New York City, USA
izmailovpavel.github.io
izmailovpavel@gmail.com
Google Scholar Profile
GitHub

Pavel Izmailov

Appointments

Starting Fall 2025 **Assistant Professor, New York University**
Tandon, Computer Science and Engineering department
Courant Institute, Computer Science department (Affiliated)

2024–Present **Research Scientist, Anthropic**
Reasoning in AI; Scalable Oversight
Contributed to the Claude 3.7 and Claude 4 models

2023–2024 **Research Scientist, OpenAI**
Reasoning in AI; Superalignment
Contributed to the o1 model

Education

2019–2023 **Ph.D, Courant Institute, New York University**
Computer Science department
- GPA: 4/4
- Supervisor: Andrew Gordon Wilson
- Thesis: Deconstructing Models and Methods in Deep Learning

2017–2019 **MSc, Cornell University**
Operations Research and Information Engineering department
- GPA: 3.9/4
- Supervisor: Andrew Gordon Wilson
- Obtained MSc as a Ph.D student and transferred to NYU

2013–2017 **BSc, Lomonosov Moscow State University**
Faculty of Computational Mathematics and Cybernetics
- GPA: 4.75/5
- Undergraduate student researcher in the Bayesian Methods Research Group supervised by D. Vetrov and D. Kropotov

Awards

2022 **ICML Outstanding Paper Award**

2021 **Harold Grad Memorial Prize**
NYU Courant Institute prize for outstanding performance and promise as a graduate student

2019 NYU MacCracken PhD Fellowship

2018-2022 **Outstanding Reviewer Awards**
UAI 2022, ICML 2022, NeurIPS 2019, NeurIPS 2018

2018-2022 Travel Awards
NeurIPS 2022 (NeurIPS Scholar Award), NeurIPS 2018, UAI 2018, AISTATS 2018

2017-2018 Cornell University PhD Fellowship

2017 **Best Undergraduate Thesis Award, Moscow State University CS Department**

Publications

* Equal Contribution

2026 Understanding Reasoning from Pretraining to Post-Training
J. Shen, A. Li, S. Rahman, Y. Sun, M. Goldblum, M. Telgarsky, **P. Izmailov**
arxiv

2026 Emergent Capabilities Arise Randomly from Learning Sparse Attention Patterns
V. Baherwani, Z. Chen, S. Qiu, A. G. Wilson, **P. Izmailov**
arxiv

2026 End-to-End Context Compression at Scale
A. Li, S. McLeish, H. Chen, N. Kalra, Z. Chen, A. Gazizov, V. A. S. K. Morisetty, B. Kailkhura, H. Menon, Z. Liu, B. R. Bartoldson, T. Goldstein, S. Lotfi, M. Goldblum, **P. Izmailov**
arxiv

2026 How's it going? Reinforcement learning in language models recruits a functional welfare axis
A. Q. Han, D. J. Chalmers, **P. Izmailov**
arxiv

2026 DiscoverPhysics: Benchmarking LLMs for Out-of-the-Box Scientific Thinking
M. L. Wiemann, L. M. Smith, P. Melchior, S. Mishra-Sharma, A. G. Wilson, **P. Izmailov**, C. Cuesta-Lázaro
arxiv

2026 Forgetting in Language Models: Capacity, Optimization, and Self-Generated Replay
M. Marek, D. Cho, S. Qiu, R. Chunara, **P. Izmailov**, A. G. Wilson
arxiv

2026 Lifting Embodied World Models for Planning and Control
European Conference on Computer Vision (ECCV)
A. N. Wang, T. Darrell, **P. Izmailov**, Y. Bai, A. Bar
arxiv

2026 When Can LLMs Learn to Reason with Weak Supervision?
S. Rahman, J. Shen, A. Mordvina, H. Palangi, S. Gabriel, **P. Izmailov**
arxiv

2026 From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence
M. Finzi, S. Qiu, Y. Jiang, **P. Izmailov**, J. Z. Kolter, A. G. Wilson
arxiv

2025 Closing the Train-Test Gap in World Models for Gradient-Based Planning
A. Parthasarathy, N. Kalra, R. Agrawal, Y. LeCun, O. Bounou, **P. Izmailov**, M. Goldblum
arxiv

2024 OpenAI o1 System Card
OpenAI
arxiv

2024 Learning to Reason with LLMs
OpenAI (contributor)
Technical Post

2024 Can a Confident Prior Replace a Cold Posterior?
M. Marek, B. Paige, **P. Izmailov**
arxiv

2023 Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision
International Conference on Machine Learning (ICML)
C. Burns, **P. Izmailov**, J. H. Kirchner, B. Baker, L. Gao, L. Aschenbrenner, Y. Chen, A. Ecoffet, M. Joglekar, J. Leike, I. Sutskever, J. Wu
arxiv

2023 Simple and Fast Group Robustness by Automatic Feature Reweighting
International Conference on Machine Learning (ICML)
S. Qiu, A. Potapczynski, **P. Izmailov**, A. G. Wilson
arxiv

2023 FlexiViT: one model for all patch sizes
Conference on Computer Vision and Pattern Recognition (CVPR)
L. Beyer, **P. Izmailov**, A. Kolesnikov, M. Caron, S. Kornblith, X. Zhai, M. Minderer, M. Tschannen, I. Alabdulmohsin, F. Pavetic
arxiv

2023 Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations
International Conference on Learning Representations (ICLR)
Spotlight Presentation, 8% accept rate
P. Kirichenko*, **P. Izmailov***, A. G. Wilson
arxiv

2022 On Feature Learning in the Presence of Spurious Correlations
Neural Information Processing Systems (NeurIPS)
P. Izmailov*, P. Kirichenko*, N. Gruver*, A. G. Wilson
arxiv

2022 On Uncertainty, Tempering, and Data Augmentation in Bayesian Classification
Neural Information Processing Systems (NeurIPS)
S. Kapoor*, Wesley J. Maddox*, **P. Izmailov***, A. G. Wilson
arxiv

2022 Bayesian Model Selection, the Marginal Likelihood, and Generalization
International Conference on Machine Learning (ICML), ***Long Talk (Oral)***
Outstanding Paper Award
S. Lotfi, **P. Izmailov**, G. Benton, M. Goldblum, A. G. Wilson
arxiv

2022 Unsupervised learning of two-component nematicity from STM data on magic angle bilayer graphene
ArXiv preprint
W. Taranto, S. Lederer, Y. Choi, **P. Izmailov**, A. G. Wilson, S. Nadj-Perge, E. Kim
arxiv

2021 Dangers of Bayesian Model Averaging under Covariate Shift
Neural Information Processing Systems (NeurIPS)
P. Izmailov, P. Nicholson, S. Lotfi, A. G. Wilson
arxiv

2021 Does Knowledge Distillation Really Work?
Neural Information Processing Systems (NeurIPS)
S. Stanton, **P. Izmailov**, P. Kirichenko, A. A. Alemi, A. G. Wilson
arxiv

2021 What Are Bayesian Neural Network Posteriors Really Like?
International Conference on Machine Learning (ICML), ***Long Talk (Oral), 3% accept rate***
P. Izmailov, S. Vikram, M. D. Hoffman, A. G. Wilson
arxiv

2020 Learning Invariances in Neural Networks from Training Data
Neural Information Processing Systems (NeurIPS)
G. Benton, M. Finzi, **P. Izmailov**, A. G. Wilson
arxiv

2020 Why Normalizing Flows Fail to Detect Out-of-Distribution Data
Neural Information Processing Systems (NeurIPS)
P. Kirichenko*, **P. Izmailov***, A. G. Wilson
arxiv

2020 Bayesian Deep Learning and a Probabilistic Perspective of Generalization
Neural Information Processing Systems (NeurIPS)
A. G. Wilson, **P. Izmailov**
arxiv

2020 Generalizing Convolutional Neural Networks for Equivariance to Lie Groups on Arbitrary Continuous Data
International Conference on Machine Learning (ICML)
M. Finzi, S. Stanton, **P. Izmailov**, A. G. Wilson A. G. Wilson, **P. Izmailov**
arxiv

2020 Semi-Supervised Learning with Normalizing Flows
International Conference on Machine Learning (ICML)
P. Izmailov*, P. Kirichenko*, M. Finzi*, A. G. Wilson
arxiv

2019 A Simple Baseline for Bayesian Uncertainty in Deep Learning
Neural Information Processing Systems (NeurIPS)
W. Maddox*, T. Garipov*, **P. Izmailov***, D. Vetrov, A. G. Wilson
arxiv

2019 Subspace Inference for Bayesian Deep Learning
Uncertainty in Artificial Intelligence (UAI)
P. Izmailov*, W. Maddox*, P. Kirichenko*, T. Garipov*, D. Vetrov, A. G. Wilson
arxiv

2019 There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average
International Conference on Learning Representations (ICLR)
B. Athiwaratkun, M. Finzi, **P. Izmailov**, A. G. Wilson
OpenReview

2018 Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs
Neural Information Processing Systems (NeurIPS),
Spotlight Presentation, 3% accept rate
T. Garipov*, **P. Izmailov***, D. Podoprikhin*, D. Vetrov, A. G. Wilson
arxiv

2018 Averaging Weights Leads to Wider Optima and Better Generalization
Uncertainty in Artificial Intelligence (UAI), ***Oral Presentation, 9% accept rate***
P. Izmailov*, D. Podoprikhin*, T. Garipov*, D. Vetrov, A. G. Wilson
arxiv

2018 Scalable Gaussian Processes with Billions of Inducing Inputs via Tensor Train Decomposition
Artificial Intelligence and Statistics (AISTATS), ***Oral Presentation, 5% accept rate***
P. Izmailov, A. Novikov, D. Kropotov
arxiv

2018 Tensor Train decomposition on TensorFlow (T3F)
Journal of Machine Learning Research (JMLR, published in 2020)
A. Novikov, **P. Izmailov**, V. Khrulkov, M. Figurnov, I. Oseledets
arxiv

2017 Faster variational inducing input Gaussian process classification
Journal of Machine Learning and Data Analysis (JMLDA)
P. Izmailov, D. Kropotov
arxiv

Workshop Papers

2022 On Feature Learning in the Presence of Spurious Correlations
ICML Principles of Distribution Shift (PODS) Workshop
P. Izmailov*, P. Kirichenko*, N. Gruver*, A. G. Wilson

2022 Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations
ICML Workshop on Spurious Correlations, Invariance, and Stability, ***Oral Presentation***
P. Kirichenko*, **P. Izmailov***, A. G. Wilson
arxiv

2019 Subspace Inference for Bayesian Deep Learning
ICML Workshop on Uncertainty and Robustness in Deep Learning, ***Oral Presentation***
P. Izmailov*, W. Maddox*, T. Garipov*, P. Kirichenko*, A. G. Wilson
PDF

2019 Semi-Supervised Learning with Normalizing Flows
ICML Workshop on Invertible Neural Nets and Normalizing Flows
P. Izmailov*, P. Kirichenko*, M. Finzi*, A. G. Wilson
PDF

2019 Invertible Convolutional Networks
ICML Workshop on Invertible Neural Nets and Normalizing Flows, ***Spotlight Presentation***
M. Finzi*, **P. Izmailov***, W. Maddox*, P. Kirichenko*, A. G. Wilson
PDF

2018 Fast Uncertainty Estimates and Bayesian Model Averaging of DNNs
UAI Workshop on Uncertainty in Deep Learning, ***Oral Presentation***
W. Maddox, T. Garipov, **P. Izmailov**, A. G. Wilson
PDF

2018 Improving Stability in Deep Reinforcement Learning with Weight Averaging
UAI Workshop on Uncertainty in Deep Learning
E. Nikishin, **P. Izmailov**, B. Athiwaratkun, P. Shvechikov, D. Podoprikhin, T. Garipov, D. Vetrov, A. G. Wilson
PDF

Talks

2025 NYU CILVR Seminar
Research Questions, Directions and Thoughts in LLM Reasoning; Event

2025 NeurIPS, New in ML Workshop
Invited talk
Thoughts on AI research and careers; Event

2025 Aspen Meeting on Foundations Models
Research Questions, Directions and Thoughts; Event

2025 ICLR, SCLS Workshop
Invited talk
Weak-to-strong generalization; Event

2025 ICLR, BiAlign Workshop
Invited talk, panel discussion on Human-AI Bidirectional Alignment
Weak-to-strong generalization; Event

2025 University of Washington, RAISE Seminar
Weak-to-strong generalization

2025 University of California, Los Angeles, NLP Seminar
Weak-to-strong generalization

2024 Simons Institute for the Theory of Computing, UC Berkeley
Debate: Sparks versus embers (discussant)
Video

2024 Simons Institute for the Theory of Computing, UC Berkeley
Weak-to-strong generalization
Video

2024 A Bayesian Odyssey in Uncertainty: from Theoretical Foundations to Real-World Applications
ECCV 2024 Tutorial with Gianni Franchi, Adrien Lafage, Olivier Laurent, Alexander Immer and Andrei Bursuc
Event; Video

2024 NYU AI Safety Reading Group
Weak-to-strong generalization

2024 Symposium on the Impact of Generative AI in the Physical Sciences (**Panelist**)
IAIFI, MIT; Event

2024 Columbia Human-Guided Machine Learning Seminar
Weak-to-strong generalization

2024 OpenAI Forum
Weak-to-strong generalization
Event

2023 Caltech, course on uncertainty quantification in machine learning (**Invited Lecture**)
Neural network loss surfaces and Bayesian neural nets

2022 Stanford, Chelsea Finn's group
Feature Learning and Distribution Shift

2022 MIT, Tommi Jaakkola's group
Understanding Knowledge Distillation

2022 University of Washington, Ludwig Schmidt's group
Feature Learning and Spurious Correlations

2022 Google Research, Shannon's Bandwagon meeting
On Uncertainty, Tempering, and Data Augmentation in Bayesian Classification

2022 Google Research, Sample Efficient Learning meeting
Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations

2021 Advances in Approximate Bayesian Inference, AABI (**Invited Talk**)
What Are Bayesian Neural Network Posteriors Really Like?
Video

2021 Max Plank Institute MIS and UCLA joint Seminar: Math Machine Learning
What Are Bayesian Neural Network Posteriors Really Like?
Video

2021 Teams at Google Brain and Perception
What Are Bayesian Neural Network Posteriors Really Like?

2021 Oxford Applied and Theoretical Machine Learning Group
What Are Bayesian Neural Network Posteriors Really Like?

2021 Teams at Google Brain, Translate and Perception
Does Knowledge Distillation Really Work?

2021 Oxford Applied and Theoretical Machine Learning Group
What Are Bayesian Neural Network Posteriors Really Like?

2021 International Conference on Machine Learning (ICML)
What Are Bayesian Neural Network Posteriors Really Like?

2021 Bayesgroup seminar, Moscow
What Are Bayesian Neural Network Posteriors Really Like?

2021 University of Freiburg, Frank Hutter's group
Bayesian Deep Learning and a Probabilistic Perspective of Generalization

2019 Bayesgroup seminar, Moscow
Scalable Bayesian inference in low-dimensional subspaces

2019 Harvard, Finale Doshi-Velez group
Subspace Inference for Bayesian Deep Learning

2019 Broad Institute of MIT and Harvard
How do we build neural networks we can trust?
Video

2018 Uncertainty in Artificial Intelligence (UAI)
Averaging Weights Leads to Wider Optima and Better Generalization
Video

2018 Artificial Intelligence and Statistics (AISTATS)
Scalable Gaussian Processes with Billions of Inducing Inputs via Tensor Train Decomposition

Teaching

2019 New York University
Teaching Assistant for "Bayesian Machine Learning" course

2018 Cornell University
Teaching Assistant for "Bayesian Machine Learning" course

Organizing

2025 Organizer of the Foundations of Reasoning in Language Models at NeurIPS 2025

2025 Organizer of the 2nd Workshop on Foundation Models in the Wild at ICLR 2025

2024 Organizer of the Workshop on Trustworthy Multi-modal Foundation Models and AI Agents (TiFA) at ICML 2024

2024 Organizer of the Workshop on Reliable and Responsible Foundation Models at ICLR 2024

2023 AISTATS Communications Chair

2021 Lead student organizer of the NeurIPS competition "Approximate Inference in Bayesian Deep Learning"

Other Service

2022 Group Leader at NeurIPS 2022 Education Outreach Day

2022 Facilitator at Women in ML (WiML) unworkshop, ICML 2022

Service

Action Editor	TMLR
Area Chair	NeurIPS 2026, COLM 2026, ICML 2026, ICLR 2026, ICLR 2025, ICLR 2024, COLM 2024
Reviewer (Journals)	JMLR, TMLR, Pattern Recognition
Reviewer (Conferences)	ECCV 2026, CVPR 2026, . . . , AAAI 2024, NeurIPS 2024, CVPR 2024, ICML 2024, NeurIPS 2023, ICML 2023, CVPR 2023, ICLR 2023, NeurIPS 2022, UAI 2022 (top 84 highest-scoring reviewers), ICML 2022 (top 10% reviewers), NeurIPS 2021, NeurIPS 2020, UAI 2020, ICML 2020, ICLR 2020, NeurIPS 2019 (top 400 highest-scoring reviewers), UAI 2019, ICML 2019, UAI 2018, NeurIPS 2018 (top 218 highest-scoring reviewers), AISTATS 2018, ICML 2018

Internship Experience

2022 Research Intern, Google Research (Brain)
- Supervisors: Lucas Beyer and Simon Kornblith
- Knowledge distillation for upstream large-scale image classifiers

2021–2022 Research Intern → Student Researcher, Google Research (Perception)
- Supervisors: Alex Alemi and Ben Poole
- Representation learning with information-constrained CLIP models

Summer 2020 Research Intern, Google Research (Perception)
- Supervisor: Matt Hoffman
- Large-scale Hamiltonian Monte Carlo for Bayesian Neural Networks

Summer 2019 Research Intern, Amazon AWS
- Supervisors: Yuyang (Bernie) Wang, Alexander J. Smola
- Multi-scale time series forecasting

Skills

Proficient Python, SciPy stack, PyTorch, JAX, Git, LaTeX

Used before MXNet, Gluon, Tensorflow