Prasanna Sattigeri

Principal Research Scientist and Manager, IBM Research · Principal Investigator, MIT-IBM Watson AI Lab

Cambridge, MA · linkedin.com/in/prasannasattigeri · Google Scholar · github.com/pronics2004

SUMMARY

I lead post-training for agent reliability at IBM Research and the team behind Granite Guardian, an open judge model. I set its direction and own its releases, from data to open weights: five versions in eighteen months, each with new capabilities and improvements. The team also adds capabilities to Granite models through modular post-training for reasoning, safety and agentic workloads, such as tool-call checks, groundedness, requirements, factuality and confidence, which Granite Switch composes into one model. Granite Guardian 4.1 ranks fifth of 20 on the Artificial Analysis guardrail benchmark at 59 ms per check. The models run in watsonx.ai, watsonx.governance and Red Hat's Ask Red Hat assistant, with 1.16M downloads on Hugging Face and 371K on Ollama.

I have deep expertise in generative modeling, starting with VAEs and GANs (ICLR 2018, NeurIPS 2017) and now diffusion models, LLMs and agentic systems (COLM 2026, ICLR 2026, ACL 2026), with calibration and uncertainty running through all of it (ICML 2024, NeurIPS 2024). I am a principal investigator on joint projects with MIT and RPI, and a core contributor to AI Fairness 360, AI Explainability 360 and UQ360. The work has 10,043 citations and an h-index of 44 (Google Scholar, October 2026).

EXPERIENCE

IBM Research and MIT-IBM Watson AI Lab

December 2015 – present · Cambridge, MA

Principal Research Scientist and Manager

- Lead the Granite Guardian team. Five releases in 18 months, from 3.0 (October 2024) to 4.1 (April 2026).
- Lead Granite Guardian, IBM's open-source judge and guardrail models, from training data and post-training through evaluation and release. Granite Guardian 4.1 8B (April 2026) covers harm and jailbreaks, RAG groundedness, function-call hallucination and user-written criteria.
- Lead modular post-training for the Granite 4.0 and 4.1 base models, composed into one model by Granite Switch, including its general-purpose Decision adapter, covering policy guardrails, factuality detection and correction, requirement checking, calibrated uncertainty and context attribution.
- Artificial Analysis guardrail benchmark (June 2026): Granite Guardian 4.1 8B scores F1 0.859, fifth of 20 configurations and first among models that take their policy at runtime, at 59 ms average latency.
- GuardBench (April 2025): Granite Guardian models held six of the top ten places across 40 datasets. LLM-AggreFact leaderboard: Granite Guardian 3.3 ranked third of 39 models in 2025, ahead of GPT-4o. Also ranked first on REVEAL (2025).
- Granite Guardian 4.1 as a reward model: 70.29 overall on JETTS best-of-N selection, ahead of every reward model tested, up to 70B parameters.
- Adoption (October 2026): 1.15M all-time Hugging Face downloads across the Guardian models and the two capability libraries, 371K Ollama pulls and 45 community quantized builds. Integrated in Traefik Hub, deepset Haystack and Mozilla.ai any-guardrail.
- Research that shipped: the calibrated-uncertainty capability is trained to reproduce a Thermometer calibrator (ICML 2024), and context attribution to reproduce MExGen attributions (ACL 2025).

- Principal investigator, MIT-IBM Watson AI Lab: “Principles and Methods for Trustworthy Learning With Limited Data” with Prof. Gregory Wornell, and “Human-Centric AI: Novel Algorithms for Shared Decision Making” with Prof. David Sontag. Principal investigator on joint work with RPI on causal inference and agent safety, with Prof. Ali Tajer and Prof. Mohammad Mohammadi Amiri.
- Research on calibration and uncertainty quantification for foundation models (Thermometer, ICML 2024; NeurIPS 2024), abstention (ACL 2026) and guardrails for agentic systems (ICLR 2026).
- Core contributor to the open-source toolkits AI Fairness 360 (2,878 GitHub stars; paper cited 2,937 times), AI Explainability 360 (1,809 stars), Uncertainty Quantification 360 and ICX360.
- Generative modeling: DIP-VAE for disentangled representation learning (ICLR 2018), implemented in Google’s disentanglement_lib and in PyTorch-VAE; semi-supervised learning with GANs (NeurIPS 2017); Fairness GAN; wavelet-domain image synthesis (not-so-BigGAN).
- Earlier work: co-regularized alignment for unsupervised domain adaptation (NeurIPS 2018) and domain shift in clinical models (Best Paper, KDD healthcare workshop, 2019).

Yelp Inc.

December 2014 – December 2015 · San Francisco, CA

Software Engineer

- Built machine learning models for products used by a large consumer base.
- Applied natural language processing and image understanding to improve search results.
- Prototyped systems for measuring deployed data science algorithms.

Arizona State University, School of ECEE

January 2009 – December 2014 · Tempe, AZ

Research Assistant; Teaching Assistant (2011–2013)

- Developed methods for learning hierarchical representations from low-rank and sparse structure, and ensemble sparse models for image classification and retrieval.
- Lead student researcher on an NSF project on ion-channel sensors.
- Teaching assistant for Signals and Systems, about 400 students.

EDUCATION

2014 Ph.D., Electrical Engineering, Arizona State University

2008 B.Tech., Electronics Engineering, National Institute of Technology, India

HONORS AND AWARDS

2025 IBM Research Outstanding Accomplishment: AI Governance Innovations to watsonx.governance and Open Source

2025 IBM Research Technical Achievement Award: Granite Guardian

2024 IBM Research Outstanding Accomplishment: Granite 3.0 Release

2024 IBM Research Technical Achievement Award: Causal Inference for Business Decision Making

2021 IBM Research Technical Achievement Award: Science of Uncertainty Quantification

2021 IBM Research Technical Achievement Award: Science of Accurate, Robust, and Generalizable AI

2020 IBM Research Technical Achievement Award: Science in Learning with Less Labels

2020 IBM Research Technical Achievement Award: Dynamic Neural Networks for Efficient AI

2020 Harvard Belfer Center Tech Spotlight, runner-up, for AI Fairness 360

2019 IBM Outstanding Technical Achievement Award: Contributions to Trustworthy AI

2019 Best Paper, KDD Applied Data Science for Healthcare Workshop

2014 University Graduate Fellowship, Arizona State University

PROFESSIONAL SERVICE

- Associate Editor: Pattern Recognition; Digital Signal Processing
- Area chair and senior program committee: NeurIPS, ICML, ICLR, AAAI
- Organizer: Workshop on Practical Bayesian Methods for Big Data, AI Research Week, MIT (2019); AAAI Fall Symposium on Gathering for Artificial Intelligence and Natural Systems (2018)
- Reviewer: NeurIPS, ICML, ICLR, AAAI, ACL, EMNLP, IEEE TPAMI, IEEE Transactions on Signal Processing
- IEEE Senior Member

MENTORING

- Thesis committee member for Joshua Lee and Maohao Shen, PhD students of Gregory Wornell at MIT, and for Burak Varıcı at RPI.
- Research interns hosted throughout ten years at IBM Research.

ACADEMIC COLLABORATIONS

MIT	Gregory Wornell, David Sontag, William Freeman, Aude Oliva; with Maohao Shen, Joshua Lee, Abhin Shah, J. Jon Ryu, Hussein Mozannar, Hunter Lang, Abinitha Gourabathina
Rensselaer Polytechnic Institute	Ali Tajer, Mohammad Mohammadi Amiri; with Zirui Yan, Burak Varıcı, Sadia Asif
Stanford University	Tatsunori Hashimoto; with Mingjian Jiang, Yangjun Ruan, Zexue He
Mila and McGill University	David Rolnick; with Paula Harder, Qidong Yang, Alex Hernandez-Garcia, Venkatesh Ramesh

PUBLICATIONS

Selected publications. The full list is on Google Scholar.

2026

M. Shen, K. Zha, Z. He, Z.-W. Hong, S. Ouyang, J. J. Ryu, **P. Sattigeri**, S. Diggavi, G. Wornell. "Decocted Experience Improves Test-Time Inference in LLM Agents." *COLM 2026*

H. Chen, M. Shen, S. Ouyang, D. Zhang, Z. He, Y. Wang, Y. Xiao, X. Zhang, I. Padhi, S. Chaudhury, G. W. Wornell, **P. Sattigeri**, Z.-W. Hong, K. Youcef-Toumi. "How Should Agents Train Online from Deployment Experience? Fast and Slow Online Learning." *NeurIPS Workshop 2026* **TTCL and LCFM workshops.**

Y. Huang, H. Hua, Y. Zhou, P. Jing, M. Nagireddy, I. Padhi, G. Dolcetti, Z. Xu, et al., **P. Sattigeri**, X. Zhang. "Building a Foundational Guardrail for General Agentic Systems via Synthetic Data." *ICLR 2026* **Co-corresponding author.**

A. Gourabathina, I. Padhi, M. Nagireddy, S. Chaudhury, **P. Sattigeri**. "Answering the Wrong Question: Reasoning Trace Inversion for Abstention in LLMs." *ACL 2026* **Senior author.**

Z. Yan, D. Wei, D. A. Katz, **P. Sattigeri**, A. Tajer. "Multi-component Causal Tracing in Large Language Models." *ACL 2026*

G. Cornacchia, I. Padhi, M. Nagireddy, S. Chaudhury, P. Dognin, T. Pedapati, A. Rawat, M. Purcell, K. R. Varshney, **P. Sattigeri**. “Calibrated Deferral Routing for Cost-Efficient Guardrails.” *ICML Workshop 2026* **Agents in the Wild workshop.**

Y. Huang, C. Gao, S. Wu, H. Wang, X. Wang, et al., incl. **P. Sattigeri**. “TrustGen: A Platform of Dynamic Benchmarking on the Trustworthiness of Generative Foundation Models.” *ICLR 2026*

S. Asif, M. M. Amiri, M. Abbas, T. Pedapati, **P. Sattigeri**. “BLINDSPOT: A Benchmark for Safety and Refusal Calibration in Long-Horizon Tool-Using Agents.” *arXiv 2026* **Senior author.**

S. Asif, M. M. Amiri, M. Abbas, **P. Sattigeri**, K. N. Ramamurthy. “LCGuard: Latent Communication Guard for Safe KV Sharing in Multi-Agent Systems.” *arXiv 2026*

2025

I. Padhi, M. Nagireddy, G. Cornacchia, S. Chaudhury, T. Pedapati, P. Dognin, K. Murugesan, E. Miehling, et al., **P. Sattigeri**. “Granite Guardian: Comprehensive LLM Safeguarding.” *NAACL (Industry) 2025* **Senior author.**

L. M. Paes, D. Wei, H. J. Do, H. Strobelt, R. Luss, A. Dhurandhar, M. Nagireddy, K. N. Ramamurthy, **P. Sattigeri**, W. Geyer, S. Ghosh. “Multi-Level Explanations for Generative Language Models.” *ACL 2025* **Oral.**

E. Miehling, M. Desmond, K. N. Ramamurthy, E. M. Daly, K. R. Varshney, et al., **P. Sattigeri**. “Evaluating the Prompt Steerability of Large Language Models.” *NAACL 2025*

M. Nagireddy, I. Padhi, S. Ghosh, **P. Sattigeri**. “When in Doubt, Cascade: Towards Building Efficient and Capable Guardrails.” *AIES 2025* **Senior author.**

P. Dognin, J. Rios, R. Luss, **P. Sattigeri**, M. Liu, I. Padhi, M. Riemer, et al. “Contextual Value Alignment.” *ICASSP 2025*

E. Miehling, K. N. Ramamurthy, K. R. Varshney, M. Riemer, D. Bouneffouf, et al., incl. **P. Sattigeri**. “Agentic AI Needs a Systems Theory.” *arXiv 2025*

G. Zizzo, G. Cornacchia, K. Fraser, M. Z. Hameed, A. Rawat, B. Buesser, et al., incl. **P. Sattigeri**. “Adversarial Prompt Evaluation: Systematic Benchmarking of Guardrails Against Prompt Input Attacks on LLMs.” *arXiv 2025*

2024

M. Shen, S. Das, K. Greenewald, **P. Sattigeri**, G. Wornell, S. Ghosh. “Thermometer: Towards Universal Calibration for Large Language Models.” *ICML 2024*

M. Jiang, Y. Ruan, **P. Sattigeri**, S. Roukos, T. Hashimoto. “Graph-based Uncertainty Metrics for Long-form Language Model Generations.” *NeurIPS 2024*

M. Shen, J. J. Ryu, S. Ghosh, Y. Bu, **P. Sattigeri**, S. Das, G. W. Wornell. “Are Uncertainty Quantification Capabilities of Evidential Deep Learning a Mirage?” *NeurIPS 2024*

Y. Hou, A. Pascale, J. Carnerero-Cano, T. Tchrakian, R. Marinescu, E. Daly, I. Padhi, **P. Sattigeri**. “WikiContradict: A Benchmark for Evaluating LLMs on Real-World Knowledge Conflicts from Wikipedia.” *NeurIPS (Datasets & Benchmarks) 2024*

B. Varıcı, D. A. Katz, D. Wei, **P. Sattigeri**, A. Tajer. “Interventional Causal Discovery in a Mixture of DAGs.” *NeurIPS 2024*

E. Miehling, M. Nagireddy, **P. Sattigeri**, E. M. Daly, D. Piorkowski, J. T. Richards. “Language Models in Dialogue: Conversational Maxims for Human-AI Interactions.” *EMNLP (Findings) 2024*

I. Padhi, K. N. Ramamurthy, **P. Sattigeri**, M. Nagireddy, P. Dognin, K. R. Varshney. “Value Alignment from Unstructured Text.” *EMNLP (Industry) 2024*

Z. Yan, D. Wei, D. A. Katz, **P. Sattigeri**, A. Tajer. “Causal Bandits with General Causal Models and Interventions.” *AISTATS 2024*

I. Padhi, P. L. Dognin, J. Rios, R. Luss, S. Achintalwar, M. Riemer, M. Liu, et al., incl. **P. Sattigeri**. “ComVas: Contextual Moral Values Alignment System.” *IJCAI 2024*

T. Pedapati, A. Dhurandhar, S. Ghosh, S. Dan, **P. Sattigeri**. “Large Language Model Confidence Estimation via Black-Box Access.” *arXiv 2024*

H. Mozannar, V. Chen, M. Alsobay, S. Das, S. Zhao, D. Wei, M. Nagireddy, et al., incl. **P. Sattigeri**. “The RealHumanEval: Evaluating Large Language Models’ Abilities to Support Programmers.” *arXiv 2024*

S. Achintalwar, I. Baldini, D. Bouneffouf, et al., incl. **P. Sattigeri**. “Alignment Studio: Aligning Large Language Models to Particular Contextual Regulations.” *IEEE Internet Computing 2024*

Q. Yang, A. Hernandez-Garcia, P. Harder, V. Ramesh, **P. Sattigeri**, et al. “Fourier Neural Operators for Arbitrary Resolution Climate Data Downscaling.” *JMLR 2024*

B. Varıcı, D. Katz, D. Wei, **P. Sattigeri**, A. Tajer. “Separability Analysis for Causal Discovery in Mixture of DAGs.” *TMLR 2024*

2023

H. Mozannar, J. J. Lee, D. Wei, **P. Sattigeri**, S. Das, D. Sontag. “Effective Human-AI Teams via Learned Natural Language Rules and Onboarding.” *NeurIPS 2023* **Spotlight.**

H. Mozannar, H. Lang, D. Wei, **P. Sattigeri**, S. Das, D. Sontag. “Who Should Predict? Exact Algorithms for Learning to Defer to Humans.” *AISTATS 2023*

S. Basu, P. Katdare, **P. Sattigeri**, V. Chenthamarakshan, K. Driggs-Campbell, P. Das, L. R. Varshney. “Efficient Equivariant Transfer Learning from Pretrained Models.” *NeurIPS 2023*

M. Shen, Y. Bu, **P. Sattigeri**, S. Ghosh, S. Das, G. Wornell. “Post-hoc Uncertainty Learning Using a Dirichlet Meta-Model.” *AAAI 2023*

S. Basu, **P. Sattigeri**, K. N. Ramamurthy, V. Chenthamarakshan, K. R. Varshney, et al. “Equi-Tuning: Group Equivariant Fine-Tuning of Pretrained Models.” *AAAI 2023*

M. Shen, S. Ghosh, **P. Sattigeri**, S. Das, Y. Bu, G. Wornell. “Reliable Gradient-free and Likelihood-free Prompt Tuning.” *EACL (Findings) 2023*

B. Richardson, **P. Sattigeri**, D. Wei, K. N. Ramamurthy, K. Varshney, et al. “Add-Remove-or-Relabel: Practitioner-Friendly Bias Mitigation via Influential Fairness.” *FAccT 2023*

P. Harder, A. Hernandez-Garcia, V. Ramesh, Q. Yang, **P. Sattigeri**, et al. “Hard-Constrained Deep Learning for Climate Downscaling.” *JMLR 2023*

B. Varici, K. Shanmugam, **P. Sattigeri**, A. Tajer. “Causal Bandits for Linear Structural Equation Models.” *JMLR 2023*

K. I. Gero, P. Das, P. Dognin, I. Padhi, **P. Sattigeri**, K. R. Varshney. “The Incentive Gap in Data Work in the Era of Large Models.” *Nature Machine Intelligence 2023*

2022

P. Sattigeri, S. Ghosh, I. Padhi, P. Dognin, K. R. Varshney. "Fair Infinitesimal Jackknife: Mitigating the Influence of Biased Training Data Points Without Refitting." *NeurIPS 2022* **First author.**

A. Shah, Y. Bu, J. K. Lee, S. Das, R. Panda, **P. Sattigeri**, G. W. Wornell. "Selective Regression Under Fairness Criteria." *ICML 2022*

B. Varici, K. Shanmugam, **P. Sattigeri**, A. Tajer. "Intervention Target Estimation in the Presence of Latent Variables." *UAI 2022*

S. Galhotra, K. Shanmugam, **P. Sattigeri**, K. R. Varshney. "Causal Feature Selection for Algorithmic Fairness." *SIGMOD 2022*

V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, et al., incl. **P. Sattigeri**. "AI Explainability 360: Impact and Design." *AAAI 2022*

2021

U. Bhatt, J. Antorán, Y. Zhang, Q. V. Liao, **P. Sattigeri**, R. Fogliato, et al. "Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty." *AIES 2021*

J. K. Lee, Y. Bu, D. Rajan, **P. Sattigeri**, R. Panda, S. Das, G. W. Wornell. "Fair Selective Classification via Sufficiency." *ICML 2021*

B. Varici, K. Shanmugam, **P. Sattigeri**, A. Tajer. "Scalable Intervention Target Estimation in Linear Models." *NeurIPS 2021*

Y. Meng, R. Panda, C.-C. Lin, **P. Sattigeri**, L. Karlinsky, K. Saenko, A. Oliva, R. Feris. "AdaFuse: Adaptive Temporal Fusion Network for Efficient Action Recognition." *ICLR 2021*

R. Luss, P.-Y. Chen, A. Dhurandhar, **P. Sattigeri**, Y. Zhang, K. Shanmugam, et al. "Leveraging Latent Features for Local Explanations." *KDD 2021*

K. Ahuja, **P. Sattigeri**, K. Shanmugam, D. Wei, K. N. Ramamurthy, M. Kocaoglu. "Conditionally Independent Data Generation." *UAI 2021*

2020

S. Han, A. Srivastava, C. Hurwitz, **P. Sattigeri**, D. D. Cox. "not-so-BigGAN: Generating High-Fidelity Images on Small Compute with Wavelet-based Super-Resolution." *arXiv 2020*

N. Tatro, P.-Y. Chen, P. Das, I. Melnyk, **P. Sattigeri**, R. Lai. "Optimizing Mode Connectivity via Neuron Alignment." *NeurIPS 2020*

Y. Meng, C.-C. Lin, R. Panda, **P. Sattigeri**, L. Karlinsky, A. Oliva, K. Saenko, R. Feris. "AR-Net: Adaptive Frame Resolution for Efficient Action Recognition." *ECCV 2020*

M. Lichtenstein, **P. Sattigeri**, R. Feris, R. Giryes, L. Karlinsky. "TAFSSL: Task-Adaptive Feature Sub-Space Learning for Few-Shot Classification." *ECCV 2020*

N. M. Kinyanjui, T. Odonga, C. Cintas, N. C. F. Codella, R. Panda, **P. Sattigeri**, K. R. Varshney. "Fairness of Classifiers Across Skin Tones in Dermatology." *MICCAI 2020*

J. J. Thiagarajan, B. Venkatesh, **P. Sattigeri**, P.-T. Bremer. "Building Calibrated Deep Models via Uncertainty Matching with Auxiliary Interval Predictors." *AAAI 2020*

2019

J. Lee, **P. Sattigeri**, G. Wornell. "Learning New Tricks From Old Dogs: Multi-Source Transfer Learning From Pre-Trained Networks." *NeurIPS 2019*

R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, et al., incl. **P. Sattigeri**. "AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias." *IBM Journal of Research and Development 2019*

P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, K. R. Varshney. "Fairness GAN: Generating Datasets with Fairness Properties Using a Generative Adversarial Network." *IBM Journal of Research and Development 2019* **First author.**

J. J. Thiagarajan, D. Rajan, **P. Sattigeri**. "Understanding Behavior of Clinical Models under Domain Shifts." *KDD Applied Data Science for Healthcare Workshop 2019* **Best Paper.**

2018

A. Kumar, **P. Sattigeri**, A. Balakrishnan. "Variational Inference of Disentangled Latent Concepts from Unlabeled Observations." *ICLR 2018*

A. Kumar, **P. Sattigeri**, K. Wadhawan, L. Karlinsky, R. Feris, W. T. Freeman, G. Wornell. "Co-regularized Alignment for Unsupervised Domain Adaptation." *NeurIPS 2018*

2017

A. Kumar, **P. Sattigeri**, P. T. Fletcher. "Semi-supervised Learning with GANs: Manifold Invariance with Improved Inference." *NeurIPS 2017*

PATENTS

Granted

US 12,725,081	Causal learning model (2026)
US 12,724,965	Reliable gradient-free and likelihood-free prompt tuning (2026)
US 12,639,623	Fair selective classification via a variational mutual information upper bound for imposing sufficiency (2026)
US 12,561,612	Mitigating the influence of biased training instances without refitting (2026)
US 11,514,318	Multi-source transfer learning from pre-trained networks (2022)
US 11,222,242	Contrastive explanations for images with monotonic attribute functions (2022)
US 9,875,428	Ensemble sparse models for image analysis and restoration (2018)

Applications

App. 18/925,251	Contextual moral value alignment through context-based aggregation
App. 18/595,630	Context large language model output explanation
App. 18/538,726	Post-hoc uncertainty quantification for machine learning systems
App. 18/503,166	Sensitive attribute driven predictive modeling
App. 17/368,925	Conditionally independent data generation for training machine learning systems
App. 16/926,407	Efficient search of robust accurate neural networks

INVITED TALKS, TUTORIALS AND PANELS

2024 MIT AI Conference: AI Ethics and Change Management

2024 NAACL TrustNLP Workshop: LLM Governance and Alignment
2024 National Academy of Sciences: Reliable AI-assisted Decision Making
2024 CHI TREW Workshop: Trust and Reliance in Evolving Human-AI Workflows
2023 KDD Workshop on Uncertainty Reasoning: Uncertainty Calibration and AI-assisted Decision Making
2023 DSHealth Workshop, KDD: Generative AI and Safety (panel)
2023 AI for Open Society Day, KDD: Trustworthy LLMs (panel)
2021 KDD Responsible AI Workshop: Trustworthy AI Toolkits
2021 PyData Global: AI Uncertainty Quantification (tutorial)
2021 ACM CODS-COMAD: UQ360 (hands-on tutorial)
2021 ARC Training Centre, Australia: AI Uncertainty Quantification
2019 AI Research Week, MIT: Trusted AI (talk and panel)
2019 Harvard ComputeFest: AI Fairness (talk and tutorial)
2018 AAAI Fall Symposium: AI and Trust (talk and panel)

SELECTED COVERAGE

- Artificial Analysis, “Benchmarking guardrail models for safety, refusal, and latency” (June 2026)
- IBM Research Blog, “Granite Guardian tops third-party AI benchmark” (April 2025)
- IBM Research Blog, “IBM Granite now has adapters designed to control AI outputs” (March 2025)
- IBM Research Blog, “How we slimmed down Granite Guardian” (February 2025)
- MIT IDE, “What Does Ethical AI Look Like?” (2024)
- VentureBeat, “What to do when AI brings more questions than answers?” (July 2021)
- SiliconANGLE, “IBM is teaching AI models to express how confident they are about their decisions” (June 2021)
- BBC News, “IBM launches tool aimed at detecting AI bias” (September 2018)