I am an alumnus of Peking University, currently working as an AI engineer specializing in large language model pretraining at a technology company. My passion for AI, coupled with the urgent need to enhance its safety and trustworthiness, drives my pursuit of a PhD in this field. While AI systems have demonstrated remarkable potential across domains, recent challenges in reliability, fairness, and ethical use call for rigorous research into building trustworthy models. These concerns inspire me to contribute to the development of systems that are not only intelligent but also safe and trustworthy.

My journey in AI research began during my master's studies, where I worked on building trustworthy legal AI systems under the supervision of Prof. Yansong Feng. We focused on integrating legal knowledge into AI models, improving their reliability by ensuring that the system's decisions were based on solid legal basis. Our project culminated in developing Lawyer-LLaMA—the first conversational Chinese legal consulting model, designed to provide accurate legal advice with enhanced trustworthiness. This experience fostered my research skills, from problem analysis and hypothesis formulation to experiment design. It also highlighted the need for AI systems to be transparent and accountable when applied to high-stakes domains such as law.

At my current company, I am responsible for independently training Mixture of Experts (MoE) models—a sparse architecture designed to reduce both training and inference costs while maintaining high performance. This role requires strong research capabilities, as I survey a wide range of academic papers, technical reports, and open-source projects to identify the most effective methods. Additionally, it demands advanced coding skills to implement these methods in robust, distributed systems that operate across hundreds of nodes. I am also a core member of out LLM data group, where I focus on improving data quality, detecting harmful data, and conducting data synthesis to ensure our models are trained on reliable datasets, . This work deepens my understanding about how data influence the ability of LLM.

In this role, I have contributed to the development of large language models ranging from 8B to 112B parameters. These efforts have allowed me to develop a strong foundation in model training and practical problem-solving, preparing me for the challenges of PhD-level research.

My interest in building trustworthy AI systems extends beyond model efficiency. As a core member of our LLM data group, I focus on improving data quality, detecting harmful data, and conducting data synthesis to ensure that our models are trained on reliable, diverse datasets. This work underscores the importance of creating AI models

that are not only accurate but also fair and unbiased, aligning with my long-term goal of advancing the trustworthiness of AI systems.

In addition to my technical experience, I have worked as a core technical developer on a commercial role-play conversational product, which broadened my understanding of human-AI interaction. Beyond the technical challenges of building chat models, I gained valuable insights into how people engage with AI systems. These interactions deepened my appreciation for the need for AI models that can reason logically, engage meaningfully with users, and verify the trustworthiness of their responses using external sources of information. This is the core of my research interest: developing conversational AI models that are not only capable of understanding language and reasoning but also able to verify the accuracy and reliability of their outputs.

Looking forward, my research interests lie in two main areas. First, I am dedicated to developing trustworthy conversational AI systems that can engage in meaningful and safe interactions with humans. I believe that achieving this requires advancements in three areas: improving language understanding and reasoning, enhancing interactivity, and verifying the trustworthiness of AI-generated responses. Second, I am deeply interested in the intersection of AI and social science. As AI systems

become more integrated into human society, understanding how they interact with humans and each other will be critical. I aim to contribute to research that explores the dynamics of human-AI societies, testing social science theories to ensure harmonious and effective integration.

This PhD program offers the ideal environment to pursue these goals. I am particularly excited about the work being done by [mention specific professors or research areas], whose research aligns closely with my interests in trustworthy AI and its broader societal implications. The interdisciplinary nature of the program, combined with access to cutting-edge resources, will provide the perfect platform for me to develop the skills necessary to tackle the challenges of AI trustworthiness.

In the long term, I aim to contribute to research that ensures AI technology is developed with safety, fairness, and ethical considerations at its core. While AI has the potential to transform society in positive ways—such as by improving daily life through intelligent automation—it also presents risks, including misinformation, fraud, and privacy concerns. By working in research institutions, I hope to lead projects that address these risks and help shape the future of AI in a way that benefits society. Through this PhD program, I am eager to develop the research skills and knowledge needed to make a meaningful impact in the field of

trustworthy AI.

This version adds depth and specificity, connecting your experiences to your research goals, and highlights your personal motivation for pursuing trustworthy AI research. It also emphasizes the relevance of your previous work and how it has prepared you for PhD-level research.

Name: Dynamic Mixture of Experts Models
Content: Most of MoE models activate fixed number of parameters during training and inference, we propose to dynamically allocated activated parameters based on the difficulty of input tokens, therefore implenting better sparsity.