Anton Xue

antonxue@seas.upenn.edu
https://antonxue.github.io/

Research Interests

Building interpretable, safe, and efficient ML systems with a flavor of formal methods.

Work Experience

2025 – Present ◇ **The University of Texas at Austin**
Postdoctoral Research Fellow, GenAI Center

Education

2019 – 2025 ◇ **University of Pennsylvania, Ph.D. in Computer Science**
Advisors: Rajeev Alur and Eric Wong
Thesis: Trustworthy Machine Learning: Specification, Verification, and Explanation

2015 – 2019 ◇ **Yale University, B.S. in Computer Science and Mathematics**
Advisor: Ruzica Piskac
Double major, undergraduate research on program analysis and Haskell.

Publications

Preprints

1. Helen Jin, Anton Xue, Weiqiu You, Surbhi Goel, and Eric Wong, "Probabilistic Stability Guarantees for Feature Attributions," *arXiv preprint arXiv:2504.13787*, 2025.
2. Weiqiu You, Anton Xue, Shreya Havaldar, Delip Rao, Helen Jin, Chris Callison-Burch, and Eric Wong, "Probabilistic Soundness Guarantees in LLM Reasoning Chains," *arXiv preprint arXiv:2507.12948*, 2025.
3. Thomas Zhang, Behrad Moniri, Ansh Nagwekar, Faraz Rahman, Anton Xue, Hamed Hassani, and Nikolai Matni, "On the Concurrence of Layer-wise Preconditioning Methods and Provable Feature Learning," *arXiv preprint arXiv:2502.01763*, 2025.
4. Helen Jin, Shreya Havaldar, Chaehyeon Kim, Anton Xue, Weiqiu You, Helen Qu, Marco Gatti, Daniel A Hashimoto, Bhuvnesh Jain, Amin Madani, *et al.*, "The FIX Benchmark: Extracting Features Interpretable to eXperts," *arXiv preprint arXiv:2409.13684*, 2024.

Conferences and Journals

1. Xiayan Ji, Anton Xue, Eric Wong, Oleg Sokolsky, and Insup Lee, "AR-Pro: Anomaly Explanation and Repair with Formal Properties," *Advances in Neural Information Processing Systems*, vol. 37, 2025.
2. Anton Xue, Avishree Khare, Rajeev Alur, Surbhi Goel, and Eric Wong, "Logicbreaks: A Framework for Understanding Subversion of Rule-based Inference," *International Conference on Learning Representations*, 2025.
3. Anton Xue, Rajeev Alur, and Eric Wong, "Stability Guarantees for Feature Attributions with Multiplicative Smoothing," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
4. Anton Xue, Lars Lindemann, and Rajeev Alur, "Chordal Sparsity for SDP-based Neural Network Verification," *Automatica*, vol. 161, p. 111 487, 2024.

5. Chang Zhu, Ziyang Li, Anton Xue, Ati Priya Bajaj, Wil Gibbs, Yibo Liu, Rajeev Alur, Tiffany Bao, Hanjun Dai, Adam Doupé, *et al.*, "{TYGR}: Type Inference on Stripped Binaries using Graph Neural Networks," in *33rd USENIX Security Symposium (USENIX Security 24)*, 2024, pp. 4283–4300.
6. Anton Xue, Lars Lindemann, Alexander Robey, Hamed Hassani, George J Pappas, and Rajeev Alur, "Chordal Sparsity for Lipschitz Constant Estimation of Deep Neural Networks," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, IEEE, 2022, pp. 3389–3396.
7. Rajeev Alur, Phillip Hilliard, Zachary G Ives, Konstantinos Kallas, Konstantinos Mamouras, Filip Niksic, Caleb Stanford, Val Tannen, and Anton Xue, "Synchronization Schemas," in *Proceedings of the 40th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 2021, pp. 1–18.
8. Osbert Bastani, Shuo Li, and Anton Xue, "Safe Reinforcement Learning via Statistical Model Predictive Shielding," in *Robotics: Science and Systems*, 2021.
9. Kedar S Namjoshi and Anton Xue, "A Self-certifying Compilation Framework for Webassembly," in *Verification, Model Checking, and Abstract Interpretation: 22nd International Conference, VMCAI 2021, Copenhagen, Denmark, January 17–19, 2021, Proceedings 22*, Springer, 2021, pp. 127–148.
10. Anton Xue and Nikolai Matni, "Data-driven System Level Synthesis," in *Learning for dynamics and control*, PMLR, 2021, pp. 189–200.
11. William T Hallahan, Anton Xue, Maxwell Troy Bland, Ranjit Jhala, and Ruzica Piskac, "Lazy Counterfactual Symbolic Execution," in *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, 2019, pp. 411–424.
12. William T Hallahan, Anton Xue, and Ruzica Piskac, "G2Q: Haskell Constraint Solving," in *Proceedings of the 12th ACM SIGPLAN International Symposium on Haskell*, 2019, pp. 44–57.

Awards

- 2023 ◇ Amazon Web Services grant to the University of Pennsylvania ASSET fund
- 2019 ◇ University of Pennsylvania ENIAC Fellowship
- 2019 ◇ Yale Computer Science Award
- 2019 ◇ National Science Foundation Graduate Research Fellowship
- 2016 ◇ Yale College Freshman Summer Research Fellowship

Teaching Experience

- Spring 2023 ◇ **Convex Optimization**
  UPenn ESE 605, TA
- 2020 – 2021 ◇ **Fundamentals of Linear Algebra and Optimization**
  UPenn CIS 515, TA, Fall 2020, Spring 2021
- Summer 2020 ◇ **Mathematical Foundations of Computer Science**
  UPenn CIS 160, TA
- Spring 2019 ◇ **Real Analysis**
  Yale MATH 305, Course Grader
- Spring 2018 ◇ **Intensive Algorithms**
  Yale CPSC 366, TA
- Spring 2017 ◇ **Design and Analysis of Algorithms**
  Yale CPSC 365, TA

Teaching Experience (continued)

2016 – 2018 ◇ **Mathematical Tools for Computer Science**
Yale CPSC 202, TA, Fall 2016, Fall 2017, Fall 2018

Projects Supervised

2023 – ◇ **David Zhang** (UPenn Undergraduate)
Topic: Adversarial Robustness of In-context Learning

2023 – 2024 ◇ **Faraz Rahman** (UPenn Undergraduate)
Topic: Explainability and Interpretability of Diffusion Models

2023 ◇ **Jawad Ahmad** (Friends Select School, Philadelphia)
Topic: Programmer-friendly Code Synthesis with Chat-GPT.

Research Community Service

2024 – ◇ **NeurIPS, ICLR, ICML**
Reviewer

2023 – ◇ **SATA, SciForDL, AdvML**
Reviewer

2021 – ◇ **IEEE L-CSS, CDC, ACC, Automatica, L4DC**
Reviewer

2023 ◇ **Tools and Algorithms for the Construction and Analysis of Systems**
Artifact Evaluation Committee

2022 ◇ **Principles of Programming Languages**
Student Volunteer

2021 ◇ **Static Analysis Symposium**
Artifact Evaluation Committee

2021 ◇ **Conference on Verification, Model Checking, and Abstract Interpretation**
Artifact Evaluation Committee

2019 – 2021 ◇ **Programming Language Design and Implementation**
Student Volunteer (2019), Artifact Evaluation Committee (2020, 2021)

2017, 2019 ◇ **Computer Aided Verification**
Student Volunteer

2017 – 2018 ◇ **Yale University Computer Science Department**
Department Student Advisory Committee

Internship Experience

2022 ◇ **SRI International**
Summer Research Intern. Manager: Susmit Jha
Designed and implemented sum-of-squares-based convex optimization methods for verifying neural networks.

2019 ◇ **Nokia Bell Labs**
Summer Research Intern. Manager: Kedar Namjoshi
Developed a self-certifying compiler framework for WebAssembly. Paper accepted at the Conference on Verification, Model Checking, and Abstract Interpretation, 2021.

Internship Experience (continued)

2018 ◇ **Harvard University School of Engineering and Applied Sciences**
Summer Research Intern. Advisor: Stephen Chong
Formalized execution semantics of the R programming language. Workshop submission accepted at Formal Methods in Computer-Aided Design, 2018.

2017 ◇ **Max Planck Institute for Software Systems**
Summer Research Intern. Advisors: Rupak Majumdar and Damien Zufferey
Applied model-checking to 3D printer firmware and contributed to the dReal SMT solver.

2016 – 2019 ◇ **Yale University Computer Science Department**
Undergraduate Research Assistant. Advisor: Ruzica Piskac
Developed a symbolic execution engine for Haskell. Paper submission accepted at Programming Language Design and Implementation, 2019.

2014 – 2015 ◇ **Harvard Medical School**
Software Engineering Intern. Advisor: Nils Gehlenborg
Developed visualization software with D3 and contributed to the refinery platform for web-based data visualization.

2014 ◇ **Vertex Pharmaceuticals**
Summer Software Engineering Intern. Manager: Jason Yuen
Developed an inventory scanner and manager for iOS.

2013 ◇ **Vertex Pharmaceuticals**
Summer Software Engineering Intern. Manager: Jason Yuen
Developed a bug-tracking database using Google Web Toolkit.