Research Statement: The Science of Large Language Models

Xinyi Wang (xinyi_wang@ucsb.edu)

The advent of large language models (LLMs) has significantly transformed machine learning and AI research, achieving remarkable performance and generalization across numerous application domains [o1]. Much of the current focus has been on advancing LLM capabilities through empirical optimization of inference algorithms, training techniques, and dataset construction. Despite these strides, there remains a considerable gap between the rapidly improving empirical performance of LLMs and our fundamental understanding of their behavior. This gap presents critical challenges and risks: without deeper insights, LLMs may exhibit unpredictable behaviors in sensitive applications, leading to undesired outcomes. Moreover, the limitations of our current understanding restrict the trajectory of AI development. Achieving greater artificial intelligence is unlikely to result merely from scaling up models and data alone.

My research aims to bridge this gap by investigating the inner workings of LLMs through carefully designed experiments and supporting theories. My research interests lie at the intersection of theory, data, and algorithms, as illustrated in Figure 1. I am a strong advocate for grounding theoretical insights in real-world scenarios: As deep learning models increase in complexity, the connection between theoretical analysis and empirical behavior becomes increasingly obscure. My prior research focuses on abstracting a neural network as a perfect training data distribution estimator, with theories centered on data distribution modeling. This perspective has allowed me to identify the potential underlying causes of many deep learning phenomena—such as in-context learning, reasoning, zero-shot prompting, and spurious correlations—in the characteristics of training data distributions. My prior research can be categorized into the following main aspects:

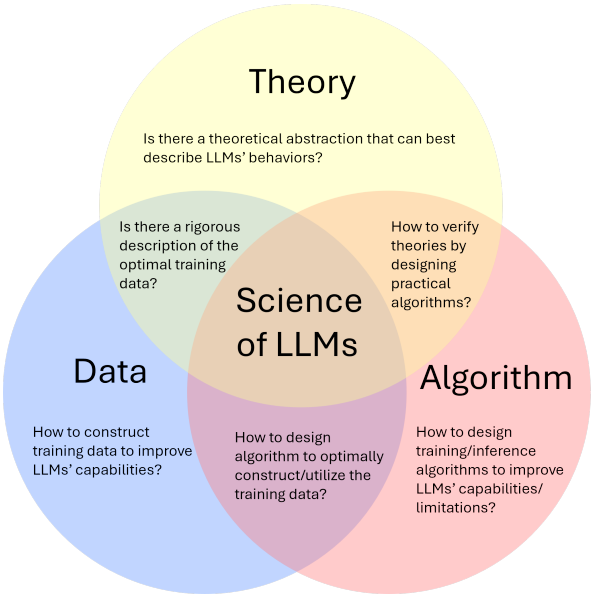

Figure 1: Illustration of my research interest.

- **Understanding and improving LLMs' capabilities with data distribution modeling**: Investigating LLM capabilities like in-context learning, reasoning, and zero-shot prompting through modeling pre-training data distributions, and proposing actionable algorithms to enhance these capabilities.
- **Designing algorithms to address fundamental limitations of deep learning**: Tackling core challenges like spurious correlations and weak structural reasoning by proposing novel training and inference algorithms.
- **Contributing resources to the LLM/AI research community**: Developing and sharing resources such as surveys and datasets for the benefit of the research community.

1 Understanding and Improving LLMs' Capabilities with Data Distribution Modeling

The success of LLMs is closely tied to the vast datasets used during their pre-training. It naturally follows that the properties of these data distributions can provide insights into the models' capabilities. My research explores the question: *What characteristics of the training data distribution give rise to specific LLM capabilities?*

In our NeurIPS paper [p1], we provided both theoretical and empirical evidence that topic-model-like latent variable structures within pre-training data are fundamental to LLMs' in-context learning abilities. Our theoretical analysis demonstrates that LLMs infer latent variables from prompts to generate appropriate continuations, by **proving that the in-context learning classifier can reach Bayes optimality by choosing the correct set of demonstrations**. Building on this insight, we proposed a novel demonstration selection algorithm based on soft-prompt tuning, which significantly enhances in-context learning performance across various text classification tasks and LLM architectures compared to random selection by more than 7% on average. **This paper has accumulated over 100 highly influential citations within a year of publication.**

In our ICML work [p2], we introduced a novel hypothesis that random-walk-like reasoning paths present in pre-training corpora enable LLMs to perform complex multi-step reasoning without fine-tuning. Inspired by the classic knowledge graph (KG) completion algorithm, Path Ranking Algorithm (PRA) [o2], we observed that an LLM pre-trained on a KG can generate distributions that closely mirror those produced by PRA when prompted to complete unseen triples. Furthermore, we demonstrated that for general reasoning tasks like solving math word problems with chain-of-thoughts (CoTs), augmenting the CoT training set with random-walk-like reasoning paths substantially improves performance across diverse datasets, including multihop factual question answering, logical reasoning, and math reasoning. **Our proposed reasoning mechanism for LLMs is among the first to show empirical effectiveness in real-world CoT tasks.**

Our recent investigation into LLMs' zero-shot prompting abilities [p3] reveals that LLMs exhibit varying levels of distributional memorization and generalization when prompted with different tasks. To quantify distributional memorization, we measured the correlation between pretraining data probabilities, estimated by a task-semantic-related n-gram language model, and the LLM-predicted probabilities on testing data. Our analysis indicates that knowledge-intensive tasks benefit from memorization, while reasoning-intensive tasks benefit from generalization—i.e., generating more novel content. Additionally, we traced the influence of relevant pre-training documents throughout the LLMs' training process to support our findings from a causal perspective. **Our work is among the first to comprehensively analyze LLMs' memorization effects across the entire pre-training corpus for a wide range of tasks**, including translation, factual question answering, and reasoning.

These studies demonstrate that examining the training data distribution provides a valuable perspective for understanding diverse LLM phenomena, such as in-context learning, multi-step reasoning, and memorization. This approach not only offers a Bayesian framework for theoretical analysis but also establishes connections to real-world tasks for empirical validation. Beyond studying what deep learning models can achieve, I am also dedicated to addressing their limitations. In the next section, I will focus on my work related to mitigating the challenges of spurious correlations and structured reasoning.

2 Designing Algorithms to Address Fundamental Limitations of Deep Learning

While deep learning models, including LLMs, have achieved significant success, they also exhibit several well-known fundamental limitations. A key issue is the use of maximum likelihood estimation (MLE) as the standard training objective for nearly all deep learning models. MLE primarily captures associations rather than causal relationships between variables, making these models susceptible to spurious correlations present in the training data. To address this, we first model the training data distribution using a causal graph comprising input variables, output variables, and confounders. We then seek to eliminate or reduce the impact of confounders by proposing new training objectives or by resampling the data to augment the observed data distribution into a spurious-free one.

In our NeurIPS paper [p4], we introduced a novel *counterfactual maximum likelihood estimation (CMLE)* training framework, which mitigates spurious correlations by augmenting the MLE training objective with implicitly or explicitly generated counterfactual examples. This approach aims to mimic a spurious-free, randomized data distribution. **We proved that the CMLE objectives serve as upper bounds of the expected negative log-likelihood of the spurious-free data.** CMLE demonstrated enhanced robustness and faithfulness across both synthetic and real-world tasks, such as natural language inference and image captioning.

In our ICLR paper [p5], we proposed an alternative approach targeting out-of-domain (OOD) generalization. We applied a similar principle by resampling training mini-batches, pairing each training example with its approximate counterfactual examples. This resampling strategy closely approximates the distribution of a spurious-free randomized trial. **We proved that the Bayes-optimal classifiers trained on this balanced distribution are minimax optimal across different domains.** Our mini-batch sampling method can be seamlessly integrated with other OOD training techniques and **achieved state-of-the-art (SoTA) performance on *DomainBed*** [o3], a widely-used OOD generalization benchmark, at the time of submission.

Another significant limitation of deep learning models, particularly LLMs, lies in their difficulty with structured formal reasoning, such as logical and mathematical reasoning. This challenge is partly due to the lack of enforced reasoning structures during training and inference. While chain-of-thought (CoT) techniques have substantially improved LLMs' multi-step reasoning abilities, the generated reasoning steps often lack coherence and struggle with precise mathematical or logical computations.

In our COLM paper [p6], we proposed a hierarchical generation scheme that incorporates planning tokens—verbalizations of discrete random variables generated prior to each reasoning step. This structured guidance improved reasoning quality across multiple datasets and LLM architectures, particularly for questions requiring long reasoning chains.

In addition to training-time enhancements, we introduced a simple yet effective inference technique called *program-of-thoughts prompting* [p7], which guides LLMs to generate an executable program to solve a given problem. The generated program is then compiled and executed by an external compiler. By generating structured code instead of free-form text, the reasoning coherence and accuracy are greatly enhanced. **We achieved SoTA performance at the time of submission across various reasoning benchmarks requiring mathematical reasoning, with this work garnering over 400 citations within two years.** Our follow-up work, targeting logical reasoning tasks, *Logic-LM* [p8], enables an LLM to generate symbolic formalizations of a problem and then solve these formalizations using an external logic engine. Logic-LM demonstrated substantial improvements over CoT techniques across various logical reasoning benchmarks, **attracting over 100 citations within two years.**

3 Contributing Resources to the LLM/AI Research Community

Beyond my research into the capabilities and limitations of LLMs, I am also committed to contributing valuable resources to the LLM/AI research community. I have actively participated in collaborative efforts to write survey articles on topics such as automated correction for LLMs [p9] and LLM data selection strategies [p10]. Additionally, I have contributed to the development of widely used datasets, including the first time-sensitive question-answering dataset [p11] and the first theorem-driven question-answering dataset [p12]. Through these open-source contributions, **our works have served as foundational resources for many subsequent studies in the field, with more than 300 citations combined.**

4 My Research Vision

In the long term, I am dedicated to building transparent super-intelligence, where every element of its design is fully understood. The three most critical components of this vision, as illustrated in Figure 1, are **theory**—understanding the inner workings of models, **data**—the source of intelligence, and **algorithm**—the methods by which models are trained and utilized. My future research will explore these directions as follows:

- **Theory: Systematic Abstraction of Deep Neural Networks.** While current mechanistic interpretations of LLMs focus on low-level circuits in models like Transformers, such as those handling repetition or grammar, the higher-level mechanisms that govern complex real-world behaviors like in-context learning and reasoning remain elusive. Causal abstraction provides a principled approach to simplifying a complex large causal graph, such as a neural network, into a more comprehensible smaller causal graph [o4, o5]. With my expertise in causality and LLM understanding, I aim to leverage causal abstraction as a systematic framework to elucidate these higher-level mechanisms and to develop a universal abstraction of LLM behaviors.

- **Data: Realistic Synthetic Data Construction.** Controlled experiments on synthetic data are essential for deriving scientific insights into LLMs. However, overly simplistic synthetic data can lead to conclusions that do not generalize well to real-world scenarios. I plan to create controllable, realistic synthetic datasets that emulate real-world data distributions by analyzing and simplifying real-world data. Such synthetic data can also have practical applications: as high-quality real-world text data becomes increasingly scarce, generating synthetic data that accurately reflects real-world distributions could be pivotal for further scaling LLMs. For instance, the random walk augmentation we proposed in [p2] may be effective for scaling up general training data.

- **Algorithm: Cross-Domain, Cross-Modal Training Schemes.** A key aspect of achieving super-intelligence is the seamless integration of language with other modalities and domains. Currently, Transformers excel in language generation tasks, while diffusion models lead in image and video generation. However, existing cross-modal training objectives often struggle to balance language and vision capabilities within a single model. For example, Transformer-based vision-language models frequently lag behind in vision capabilities. My goal is to develop algorithms that unify AI models, enabling them to process and generate data across different domains and modalities seamlessly. As an example, in our NeurIPS paper [p13], we integrated a mixture of differentiable rewards into the consistency distillation process of a pre-trained text-to-video model, improving the quality of generated videos.

My Work

[p1] Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 15614–15638. Curran Associates, Inc., 2023.

[p2] Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, and William Yang Wang. Understanding reasoning ability of language models from the perspective of reasoning paths aggregation. *Forty-first International Conference on Machine Learning*, 2024.

[p3] Xinyi Wang, Antonis Antoniades, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. Generalization vs memorization: Tracing language models' capabilities back to pretraining data. *arXiv preprint arXiv:2407.14985*, 2024.

[p4] Xinyi Wang, Wenhu Chen, Michael Saxon, and William Yang Wang. Counterfactual maximum likelihood estimation for training deep networks. *Advances in Neural Information Processing Systems*, 34:25072–25085, 2021.

[p5] Xinyi Wang, Michael Saxon, Jiachen Li, Hongyang Zhang, Kun Zhang, and William Yang Wang. Causal balancing for domain generalization. In *The Eleventh International Conference on Learning Representations*, 2023.

[p6] Xinyi Wang, Lucas Caccia, Oleksiy Ostapenko, Xingdi Yuan, and Alessandro Sordoni. Guiding language model reasoning with planning tokens. *Conference on Language Modeling*, 2024.

[p7] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023.

[p8] Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[p9] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically Correcting Large Language Models: Surveying the Landscape of Diverse Automated Correction Strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506, 05 2024.

[p10] Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. A survey on data selection for language models. *Transactions on Machine Learning Research*, 2024. Survey Certification.

[p11] Wenhu Chen, Xinyi Wang, and William Yang Wang. A dataset for answering time-sensitive questions. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[p12] Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. Theoremqa: A theorem-driven question answering dataset. In *Proceedings of*

the 2023 Conference on Empirical Methods in Natural Language Processing, pages 7889–7901, 2023.

[p13] Anonymous. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Related Work

[o1] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[o2] Ni Lao, Tom Mitchell, and William W. Cohen. Random walk inference and learning in a large scale knowledge base. In Regina Barzilay and Mark Johnson, editors, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 529–539, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[o3] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

[o4] Atticus Geiger, Hanson Lu, Thomas F. Icard, and Christopher Potts. Causal abstractions of neural networks. In *Neural Information Processing Systems*, 2021.

[o5] Atticus Geiger, Chris Potts, and Thomas Icard. Causal abstraction for faithful model interpretation. *arXiv preprint arXiv:2301.04709*, 2023.