance of Professor An Zhang, who previously served as a research fellow at the NUS Next++ Research Group and recently joined USTC as a professor, and with the support of PhD student Leheng Sheng (NUS Next++). Our work focuses on AI safety, specifically exploring the mechanisms behind jailbreak scenarios in (LLMs) using advanced theoretical frameworks and innovative tools. My contributions include constructing datasets tailored to the research problem, designing and implementing experimental setups to investigate the mechanisms behind LLM jailbreak scenarios, and collaborating on the analysis of preliminary findings. Moving forward, I will continue to actively participate in refining the research methodology and co-authoring the resulting paper.

This project has not only deepened my understanding of AI safety but also enhanced my technical and critical thinking skills as I work on addressing real-world challenges in this critical area. My involvement in this research has further solidified my passion for AI safety, and I am eager to contribute to advancing the robustness and trustworthiness of AI systems through innovative solutions.

Beyond Academics

Outside of my academic pursuits, I have actively engaged in diverse experiences that developed my leadership, communication, and teamwork skills. Leading the code team for USTC-Software at the IGEM Jamboree in Paris, I oversaw platform development and presented our work to a global audience, fostering cross-cultural collaboration and problem-solving. Similarly, participating in an exchange program at HKUST allowed me to explore the intersection of machine learning and physics, broadening my interdisciplinary perspective.

As president of the USTC English Club, I organized impactful events like the "USTC Mystery Hunt" and managed the club's communications, honing my organizational and leadership abilities. My involvement in the USTC Hosting Club and speech competitions further enhanced my public speaking and interpersonal skills, while volunteer initiatives strengthened my commitment to community engagement.

In addition, I balanced these efforts with an active lifestyle, competing in basketball tournaments and participating in fitness and taekwondo clubs, which reinforced my resilience and teamwork. These experiences beyond academics have shaped me into a well-rounded individual, ready to contribute in dynamic and collaborative environments.

Research Interests and Goals

My current research interests focus on understanding the mechanisms behind large language models (LLMs), particularly exploring their internal processes from the perspectives of representation spaces or model parameters. I am fascinated by how these models encode, process, and represent information, and I aim to leverage interpretability tools, such as Sparse Autoencoders, to gain deeper insights into these mechanisms. This exploration not only satisfies my intellectual curiosity but also holds the potential to improve the reliability and transparency of LLMs.

In the future, I hope to contribute to advancing the understanding and usability of LLMs by addressing fundamental questions about their design and behavior. My goal is to produce meaningful work in this domain, enhancing both the theoretical foundations and practical applications of these models. This is a field I am deeply passionate about and one to which I am committed to dedicating significant effort.

Conclusion

As someone who is deeply self-motivated and thrives on intellectual challenges, I am eager to delve deeper into the mechanisms of large language models and make meaningful contributions to this rapidly evolving field. My passion for exploring complex ideas, combined with my dedication to research, drives me to push boundaries and seek innovative solutions. I believe that with my curiosity, determination, and commitment to excellence, I can not only grow as a researcher but also contribute to advancing the understanding and development of intelligent systems. I am excited to embrace the opportunities and challenges ahead, confident in my ability to make a lasting impact.