Curriculum Vitae

Andrew Gritsevskiy

January 21, 2025

Citizenship: United States

Address: 5413 Regent St
Madison, WI 53705
United States

Email: agritsevskiy@gmail.com
Personal website: andrew.gr

Education

2024 –	University of Wisconsin–Madison PhD candidate, computer science
2019 – 2022	University of Toronto, Honours B.S. Mathematics, computer science, stats, biology
2018 – 2019	University of California, Los Angeles Mathematics and computer science
2018	Canada/USA Mathcamp
2017 – 2018	Harvard University Extension Mathematics
2014 – 2018	Lexington High School

Research Positions

2025—pres. Researcher
MATS
Conducting AI safety research in white-box monitoring, chain-of-thought faithfulness, and scalable oversight

2024—pres. PhD candidate
UW–Madison
AI safety, scalable oversight, robotics

2023—pres. Researcher
MATS
Conducting AI safety research on model backdoors, adversarial robustness, scalable oversight, and models of misalignment with Jeffrey Ladish

2022—pres. Research Director
Cavendish Labs
Director of Artificial Intelligence research at Cavendish Labs, a 501(c)(3) research nonprofit. Leading the sparse autoencoder interpretability projects, the multimodal evaluation project, and the inverse scaling project.

2020—2022 Research Assistant
Vector Institute
I worked on distance-based planning for reinforcement learning, co-supervised by Silviu Pitis and Harris Chan in Prof. Jimmy Ba's lab.

2022 Research Assistant
The Hospital for Sick Children
I was a Data Sciences Institute Scholar at the Josselyn Frankland Lab, where I investigated how the brain encodes memory.

2021 Research Fellow
Institute for Advanced Research in Artificial Intelligence
Worked with Dr. Michael Kopp on reinforcement learning for Ramsey Theory.

2020—2021 UofT iGEM team
Generative modelling track

2019 Research lead
UCLA iGEM team
Worked with Mark Arbing at the Protein Expression Lab at the UCLA-DOE Institute and with Todd Yeates at the Yeates Lab.

2017—2018 Student researcher
MIT Media Lab
Worked with Maksym Korablyov and Dr. Joseph Jacobson on low-data transfer learning using capsule networks.

2016—2018 Student researcher
MIT Affinity project
Worked with Maksym Korablyov, Dr. Joseph Jacobson, Kfir Schreiber, Isaac Wolverton, Aditi Harini, and Manvitha Ponapatti on developing a deep learning library for molecular geometry.

2016—2017 Student researcher
Biomedical Cybernetics Laboratory, Harvard University
Conducted research on predicting biological properties of genomes with deep learning with Adithya Vellal and Dr. Gil Alterovitz

2016—2018 Student researcher
MIT PRIMES program

2015 Student researcher
Draper laboratory
Created personalized biosurveillance software with Albert Gerovitch and Gregory Barboy at Dr. Natasha Markuzon's lab.

Industry experience

2025 RunDPO.com—created an entire DPO platform in under 10 days, complete with automatic GPU scaling, user management, and a python library

2019 Developed blind-spot vehicle radars at Veoneer

Papers

1. Cavanagh, J., Sun, K. **Gritsevskiy, A.**, Bagni, D., Bannister, T., Head-Gordon (2024) SmileyLlama: Modifying Large Language Models for Directed Chemical Space Exploration. *NeurIPS 2024 Workshop on AI for New Drug Modalities*

2. Draguns, A.*, **Gritsevskiy, A.***, Motwani, S. R., Rogers-Smith, C., Ladish, J., de Witt, C. S. (2024) Unelicitable Backdoors in Language Models via Cryptographic Transformer Circuits. *NeurIPS 2024*

3. **Gritsevskiy, A.**, Panickssery, A., Kauffman, D., Gritsevskaya, I., Cavanagh, J., Chiang, J., La Roux, L., and Hung, M. (2024) REBUS: A Robust Evaluation Benchmark of Understanding Symbols. Preprint at arXiv:2401.05604

4. Krenn, M., Buffoni, L., Coutinho, B., Eppel, S., Foster, J. G., **Gritsevskiy, A.**, Lee, H., Lu, Y., Moutinho, J., Sanjabi, M., Sonthalia, R., Tran, N. M., Valente, F., Xie, Y., Yu, R., Kopp, M. (2023) Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network. Published in *Nature Machine Intelligence*

5. McKenzie, I., Lyzhov, A., Pieler, M., Parrish, A., Mueller, A., Prabhu, A., McLean, E., Kirtland, A., Ross, A., Liu, A., **Gritsevskiy, A.**, Wurgraft, D., Kauffman, D., Recchia, G., Liu, J., Cavanagh, J., Weiss, M., Huang, S., Droid, T. F., Tseng, T., Korbak, T., Shen, X., Zhang, Y., Zhou, Z., Kim, N., Bowman, S. R., Perez, E. (2023) Inverse Scaling: When Bigger Isn't Better. Published in TMLR; Featured paper

6. Kauffman, D., **Gritsevskiy, A.**, and Cavanagh, J. (2022) Finding Human Simulators by Varying Data Quality. *First place at the Prometheus ELK Prize*

7. **Gritsevskiy, A.** (2022) Control Theory and Efficient Heuristic Reinforcement Learning. *Submitted as part of MAT495 research course*

8. **Gritsevskiy, A.** and Korablyov, M. (2018) Capsule networks for low-data transfer learning. Preprint at arXiv:1804.10172

9. **Gritsevskiy, A.** (2017) Towards Generative Drug Discovery: Metric Learning using Variational Autoencoders. Preprint at math.mit.edu.

10. **Gritsevskiy, A.** and Vellal, A. (2016) Development and Biological Analysis of a Neural Network Based Genomic Compression System. Preprint at math.mit.edu.

11. Gerovitch, A., **Gritsevskiy, A.**, and Barboy, G. (2015) Mobile Health Surveillance: The Development of Software Tools for Monitoring the Spread of Disease. Preprint at math.mit.edu.

* *denotes equal contribution*

Teaching & Supervision

2025	MIT PRIMES Mentor in Computer Science, project TBD
2024	MIT PRIMES Mentor in Computer Science, supervising a project on sparse autoencoders for extracting FSMs from RL agents
2023	MIT PRIMES Mentor in Computer Science, supervising a project on the algebraic value-editing conjecture in AI alignment
2022	Taught a course on quantum algorithms at Camp Cape Cod
2022	MIT PRIMES Mentor in Computer Science, supervising a project on deep learning for kinematics
2022	Leading and facilitating an introductory effective altruism fellowship at the University of Toronto
2020	Taught a course on neural networks and deep learning at Camp Cape Cod
2019	Taught two one-day courses on deep learning and the curse of dimensionality at UCLA Splash
2018	Co-taught a class on the curse of dimensionality with Michelle Hung at Canada/USA Mathcamp
2018	Taught a three-day class on neural networks for visual recognition, inspired by Stanford's CS231n
2017	Taught two one-day classes on deep learning and molecular orbital theory at Lexington Splash

Talks

2020, CC	Distance-based Planning in Reinforcement Learning
2019, UCLA	Lie Groups in Physics
2018, MIT	Capsule Networks for Low-Data Transfer Learning
2017, MIT	Deep Learning Techniques for the Determination of Cross-Species Structural Gene Expression

Awards and Recognition

2024	Long-Term Future Fund Research Grant ($50000 award, for team of 3)
2023	NSF Compute Grant ($10000 value)
2023	Long-Term Future Fund Research Grant ($30000 award)
2022	Cavendish Labs Research Grant ($50000 value, until Dec. 2023)
2022	Eliciting Latent Knowledge Competition – First Place ($15000 prize, with Derik Kauffman and Joe Cavanagh)
2022	Inverse Scaling Prize – Two Third Prizes ($10000 prize, with Derik Kauffman and Joe Cavanagh)
2022	Nominated for Rhodes Scholarship for Canada
2022	Data Sciences Institute SUDS Research Scholar ($8000 award)

2019	Best Overall Hack—UCLA Hack On The Hill
2019	First place, UCLA algorithms competition
2018	National AP Scholar (5/5 on 10 exams)
2017	National Merit Scholarship Seminifinalist
2017	DOE National Science Bowl Wildcard Award
2017	Perfect SAT score in chemistry, molecular biology, and mathematics
2017	United States Computing Olympiad—Gold level
2016	Chinese-American Biomedical Association High School Research Award
2016	Musical compositions chosen for performance in Boston, MA and St. Petersburg, Russia

Expository writing

1. Gritsevskiy, A. (2024) Implementing an SHA transformer by hand
2. Gritsevskiy, A. (2024) Q-learning in RASP
3. Gritsevskiy, A. (2024) Hand-coding backdoors in transformers with RASP
4. Gritsevskiy, A. (2020) The Language of Nature.
5. Hung, M. and Gritsevskiy, A. (2018) The Curse of Dimensionality.

Relevant Projects

I have worked on dozens of artificial intelligence, reinforcement learning, and robotics projects. Details available upon request.