Statement of Purpose

When I began research it has been a wild time—established knowledge often yielded ground to the new era of foundation models. For example, a language model agent that can browse the web and use tools was considered nearly impossible. With more models coming out every single day, I have learned to focus on research directions that can stay relevant in the long term. One such direction I find worth pursuing is **Reliable AI**. In particular, *(i)* How can we ensure **AI safety** such as protecting model ownership [1], aligning models with human preferences [2], and defending against malicious exploitation of model vulnerabilities [3, 4]? *(ii)* How can we train AI that generalizes well in **low-resource** regime where limited data are available [5–10]? I am fortunate to have had some initial attempts with many wonderful collaborators at USC, Amazon, Harvard, Stanford, and Microsoft.

Make AI Safe Again

Ensuring AI safety is increasingly important as models are reaching millions of users worldwide. As one of the users who interacts with AI to automate daily chores, I deeply care about the impact of AI on human users, which motivates me to work on safer AI, both for users and model developers.

Fingerprint for LLM Ownership [1]: As companies open-source strong LLMs and developers throughout the world build customized models derived from their releases, there's a growing need for effective protocols to protect model intellectual property and prevent malicious users from taking the released models and claiming ownership without respecting the original license. At **USC** with Prof. **Muhao Chen** and Prof. **Chaowei Xiao**, we presented a first attempt to fingerprint LLMs to safeguard ownership by using "poison" attacks benignly. Ownership verification is reduced to whether a model can recall the "poisoned" instances after significant finetuning (as verification might occur after users have finetuned on their dataset). With language models' memorization capacity, we developed a lightweight fingerprint that works on a wide range of popular LLMs. We also showcased that it is even possible for multiple organizations to implant different fingerprints on the same model, akin to MIT License.

Aligning with Human Preferences [2]: Current reward models rely on shallow encoders that often struggle with generalization in multi-turn conversations and introduce a linear head that is challenging to optimize. At **Amazon Science** with Dr. **Michael Johnston**, we showed a finetuned decoder-only LLM on publicly-available preference datasets, implicitly aligning with human preferences, can match propriety LLMs in reward scoring. Furthermore, due to benefits of the decoder architecture, we showed such reward models can zero-shot generalize to unseen dimensions and domains, facilitate effective data filtering, provide rationales to explain decisions, and generate synthetic multi-turn conversations for AI self-improvement (RLAIF).

Backdoor Vulnerabilities and Test-Time Defense [3, 4]: Data-hungry LLMs often train on crowdsourced corpus, creating vulnerabilities for malicious attackers to inject seemingly innocuous poison data to compromise the trained model. With Prof. **Muhao Chen** and Prof. **Chaowei Xiao**, we demonstrated the feasibility of using less than 1000 tokens to poison instruction-tuned models, such that models will always give wrong answers when attackers present poison in the user query. Such poison is transferable to multiple tasks, and can hardly be cured by continual learning [3]. To remedy such dangerous security risks, we then explored using in-context demonstrations as an effective test-time defense against various backdoor attacks. We found that LLMs' reasoning over the clean demonstrations can influence and rectify the model behavior such that poisoned models would not be affected even if poison is present in the user query [4].

Next Steps: Another crucial direction for safety is interpretability. Inspired by my previous works on learning from explanation [11] and ongoing project at **Harvard** with Prof. **Himabindu Lakkaraju** that applies mechanistic interpretability on LLaVA-like vision language models to understand inner workings of models to respond to complex queries (*e.g.* referring expression, object counting), I want to apply interpretability to model memorization, *i.e.* how can a model recall a specific fact/knowledge? I hope that with a deeper understanding of this aspect, I can reversely utilize the memorization capacity to prevent memorizing the poison, thereby creating a better defense for backdoor attacks.

Less Is More: Excelling in Low-Resource Regimes

While current foundation models have achieved "superhuman" performance in many academic benchmarks, they falter in less-general domains, *e.g.* biomedical relation extraction and instance segmentation in low-lighting rooms. This shortfall is often due to the scarcity and high cost of annotations in those less-explored domains. My research is driven by the challenge of enabling these models to adapt and excel in those low-resource environments, such that AI can gradually generalize across the board.

On-Demand Synthetic Data Generation [5–8]: Automatic synthetic data, deemed as a cure for annotation shortage, is limited in controllability, personalization, and diversity, thus often collapsing into common distributions (*i.e.* failing to simulate out-of-distribution scenarios). Training on such datasets leads to models with limited generalization. At **Stanford** with Prof. **Jiajun Wu**, we developed a photorealistic and physically accurate simulator to systematically curate parametric datasets for out-of-distribution scenarios (*e.g.* low lighting, object articulation). Such synthetic datasets also bolster performance when annotations are costly to obtain (*e.g.* spatial relationship between objects) [5]. At **Microsoft Research** with Dr. **Vibhav Vineet**, we leveraged prompt distribution learning to improve diffusion models' text-to-image and text-to-3D generation quality, diversity and personalization such that generalization on personalized tasks becomes possible via synthetic data [6]; we utilized diffusion models and object cut-and-paste to create coherent synthetic training datasets for zero-shot object detection and segmentation [7]; and we proposed differentiable synthetic data generation with NeRF to improve out-of-distribution object detection from varying camera views [8]. Observing the substantial performance boost, I believe through these enhanced generation pipelines, AI can bridge the "last mile" to support long-tailed but vital scenarios that downstream users worldwide can benefit from.

Indirect Supervision From Resource-Rich Tasks [9, 10]: In cases where even synthetic data are challenging to obtain such as the biomedical field, one feasible solution is data-efficient training that maximizes learning supervision from limited training data. I focused on one such method—indirect supervision, which borrows supervision signals from resource-rich tasks to enhance resource-limited tasks. With Prof. **Muhao Chen**, I studied how to apply indirect supervision to automate biomedical relation extraction given the intricacy and expense associated with manual extraction of biomedical knowledge from massive corpora. We reformulated relation extraction as a natural language inference (NLI) task, and trained general domain NLI models on few-shot examples. The core idea was to utilize the inductive bias inherent in NLI models to enable effective cross-domain and cross-task signal transfer. We observed up to 300% improvement on few-shot setting and moderate zero-shot performance [9]. Further exploring the upper limits of cross-task signal transfer, we investigated transferring three semantic typing tasks for large label space inference in a multi-task framework [10].

Next Steps: Inspired by my previous work on synthetic vision data, I am interested in exploring synthetic language data, *e.g.* linguistic-grounded data generation for complex grammar structure understanding. I am also excited to apply indirect supervision in inference, investigating whether different distributions of in-context demonstrations can cross-task benefit target task, *e.g.* do inductive reasoning demonstrations benefit deductive reasoning? Do coarse-grained classification demonstrations benefit fine-grained classification?

Career Goals

My long-term career goal is to become a researcher who contributes to advancing reliable AI with tangible real-world impact. My past journey in both computer vision and natural language processing taught me to value interdisciplinary studies and collaboration between diverse groups.

At Waterloo: I look forward to collaborating with Prof. **Victor Zhong** as we both share interests in methods to generalize and adapt to new domains; Prof. **Yuntian Deng** as his research on reliable multi-LLM communication is deeply connected with my research on reliability; and Prof. **Wenhu Chen** for our joint interest in reliable multimodal model applications. Lastly, I highly value the collaborative and inclusive environment at University of Waterloo and Vector Institute, which will help me expand my research perspective and conduct highly impactful research.

References

* means equal contribution. Also see my CV for paper links. Please do not share under review papers externally.

[1] **Jiashu Xu**, Fei Wang, Mingyu Derek Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. Fingerprinting large language models. *NAACL*, 2024 (Under Review).

[2] **Jiashu Xu**, Daniel Pressel, Prasoon Goyal, Luke Dai, Reza Ghanadan, and Michael Johnston. Training large language models as reward models. *Conference on Language Modeling*, 2024 (Under Amazon Internal Review).

[3] **Jiashu Xu**, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. *NAACL*, 2024 (Under Review).

[4] Wenjie Mo, **Jiashu Xu**, Qin Liu, Jiongxiao Wang, Jun Yan, Chaowei Xiao, and Muhao Chen. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. *NAACL*, 2024 (Under Review).

[5] Yunhao Ge*, Yihe Tang*, **Jiashu Xu***, Cem Gokmen*, Chengshu Li, Wensi Ai, Benjamin Jose Martinez, Arman Aydin, Mona Anvari, Ayush K Chakravarthy, Hong-Xing Yu, Josiah Wong, Sanjana Srivastava, Sharon Lee, Shengxin Zha, Laurent Itti, Yunzhu Li, Roberto Martín-Martín, Miao Liu, Pengchuan Zhang, Ruohan Zhang, Fei-Fei Li, and Jiajun Wu. Behavior vision suite: Customizable dataset generation via simulation. *CVPR*, 2024 (Under Review).

[6] Brian Nlong Zhao, **Jiashu Xu***, Yuhang Xiao*, Xinyang Jiang, Yifan Yang, Dongsheng Li, Laurent Itti, Yunhao Ge, and Vibhav Vineet. Dreamdistribution: Prompt distribution learning for text-to-image diffusion models. *CVPR*, 2024 (Under Review).

[7] Yunhao Ge*, **Jiashu Xu***, Brian Nlong Zhao, Laurent Itti, and Vibhav Vineet. Dall-e for detection: Language-driven context image synthesis for object detection. *Arxiv*, 2022.

[8] Yunhao Ge, Harkirat Behl*, **Jiashu Xu***, Suriya Gunasekar, Neel Joshi, Yale Song, Xin Wang, Laurent Itti, and Vibhav Vineet. Neural-sim: Learning to generate training data with nerf. *ECCV*, 2022.

[9] **Jiashu Xu**, Mingyu Derek Ma, and Muhao Chen. Can NLI provide proper indirect supervision for low-resource biomedical relation extraction? *ACL*, 2023 (**Oral**).

[10] James Y. Huang, Bangzheng Li*, **Jiashu Xu***, and Muhao Chen. Unified semantic typing with meaningful label inference. *NAACL*, 2022.

[11] Aaron Chan, **Jiashu Xu**, Boyuan Long, Soumya Sanyal, Tanishq Gupta, and Xiang Ren. Salkg: Learning from knowledge graph explanations for commonsense reasoning. *NeurIPS*, 2021.