

Yao Qin

Email: yaoqin@ece.ucsb.edu

Education

University of California, San Diego Doctor of Philosophy, Department of Computer Science and Engineering	2015.09 - 2020.01
University of California, San Diego Master of Science, Department of Computer Science and Engineering	2015.09 - 2017.12
Dalian University of Technology Bachelor of Science, Department of Electrical Engineering	2011.09 - 2015.06

Research Experience

Assistant Professor , Department of ECE, University of California, Santa Barbara, USA	2023.01 - present
Research Scientist , Google Research, New York, USA	2020 - 2023

Publications (Note: * below denotes equal contribution)

[Google Scholar](#)

Preprints

13. K. Tang, Y. Li and **Y. Qin**. DIY-MKG: An LLM-Based Polyglot Language Learning System. *Under Review*, 2025.
12. K. Tang*, C. Liu*, **Y. Qin** and Qi Lei. Bridging Distribution Shift and AI Safety: Conceptual and Methodological Synergies. *Under Review*, 2025.
11. A. Hua*, K. Tang*, C. Gu, J. Gu, E. Wong and **Y. Qin**. Artifact or Flaw? Rethinking Prompt Sensitivity in Evaluating LLMs. *Under Review*, 2025.
10. K. Tang, Y. Li and **Y. Qin**. SPICE: A Synergistic, Precise, Iterative, and Customizable Image Editing Workflow. *Under Review*, 2025.
9. C. Gu, A. Hua, J. Gu and **Y. Qin**. Improving Adversarial Transferability in MLLMs via Dynamic Vision-Language Alignment Attack. *Under Review*, 2025.
8. M. Dhaliwal*, K. Tang*, E. M. Aiello, D. P. Zaharieva, R. A. Lal, C. Summers, B. Arbiter, K. Watson, M. J. Connolly, L. E. Figg, I. Balistreri, A. L. Cortes, R. S. Kingman, B. Suh, M. C. Riddell and **Y. Qin**. Variation in Hypoglycemia Risk with Real-World Physical Activity in Adults with Type 1 Diabetes: Insights from the Type 1 Diabetes Exercise Initiative. *Under Review*, 2025.
7. L. Liu, R. Pourreza, S. Panchal, A. Bhattacharyya, **Y. Qin** and R. Memisevic. Enhancing Hallucination Detection through Noise Injection. *Under Review*, 2025.
6. Y. Yoon, D. Hu, I. Weissburg, **Y. Qin** and H. Jeong. Model Collapse in the Self-Consuming Chain of Diffusion Finetuning: A Novel Perspective from Quantitative Trait Modeling. *Under Review*, 2025.

5. A. Balashankar, X. Ma, A. Sinha, A. Beirami, **Y. Qin**, J. Chen and A. Beutel. Improving Few-shot Generalization of Safety Classifiers via Data Augmented Parameter-Efficient Fine-Tuning. *Under Review*, 2025.
4. J. Gu, A. Beirami, X. Wang, A. Beutel, P. Torr and **Y. Qin**. Towards Robustness of In-Context Learning on Vision-language Models. *Under Review*, 2024.
3. J. Gu, Z. Han, S. Chen, A. Beirami, B. He, G. Zhang, R. Liao, **Y. Qin**, V. Tresp and P. Torr. A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models. *Under Review*, 2023.
2. **Y. Qin**, N. Frosst, C. Raffel, G. Cottrell and G. Hinton. Deflecting Adversarial Attacks. *Preprints*, 2019.
1. Ian Goodfellow, **Yao Qin**, David Berthelot. Evaluation Methodology for Attacks Against Confidence Thresholding Models. *Preprints*, 2018.

Conferences & Journals

29. L. Liu and **Y. Qin**. Detecting Out-of-Distribution through the Lens of Neural Collapse. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
28. A. Hua*, M. Dhaliwal*, L. Pallela, R. Burke and **Y. Qin**. NutriBench: A Dataset for Evaluating Large Language Models in Nutrition Estimation from Meal Descriptions (*ICLR*), 2025.
27. E. Aiello, K. Tang, M. Dhaliwal, R. Lal, C. Summers, M. Connolly, D. Zaharieva, B. Arbiter, K. Watson, M. Friedman, L. Figg, A. Cortes-Navarro, I. Balistreri, R. Kingman, B. Suh, M. Riddell and **Y. Qin**. Modeling Metabolic Changes in Glucose Physiology during Physical Activity in T1D. *American Diabetes Association (ADA)*, 2025. ([ADA Early Career Abstract Award](#))
26. E. Aiello, K. Tang, M. Dhaliwal, R. Lal, C. Summers, M. Connolly, D. Zaharieva, B. Arbiter, K. Watson, M. Friedman, L. Figg, A. Cortes-Navarro, I. Balistreri, R. Kingman, B. Suh, M. Riddell and **Y. Qin**. Identifying Insulin and Non-Insulin-Mediated Mechanisms during Physical Activity from Real-World T1D Data. *American Diabetes Association (ADA)*, 2025.
25. K. Tang, P. Song, **Y. Qin**, X. Yan. Creative and Context-Aware Translation of East Asian Idioms with GPT-4. *Findings of Empirical Methods in Natural Language Processing (Findings of EMNLP)*, 2024.
24. L. Liu and **Y. Qin**. Fast Decision Boundary based Out-of-distribution Detection. *International Conference on Machine Learning (ICML)*, 2024.
23. M. Dhaliwal, K. Tang, E. Aiello, D. Zaharieva, R. Lal, C. Summers, B. Arbiter, K. Watson, L. Figg, I. Balistreri, R. Kingman, B. Suh and **Y. Qin**. Understanding Hypoglycemia Risk in Unstructured Real-World Physical Activities in Adults with Type 1 Diabetes. *American Diabetes Association (ADA)*, 2024.
22. M. Dhaliwal, K. Tang, E. Aiello and **Y. Qin**. Glycemic Effect of Free-Living Activities in Adults with Type 1 Diabetes. *American Diabetes Association (ADA)*, 2024.
21. Y. Lal, P. Lahoti, A. Sinha, **Y. Qin**, A. Balashankar. Automated Adversarial Discovery for Safety Classifiers. *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP at NAACL)*, 2024. ([Best Paper Runner up](#))
20. M. Song, X. Wang, T. Biradar, **Y. Qin** and M. Chandraker. A Minimalist Prompt for Zero-Shot Policy Learning. *Task Specification Workshop at The Robotics: Science and Systems (RSS)*, 2024.
19. A. Hua, J. Gu, Z. Xue, N. Carlini, E. Wong and **Y. Qin**. Initialization Matters for Adversarial Transfer Learning. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
18. S. Niazi, N. Aadit, M. Mohseni, S. Chowdhury, **Y. Qin** and K. Camsari. Training Deep Boltzmann Networks with Sparse Ising Machines. *Nature Electronics*, 2024.

17. X. Zhang, S. Li, X. Yang, C. Tian, **Y. Qin** and L. Petzold. Enhancing Small Medical Learners with Privacy-preserving Contextual Prompting. *International Conference on Learning Representations (ICLR)*, 2024.
16. B. Puranik, A. Beirami, **Y. Qin**, U. Madhow. Improving Robustness via Tilted Exponential Layer: A Communication-Theoretic Perspective. *Artificial Intelligence and Statistics (AISTATS)*, 2024.
15. A. Balashankar, X. Wang, **Y. Qin**, N. Thain, B. Packer, E. Chi and A. Beutel. Improving Robustness through Pairwise Generative Counterfactual Data Augmentation. *Findings of Empirical Methods in Natural Language Processing (Findings of EMNLP)*, 2023.
14. Z. Shi, N. Carlini, A. Balashankar, L. Schmidt, C. Hsieh, A. Beutel and **Y. Qin**. Effective Robustness against Natural Distribution Shifts for Models with Different Training Data. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
13. **Y. Qin**, X. Wang, B. Lakshminarayanan, E. Chi, A. Beutel. What are Effective Labels for Augmented Data? Improving Robustness with AutoLabel. *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2023.
12. J. Zhao, X. Wang, **Y. Qin**, J. Chen, K. Chang. Investigating Ensemble Methods for Model Robustness Improvement of Text Classifiers. *Findings of Empirical Methods in Natural Language Processing (Findings of EMNLP)*, 2022.
11. **Y. Qin**, C. Zhang, T. Chen, B. Lakshminarayanan, A. Beutel, X. Wang. Understanding and Improving Robustness of Vision Transformers through Patch-based Negative Augmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
10. J. Gu, V. Tresp, **Y. Qin**. Are Vision Transformers Robust to Patch-wise Perturbations? *European Conference on Computer Vision (ECCV)*, 2022.
9. **Y. Qin**, X. Wang, A. Beutel, E. Chi. Improving Uncertainty Estimates through the Relationship with Adversarial Robustness. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
8. T. Wang, X. Wang, **Y. Qin**, B. Packer, K. Li, J. Chen, A. Beutel, E. Chi. CAT-Gen: Improving Robustness in NLP Models via Controlled Adversarial Text Generation. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.
7. **Y. Qin***, N. Frosst*, S. Sabour, C. Raffel, G. Cottrell and G. Hinton. Detecting and Diagnosing Adversarial Examples with Class-Conditional Capsule Reconstructions. *International Conference on Learning Representations (ICLR)*, 2020.
6. **Y. Qin**, N. Carlini, I. Goodfellow, G. Cottrell and C. Raffel. Imperceptible, Robust and Targeted Adversarial Example for Automatic Speech Recognition. *International Conference on Machine Learning (ICML)*, 2019.
5. **Y. Qin**, S. Ancha, J. Nanavati, G. Cottrell, A. Criminisi and A. Nori. Autofocus Layer for Semantic Segmentation. *International Conference on Medical Image Computing & Computer Assisted Intervention (MICCAI)*, 2018. (**Oral presentation**, 4% acceptance rate)
4. **Y. Qin***, M. Feng*, H. Lu and G. Cottrell. Hierarchical Cellular Automata for Visual Saliency. *International Journal of Computer Vision (IJCV)*, 2017
3. **Y. Qin**, D. Song, H. Chen, W. Cheng, G. Jiang and G. Cottrell. A Dual- Stage Attention-Based Recurrent Neural Network for Time Series Prediction. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017
2. Q. Pan, **Y. Qin**, Y. Xu, M. Tong and M. He. Opinion Evolution in Open Community. *International Journal of Modern Physics C*, 1750003, 2016.
1. **Y. Qin**, H. Lu, Y. Xu and H. Wang. Saliency Detection via Cellular Automata. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015

Patents

1. **Y. Qin**, X. Wang, B. Lakshminarayanan, E. Chi, A. Beutel. What are Effective Labels for Augmented data? Improving Robustness with AutoLabel.
2. D. Song, H. Chen, G. Jiang, **Y. Qin**. Dual Stage Attention based Recurrent Neural Network for Time Series Prediction.

Current Funding

1. Lead PI, Safe Insulin Control for Exercise with Type 1 Diabetes with Activity-specific Presets.
Total Amount: \$2,977,229, **Project Period:** 11/2023 - 10/2026, **Funding Source:** Helmsley Trust.
2. Co-PI, REAL AI Initiative for AI for Science
Total Amount: \$185,000, **Funding Source:** Trustees donation.
3. Co-PI, Neural Collapse in Infrared Search and Track Architectures for Directed Energy Applications.
Total Amount: \$54,000 **Project Period:** 12/2024 - 6/2025, **Funding Source:** Air Force.
4. Co-PI, Toward Out-of-Distribution Aware Time Series Data Mining.
Total Amount: \$35,000 **Project Period:** 19/2024 - 9/2025, **Funding Source:** CAHSI-Google.

Teaching & Mentoring

Instructor

1. ECE180: Introduction to Deep Learning (Spring 2024, 2025), UC Santa Barbara
2. ECE194: Adversarial Robustness in Machine Learning (Winter 2024, 2025), UC Santa Barbara
3. ECE594: Robustness in Machine Learning (Winter 2023, Fall 24, Winter 25), UC Santa Barbara

Teaching Assistant

1. CSE253: Neural Networks for Pattern Recognition (Winter 2019), UC San Diego
2. CSE190: Neural Networks and Deep Learning (Fall 2017), UC San Diego

Student Mentorship

* Current PhD Students

1. Mehak Dhaliwal (PhD at UCSB)
2. Andong Hua (PhD at UCSB)
3. Kenan Tang (PhD at UCSB)
4. Youngseok Yoon (PhD at UCSB)

* Previous Students/Interns

1. Zhouxing Shi (PhD at UCLA → Assistant Prof. at UC Riverside)

2. Jieyu Zhao (PhD at UCLA → Assistant Prof. at USC)
3. Ananth Balashankar (PhD at NYU → Senior Research Scientist at Google)
4. Jindong Gu (PhD at University of Munich → Senior Research Scientist at Google DeepMind)
5. Tianlu Wang (PhD at UVA → Senior Research Scientist at FAIR)

Selected Awards

* Early Career Abstract Award	<i>American Diabetes Association (ADA), 2025</i>
* Best Paper Runner Up	<i>TrustNLP Workshop at NAACL, 2024</i>
* Regents' Junior Faculty Fellowship Award	<i>UC Santa Barbara, 2024</i>
* UCSB Faculty Research Grant Award	<i>UC Santa Barbara, 2023</i>
* Adobe Faculty Research Award	<i>Adobe, 2023</i>
* AI2000 Most Influential Scholar Honorable Mention in AAAI/IJCAI	<i>2022</i>
* Rising Star in EECS	<i>MIT, 2021</i>
* Departmental Fellowship	<i>UC San Diego, 2015</i>
* Outstanding Undergraduate Student Award	<i>Liaoning Province, China, 2015</i>
* HIWIN Elite Scholarship (<i>top 15 students university-wide</i>)	<i>China, 2014</i>
* National Scholarship	<i>China, 2013, 2012</i>

Selected Invited Talks

AI for Diabetes

@ Advanced Technologies & Treatments for Diabetes (ATTD)	<i>2025</i>
@ USC symposium on Frontiers of Machine Learning and AI: Fundamentals and Applications	<i>2025</i>
@ NIH-NIDDK Fifth Artificial Pancreas Workshop	<i>2024</i>
@ Sansum Diabetes Research Institute	<i>2024</i>
@ Endocrine Society AI in Healthcare Summit	<i>2024</i>

AI Safety

@ ICCV Workshop on Safe and Trustworthy Multimodal AI Systems	<i>2025</i>
@ ICCV Workshop on Building Foundation Models You Can Trust	<i>2025</i>
@ UCSB Center of Responsible ML Summit	<i>2023</i>
@ Information Theory and Applications Workshop	<i>2022 & 2023</i>
@ LatinX in AI at NeurIPS	<i>2022 & 2023</i>

@ WiML Un-Workshop at ICML	2022 & 2023 & 2025
@ UCSB, CMU, USC, MPI	2022
@ Google, FAIR, Amazon, Apple	2019
@ Salesforce	2019

Professional Services

Workshop/Summit Organizer

* (Workshop Organizer) AIM-FM: Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond at NeurIPS	2024
* (Workshop Organizer) The 3rd New Frontiers in Adversarial Machine Learning at NeurIPS	2024
* (Summit Organizer) Department of ECE Summit at UCSB	2023
* (Summit Organizer) Responsible Machine Learning Summit at UCSB	2023
* (Workshop Organizer) Robustness of Zero/Few-shot Learning in Foundation Models at NeurIPS	2023
* (Local Arrangement co-Chair) Knowledge Discovery and Data Mining (KDD)	2023
* (Workshop Organizer) Southern California Data Science Day at KDD	2023

Area Chair

* (Area Chair) Advances in Neural Information Processing Systems (NeurIPS)	2025
* (Area Chair) International Conference on Machine Learning (ICML)	2024-2025
* (Area Chair) International Conference on Learning Representations (ICLR)	2023-2025
* (Area Chair) International Conference on Computer Vision (ICCV)	2023, 2025
* (Area Chair) Conference on Computer Vision and Pattern Recognition (CVPR)	2025
* (Senior Program Committee) AAAI Conference on Artificial Intelligence (AAAI)	2025
* (Area Chair) Workshop for Women in Machine Learning (WiML)	2019-2022