Jiaxin Zhang

AI Senior Staff Research Scientist @ Intuit AI Research

jxzhangai@gmail.com 443-676-0871 Homepage Linkedin Github Google Scholar Bay Area, CA

Highlights

- 6 years research experience on **reliable and robust AI**, 3 years on large language models (LLMs) and generative AI research
- 4 years on building AI research team, managing engineers/scientists, and leading cross-team and cross-org collaborations
- Strong publication records in top AI conferences, e.g., NeurIPS, CVPR, etc and served as Area Chair (AC) of ACL, EMNLP, etc
- **Research interests**: LLM reliability, uncertainty, robustness and safety, alignment, optimization, RAG, and multimodality

Education

Ph.D. **Johns Hopkins University**, Computational Science 2014 – 2018

- Thesis Title: *Uncertainty Quantification Form Small Data: A Multimodel Approach*

M.S. **Johns Hopkins University**, Applied Mathematics & Statistics 2017 – 2018

- Focus: *Machine Learning, Bayesian Statistics, Probabilistic Modeling, Optimization*

Experience

AI Senior Staff Research Scientist, Intuit AI Research 2022 – Now

Leading a team of 5-10 scientists and engineers working on **LLM reliability**, **Optimization**, **Alignment** and **Multimodality**. Responsible for both the technical directions, executions of projects, and tech transfer to business products.

LLM hallucination, uncertainty and reliability

- Reliable hallucination detection via semantic-aware cross-check consistency [SAC3, *EMNLP 23'*], and divide-conquer-reasoning for consistency evaluation and improvement of LLMs [DCR, *EMNLP 24'*]
- Efficient Contextual hallucination mitigation using gradient-guided attention map editing [GAME, NAACL 25']
- Uncertainty quantification and confidence estimation in LLMs [SPUQ, *EACL 24'*]

Prompt optimization and RAG

- Unified in-context prompt optimization [PhaseEvo, *TMLR 24'*] with security and safety contraints [SoS, *EMNLP 24'*]
- Knowledge ingestion and injection via RAG, supervised fine-tuning and continual pre-training [Ski, *EMNLP 24'*]; and characterizing query-knowledge relevance for reliable retrieval augmented generation [OOK, *EMNLP 24'*]

Alignment and reasoning

- Interactive multi-fidelity learning for cost-effective LLM fine-tuning with sparse human supervision [IMFL, *NeurIPS 23'*]
- Rethinking LLMs with gradient-descent feedback alignment; and reasoning-aware contrastive inference-time decoding

Multimodal AI: vision-language models

- Holistic Evaluation for Interleaved Text-and-Image Generation [InterleavedBench, *EMNLP 24'*]
- Modality-Specialized Synergizers for Interleaved Vision-Language Generalists [MoSS, *ICLR 25'*]
- Cycle-consistent diffusion models [DECDM, *WACV 24'*] and robustness of diffusion inversion [LTOP, *ICLR 23' workshop*]

Staff Research Scientist, Computer Science and Mathematics Division, Oak Ridge National Laboratory 2018 – 2022

Led a team of 4-6 research scientists working on **Deep Generative Models**, **Blackbox Optimization**, and **Distributed Deep Learning**. Responsible for driving research directions (PI) and managing people, as well as delivering impact to *AI for Science*.

Deep Generative Models (GANs, Normalizing Flows, Energy-based Models)

- Exploit invertible neural networks (INNs) for nonlinear dimension reduction [NLL, *NeurIPS 19'*] and accelerate inverse-problem solving with INNs [iPage, *AAAI 23'*]; architecture optimization of normalizing flows [AutoNF, *AAAI 23'*];
- Auditing Privacy Defenses via Generative Gradient Leakage leveraging the latent space of GANs. [GGL, *CVPR 22'*]

High-dimensional blackbox optimization

- Novel evolution strategy with directional Gaussian smoothing for blackbox optimization [DGS, *UAI 21'*]

- Scalable gradient free method for Bayesian experimental design with implicit models [SAGABED, *AISTATS 21'*]

Distributed deep learning on ORNL Supercomputers (Summit, Frontier)

- Implemented Gaussian smoothing optimization strategy to accelerate reinforcement learning training on **1000+ CPUs**
- Developed synchronous distributed deep learning through data parallelism, achieving near-linear scaling of distributed training up to **768 NVIDIA V100 GPUs** on Summit supercomputer; and training ImageNet with ResNet-50 in **10 minutes**

Graduate Research Assistant, Johns Hopkins University 2014 – 2018

- Bayesian inference and Monte Carlo methods for uncertainty quantification and propagation [single-author survey]
- Generalization of Latin hypercube sampling for space-filling and variance reduction [LPSS, *RESS 16'*, **460+** citations]

Selected Publications

50+ peer-reviewed publications (30+ first authors), including NeurIPS, CVPR, EMNLP, AAAI, UAI, AISTATS, etc

1. [**EMNLP 24'**] Synthetic Knowledge Ingestion: Towards Knowledge Refinement and Injection for Enhancing Large Language Models. Jiaxin Zhang, Wendi Cui, Yiran Huang, Kamalika Das, and Sricharan Kumar. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2024.*
2. [**EMNLP 24'**] Do You Know What You Are Talking About? Characterizing Query-Knowledge Relevance For Reliable Retrieval Augmented Generation. Zhuohang Li, Jiaxin Zhang, Chao Yan, Kamalika Das, Sricharan Kumar, Murat Kantarcioglu, and Bradley A Malin. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2024.*
3. [**EMNLP 24'**] HyQE: Ranking Contexts with Hypothetical Query Embeddings. Weichao Zhou, Jiaxin Zhang, Hilaf Hasson, Anu Singh, and Wenchao Li. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2024.*
4. [**EMNLP 24'**] Holistic evaluation for interleaved text-and-image generation. Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2024.*
5. [**EMNLP 24'**] Divide-Conquer-Reasoning for Consistency Evaluation and Automatic Improvement of Large Language Models. Wendi Cui, Zhuohang Li, Damien Lopez, Kamalika Das, Bradley A. Malin, Sricharan Kumar, and Jiaxin Zhang. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing - Industry Track, 2024.*
6. [**EMNLP 24'**] Survival of the Safest: Towards Secure Prompt Optimization through Interleaved Multi-Objective Evolution. Ankita Sinha, Wendi Cui, Kamalika Das, and Jiaxin Zhang. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing - Industry Track, 2024.*
7. [**TMLR 24'**] PhaseEvo: Towards Unified In-Context Prompt Optimization for Large Language Models. Wendi Cui, Zhuohang Li, Hao Sun, Damien Lopez, Kamalika Das, Bradley Malin, Sricharan Kumar, and Jiaxin Zhang. *Transactions on Machine Learning Research, 2024 (in review).*
8. [**EACL 24'**] SPUQ: Perturbation-Based Uncertainty Quantification for Large Language Models. Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. *In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, 2024.*
9. [**AISTATS 24'**] Discriminant Distance-Aware Representation on Deterministic Uncertainty Quantification Methods. Jiaxin Zhang, Kamalika Das, and Sricharan Kumar. *In Proceedings of the International Conference on Artificial Intelligence and Statistics, 2024.*
10. [**WACV 24'**] On the Quantification of Image Reconstruction Uncertainty Without Training Data. Jiaxin Zhang, Sirui Bi, and Victor Fung. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024.*
11. [**WACV 24'**] DECDM: Document Enhancement using Cycle-Consistent Diffusion Models. Jiaxin Zhang, Joy Rimchala, Lalla Mouatadid, Kamalika Das, and Sricharan Kumar. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024.*
12. [**NeurIPS 23'**] Interactive Multi-fidelity Learning for Cost-effective Adaptation of Language Model with Sparse Human Supervision. Jiaxin Zhang, Zhuohang Li, Kamalika Das, and Sricharan Kumar. *In Proceedings of the Neural Information Processing Systems, 2023.*
13. [**EMNLP 23'**] SAC3: Reliable Hallucination Detection in Black-Box Language Models via Semantic-aware Cross-check Consistency. Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin and Sricharan Kumar. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2023.*
14. [**AAAI 23'**] Accelerating Inverse Learning via Intelligent Localization with Exploratory Sampling. Jiaxin Zhang, Sirui Bi and Victor Fung. *In Proceedings of the AAAI Conference on Artificial Intelligence, 2023.*

15. [**AAAI 23'**] AutoNF: Automated Architecture Optimization of Normalizing Flows Using a Mixture Distribution Formulation. Yu Wang, Jan Drgona, Jiaxin Zhang, Karthik Somayaji NS, Frank Y Liu, Malachi Schram, and Peng Li. *In Proceedings of the AAAI Conference on Artificial Intelligence, 2023.*
16. [**CVPR 22'**] Auditing Privacy Defenses in Federated Learning via Generative Gradient Leakage. Zhuohang Li, Jiaxin Zhang, Luyang Liu, and Jian Liu. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.*
17. [**NeurIPS 21'**] On the Stochastic Stability of Deep Markov Models. Jan Drgona, Sayak Mukherjee, Jiaxin Zhang, Frank Liu, and Mahantesh Halappanavar. *In Proceedings of the Neural Information Processing Systems, 2021.*
18. [**UAI 21'**] Enabling Long-range Exploration in Minimization of Multimodal Functions. Jiaxin Zhang, Hoang Tran, Dan Lu, and Guannan Zhang. *In Proceedings of Uncertainty in Artificial Intelligence, 2021.*
19. [**AISTATS 21'**] A Scalable Gradient Free Method for Bayesian Experimental Design with Implicit Models. Jiaxin Zhang, Sirui Bi, and Guannan Zhang. *In Proceedings of the International Conference on Artificial Intelligence and Statistics, 2021.*
20. [**NeurIPS 19'**] Learning Nonlinear Level Sets for Dimensionality Reduction in Function Approximation. Guannan Zhang, Jiaxin Zhang, and Jacob Hinkle. *In Proceedings of the Neural Information Processing Systems, 2019.*

Professional Activities and Services

Area Chair and Program Committee

- Area Chair: ACL, EMNLP and NAACL 2024 - now
- Reviewer of AI Conferences: NeurIPS, ICML, ICLR, CVPR, ICCV, ECCV, AAAI, AISTATS, KDD, ICASSP, SIAM SDM 2020 - now
- Reviewer of AI Journals: Journal of Machine Learning Research, Transactions on Machine Learning Research

Workshop Organizer & Chair

- SIAM Uncertainty Quantification - "Robustness, Reliability and Safety in Deep Learning" 2022
- SIAM Mathematical Aspects of Materials Science - "Machine Learning for Solving Inverse Problems" 2021

Research Project Funding

- As a Principal Investigator (PI) or co-PI, leading 7 DOE ASCR/ORNL projects, over **6.4** million in total 2018-2022

Honors and Awards

- **Intuit CTO Award**, Intuit 2024
- **Intuit A2D Innovation Award**, Intuit 2024
- **Promising Early-Career Researcher Award**, Oak Ridge National Laboratory, DOE 2020
- **NeurIPS Travel Award**, NeurIPS 2019
- **Chinese Outstanding Students Abroad Award**, Ministry of Education of the P.R. China 2019
- **Acheson J. Duncan Graduate Research Award**, Johns Hopkins University 2018
- **Dean's Fellowship**, Johns Hopkins University 2014
- **National Scholarship**, Ministry of Education of the P.R. China 2009, 2012

Selected Open-Source Libraries and Softwares

- Intuit AI Research Open Source GitHub Organization - [Project Owner and Founder] 2024 - now
- Ski:Towards Knowledge Refinement and Injection for Enhancing LLMs - [New!] 2024 - now
- Awesome-LLM-Prompt-Optimization - [230+ stars][10+ forks] 2023 - now
- Awesome-LLM-Uncertainty-Reliability-Robustness - [670+ stars][40+ forks] 2023 - now
- Awesome-LLM-Retrieval-Augmented Generation (RAG) - [950+ stars][60+ forks] 2023 - now
- SAC3: Reliable Hallucination Detection in Black-Box LLMs - [30+ stars][5+ forks] 2023 - now
- Divide-Conquer-Reasoning for Consistency Evaluation and Improvement - [20+ stars][3+ forks] 2023 - now
- GGL: Generative Gradient Leakage in Federated Learning - [60+ stars][10+ forks] 2022 - now
- MatDeepLearn: Graph Neural Networks Platform in Materials Chemistry - [170+ stars][40+ forks] 2021 - now
- MatDesINNe: Materials Inverse Design with Invertible Neural Networks - [60+ stars][10+ forks] 2021 - now
- UQpy: Uncertainty Quantification with Python - [270+ stars][80+ forks] 2018 - now