Stephan Rabanser

Postdoctoral Research Fellow, Princeton University

rabanser@princeton.edu — rabanser.dev — Last update: May 15, 2026

Research Interests

My research centers on advancing reliable and trustworthy artificial intelligence by bridging the gap between foundational model robustness and the rigorous evaluation of frontier systems. To build models that are aware of their limitations, I focus on methods that allow AI to safely abstain from predictions outside its competence (e.g., uncertainty quantification, selective prediction, and out-of-distribution detection). Complementing this foundational work, I develop realistic evaluation frameworks for advanced AI and autonomous agents (e.g., long-horizon, open-world assessments and credible log-based evaluations) to surface real-world capabilities and critical failure modes that standard benchmark metrics often miss. Ultimately, this dual approach ensures that AI systems are not only robust under shifting conditions but also verifiably safe, credible, and dependable in practical deployments.

Experience

- **Postdoctoral Research Fellow** — Since October 2025
  Princeton University, advised by Prof. Arvind Narayanan & Prof. Matthew Salganik — Princeton, NJ
- **Machine Learning Researcher** — September 2020 – August 2025
  Vector Institute for Artificial Intelligence — Toronto, Canada
- **Student Researcher** — August 2024 – January 2025
  Google Research — Zurich, Switzerland
  - Developed hierarchical selective prediction/rejection techniques for large vision-language models (VLMs).
- **Intern Applied Scientist** — June 2021 – October 2021
  Amazon, AWS AI Labs — Munich, Germany
  - Designed context-invariant time series representations using contrastive and domain-adversarial learning.
- **Intern Applied Scientist** — September 2019 – July 2020
  Amazon, AWS AI Labs — Munich, Germany
  - Systematically assessed the impact of I/O representations for deep-learning-based time-series forecasting.
- **Intern Applied Scientist** — May 2018 – August 2018
  Amazon, AWS AI Labs — Munich, Germany
  - Evaluated existing and developed new ML-based algorithms for large-scale lossless data compression.
  - Implemented autoencoder-based probability distribution estimation for arithmetic coding on tabular data.
- **Intern Software Development Engineer** — August 2017 – October 2017
  Amazon, Core Machine Learning — Berlin, Germany
  - Received an overview of standard time series analysis / forecasting techniques.
  - Implemented Bayes by Backprop (weight uncertainty quantification) for plain MLPs & RNNs in MXNet.

Education

- **PhD in Computer Science** — September 2020 – August 2025
  University of Toronto, advised by Prof. Nicolas Papernot — Toronto, Canada
  - **Supervisory Committee**: Prof. Nicolas Papernot, Prof. Rahul Krishnan, Prof. David Duvenaud, Prof. Roger Grosse, Prof. Zachary Lipton

- **Research Interests**: Machine Learning, Robustness, Safety, Reliability, Uncertainty, Causality, Generative Modeling, Representation Learning, Probabilistic Deep Learning, Anomaly Detection, Distribution Shifts, Out-of-Distribution Sample Detection.

- **Visiting Graduate Student** — June 2023 – September 2023
  University of Cambridge, advised by Prof. David Krueger — Cambridge, UK
- **M.Sc. in Computer Science** — October 2015 – July 2019
  Technical University of Munich (TUM), advised by Prof. Stephan Günnemann — Munich, Germany
- **Visiting Research Scholar** — August 2018 – January 2019
  Carnegie Mellon University (CMU), advised by Prof. Zachary Lipton — Pittsburgh, PA
- **Honours Degree in Technology Management** — August 2015 – June 2017
  Center for Digital Technology and Management (CDTM) — Munich, Germany
- **Visiting Research Student** — February 2016 – June 2016
  Massachusetts Institute of Technology (MIT), advised by Prof. Thomas Malone — Cambridge, MA
- **B.Sc. in Computer Science, Minor in Economic Sciences** — October 2012 – October 2015
  Technical University of Munich (TUM) — Munich, Germany

Awards & Honors

Top Reviewer Award — ICML 2025/2026, NeurIPS 2023, Dist. Shift Workshop @ NeurIPS 2021
Member of the Elite Network of Bavaria — Since 2016
Apple WWDC Student Scholarship — June 2013

Publications

Open-World Evaluations for Measuring Frontier AI Capabilities
Sayash Kapoor, Peter Kirgis, Andrew Schwartz, Stephan Rabanser, JJ Allaire, Rishi Bommasani, Magda Dubois, Gillian Hadfield, Andy Hall, Sara Hooker, Seth Lazar, Steve Newman, Dimitris Papailiopoulos, Shoshannah Tekofsky, Helen Toner, Cozmin Ududec, Arvind Narayanan
2026 — Paper

Log Analysis is Necessary for Credible Evaluation of AI Agents
Peter Kirgis, Sayash Kapoor, Stephan Rabanser, Nitya Nadgir, Cozmin Ududec, Magda Dubois, JJ Allaire, Conrad Stosz, Marius Hobbhahn, Jacob Steinhardt, Arvind Narayanan
arXiv preprint
2026 — Paper

Towards a Science of AI Agent Reliability
Stephan Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, Arvind Narayanan
International Conference on Machine Learning (ICML)
2026 — Paper

What Does It Take to Build a Performant Selective Classifier?
Stephan Rabanser, Nicolas Papernot
Advances in Neural Information Processing Systems (NeurIPS)
2025 — Paper

Cascadia: A Cascade Serving System for Large Language Models
Youhe Jiang, Fangcheng Fu, Wanru Zhao, Stephan Rabanser, Nicholas D. Lane, Binhang Yuan
International Conference on Learning Representations (ICLR)
2026 — Paper

Gatekeeper: Improving Model Cascades Through Confidence Tuning
Stephan Rabanser, Nathalie Rauschmayr, Achin Kulshrestha, Petra Poklukar, Wittawat Jitkrittum, Sean Augenstein, Congchao Wang, Federico Tombari
Advances in Neural Information Processing Systems (NeurIPS) & TTODLer-FM Workshop @ ICML **Best Poster**
2025 — Paper

Suitability Filter: A Statistical Framework for Model Evaluation in Real-World Deployment Settings
Angéline Pouget, Mohammad Yaghini, Stephan Rabanser, Nicolas Papernot
International Conference on Machine Learning (ICML) **Oral**
2025 — Paper

Confidential Guardian: Cryptographically Prohibiting the Abuse of Model Abstention
Stephan Rabanser, Ali Shamsabadi, Olive Franzese, Xiao Wang, Adrian Weller, Nicolas Papernot
International Conference on Machine Learning (ICML)
2025 — Paper

Selective Prediction Via Training Dynamics
Stephan Rabanser, Anvith Thudi, Kimia Hamidieh, Adam Dziedzic, Nicolas Papernot
Transactions on Machine Learning Research (TMLR)
2025 — Paper

Training Private Models That Know What They Don't Know
Stephan Rabanser, Anvith Thudi, Abhradeep Thakurta, Krishnamurthy Dvijotham, Nicolas Papernot
Advances in Neural Information Processing Systems (NeurIPS)
2023 — Paper — Slides

Robust and Actively Secure Collaborative Machine Learning
Nicholas Franzese, Adam Dziedzic, Christopher A. Choquette-Choo, Mark R. Thomas, Muhammad Ahmad Kaleem, Stephan Rabanser, Congyu Fang, Somesh Jha, Nicolas Papernot, Xiao Wang
Advances in Neural Information Processing Systems (NeurIPS)
2023 — Paper

p-DkNN: Out-of-Distribution Detection through Statistical Testing of Deep Representations
Adam Dziedzic, Stephan Rabanser, Mohammad Yaghini, Nicolas Papernot
arXiv preprint
2022 — Paper

Intrinsic Anomaly Detection in Multi-Variate Time Series
Stephan Rabanser, Tim Januschowski, Kashif Rasul, Oliver Borchert, Richard Kurle, Jan Gasthaus, Michael Bohlke-Schneider, Nicolas Papernot, Valentin Flunkert
arXiv preprint
2022 — Paper

The Effectiveness of Discretization in Forecasting: An Empirical Study on Neural Time Series Models
Stephan Rabanser, Tim Januschowski, Valentin Flunkert, David Salinas, Jan Gasthaus
7th KDD Workshop on Mining and Learning from Time Series (MiLeTS) **Oral**
2020 — Paper — Slides

Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift
Stephan Rabanser, Stephan Günnemann, Zachary Lipton
Advances in Neural Information Processing Systems (NeurIPS)
2019 — Paper — Poster — Slides

Introduction to Tensor Decompositions and their Applications in Machine Learning
Stephan Rabanser, Oleksandr Shchur, Stephan Günnemann
arXiv preprint
2017 — Paper

Community Service

Reviewing

Conferences: NeurIPS (2021–2026), ICML (2021, 2022, 2025, 2026), ICLR (2024), IEEE SaTML (2024), AAAI (2020).
Workshops: Distribution Shift @ NeurIPS (2021–2023), Distribution Shift @ ICML (2022), Time Series @ KDD (2022), Time Series @ ICML (2021), Human Evaluation of Generative Models @ NeurIPS (2022).

Invited Talks

UMN CSE Data Science Initiative ML Seminar	Apr 2026
UK AI Security Institute	Apr 2026
Google DeepMind, London	Sep 2023
MIT MIMO Research Forum	Oct 2022
Intel Private AI Institute Fall Workshop	Oct 2022
Microsoft Security Data Science Colloquium	Jul 2021