# Research Statement of Kaijie Zhu

My research is driven by a commitment to advancing trustworthy artificial intelligence (AI), with a specific focus on adversarial robustness and the evaluation of foundational models, such as large language models (LLMs) and vision foundation models. In the rapidly evolving landscape of AI, my work aims to address critical gaps in ensuring the reliable and safe use of these widely-applied technologies.

## Prior Work

My master's research (worked as the primary author) focused on two main areas: the *adversarial robustness* and the *evaluation protocol* for LLMs.

1. **Adversarial robustness.** In this line of research, we first studied a fundamental problem in adversarial training: the trade-off between models' robustness to adversarial samples and their generalization capability. We proposed a robust-critical fine-tuning technique [12] to maintain the adversarially trained models' robustness while improving their generalization ability. In another work, we investigated the resilience of LLMs to *adversarial prompts*. Specifically, we proposed 4 levels (character, word, sentence, and semantic) of textual adversarial attacks to subtle manipulate the input prompts [13]. Our findings revealed a significant performance degradation in LLMs due to subtle, human-imperceptible modifications in prompts, underlining a major obstacle for their reliable deployment.

2. **Evaluation protocol for LLMs.** The heavy reliance of LLMs on expansive training corpora highlights two primary evaluation concerns: the risk of test data contamination and the obsolescence of static benchmarks due to ongoing LLM evolution. In response, we developed DyVal [11], a dynamic evaluation protocol that facilitates *complexity-tailored, on-the-fly generation* of evaluation samples in reasoning tasks. This approach guarantees the uniqueness of evaluation samples, thereby significantly reducing the risk of test data contamination. Additionally, it offers benchmarks that are adaptable, with their complexity modulated to match the progressive development of LLM capabilities.

In addition to my primary research directions and projects, I have contributed to three projects (worked as the co-author): (1) a comprehensive evaluation survey of LLMs [1]; (2) improving LLMs performance through psychological stimuli [6]; and (3) examining competitive behaviors among AI agents [10]. These experiences position me uniquely to contribute to the evolving landscape of AI research, with a particular emphasis on the development and understanding of LLMs.

## Future Directions

Building on my previous work, my future research will continue to focus on enhancing AI trustworthiness and evaluating foundation models, emphasizing three interconnected domains vital to the advancement and safety of these models:

1. **Enhancing AI Trustworthiness:**

- **Reinforce the robustness of foundation models to unexpected inputs.** Ensuring robustness in foundation models against unexpected inputs, such as adversarial examples, jailbreak, and backdoor attacks, is crucial. This robustness is vital to prevent the generation of malicious content, particularly as these models become increasingly ubiquitous in safety-sensitive applications. Current methodologies for achieving robustness in AI models, particularly adversarial training [7], confront several critical challenges: (1) the *trade-off between robustness and accuracy* [9]; (2) the substantial *training costs* required; (3) a lack of guaranteed robustness even in certified robustness methods [2]. Given these constraints, there is an urgent need for innovative approaches that transcend traditional adversarial training methods to enhance the safety of foundation models, ensuring their reliability in diverse applications.
- **Detecting AI-Generated Content (AIGC).** Detecting AIGC is an essential *complementary strategy* to directly reinforcing the robustness of foundation models. This dual approach provides more comprehensive safeguarding of AI systems, addressing both the prevention and *identification* of potentially harmful or misleading outputs of foundation models. While current methods such as watermarking [4] techniques are promising, they are still vulnerable to perturbations, such as paraphrasing attacks [5, 8] and adversarial examples [3]. Focusing on the development of robust detection algorithms that can accurately identify AIGC is crutial for mitigating the spread of misinformation and ensuring the authenticity of digital content.

2. **Evaluating foundation models.** The progression of Artificial Intelligence (AI) is closely tied to the development and meticulous evaluation of its foundational models. Accurate and comprehensive evaluation is crucial as it verifies the true intelligence of a model. The emergence of foundation models, especially LLMs, presents three primary challenges to their evaluation:

   - **Test data contamination.** The expansive and diverse training datasets of LLMs elevate the risk of overlap between training and testing data. This contamination can lead to inflated performance metrics, giving a misleading impression of LLMs' genuine capabilities. Therefore, developing a new evaluation framework or devising methods for detecting data contamination, is imperative. Building on my previous foundation, we plan to extend DyVal's dynamic evaluation approach [11] to the realm of natural language understanding tasks and computer vision tasks.
   - **New evaluation measurements for generation models.** Traditional evaluation metrics are often inadequate for generative models such as LLMs, diffusion models. These models produce outputs that can be highly variable and yet equally valid. In contrast to tasks with definitive answers, formulating evaluation metrics that effectively assess their generative ability poses a significant challenge.
   - **Evaluation benchmarks reflecting diverse real-world scenarios.** The deployment of LLMs spans a variety of challenging environments, including coding, decision-making, and AI agent interactions. This variety highlights the limitations of current benchmarks, which often fall short in capturing the complexity and dynamism of real-world applications. To truly assess the utility and robustness of LLMs, it is imperative to develop benchmarks that accurately mirror these diverse and dynamic conditions.

## References


[1] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.

[2] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

[3] Zhengyuan Jiang, Jinghuai Zhang, and Neil Zhenqiang Gong. Evading watermark based detection of ai-generated content, 2023.

[4] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR.

[5] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the reliability of watermarks for large language models. *arXiv preprint arXiv:2306.04634*, 2023.

[6] Cheng Li, Jindong Wang, Kaijie Zhu, Yixuan Zhang, Wenxin Hou, Jianxun Lian, and Xing Xie. Emotionprompt: Leveraging psychology for large language models enhancement via emotional stimulus. *arXiv preprint arXiv:2307.11760*, 2023.

[7] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[8] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected?, 2023.

[9] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[10] Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. Competeai: Understanding the competition behaviors in large language model-based agents, 2023.

[11] Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. Dyval: Graph-informed dynamic evaluation of large language models, 2023.

[12] Kaijie Zhu, Xixu Hu, Jindong Wang, Xing Xie, and Ge Yang. Improving generalization of adversarial training via robust critical fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4424–4434, 2023.

[13] Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*, 2023.