Changho Shin

cshin23@wisc.edu
https://ch-shin.github.io
1210 W Dayton St, Madison, WI 53706

RESEARCH INTERESTS

My research focuses on **data-centric AI for foundation models**, including large language models (LLMs) and multimodal foundation models (MLLMs). I develop methods for **efficient supervision**, leveraging **weak supervision**, **data selection**, and **weak-to-strong generalization** to enhance model performance with minimal human oversight. Additionally, I explore **training-free approaches**, such as **representation editing**, to steer foundation models at inference time, enabling robust adaptation and the adoption of new capabilities. My long-term vision is to develop frameworks for **supervising superhuman-level intelligence**, where I am investigating strategies like **scalable oversight** and **self-improvement** to ensure effective guidance, adaptation, and capability expansion in increasingly powerful AI systems.

University of Wisconsin-Madison Sep. 2020 – Present

- Ph.D. Computer Science, M.S. Mathematics
- Advisor: Professor Frederic Sala

Seoul National University Mar. 2015 – Feb. 2017

- M.S. Machine Learning
- Advisor: Professor Wonjong Rhee

Seoul National University Mar. 2011 – Feb. 2015

- B.A. in Psychology, B.S. in Computer Science and Engineering
- Graduated with honors (Cum Laude)

HONORS & AWARDS

Qualcomm Innovation Fellowship Finalist	2024
Best Paper Award Honorable Mention (NeurIPS R0-FoMo Workshop)	2023
NeurIPS 2023 Scholar Award	2023
Winner in DataComp competition (Filtering Track, Small)	2023
CS Departmental Scholarship (University of Wisconsin-Madison)	2020

PREPRINTS

[P6] Sungjun Cho, **Changho Shin**, Suenggwan Jo, Xinya Yan, Shourjo Aditya Chaudhuri, Frederic Sala, "LLM-Integrated Bayesian State Space Models for Multimodal Time-Series Forecasting", *Under Submission*, 2025.

[P5] Jitian Zhao*, **Changho Shin***, Tzu-Heng Huang, Srinath Namburi, Frederic Sala, "From Many Voices to One: A Statistically Principled Aggregation of LLM Judges", *Under Submission*, 2025.

[P4] Dyah Adila, Albert Ge, Avi Trost, Alexander Yun, Srinath Namburi, **Changho Shin**, Frederic Sala, Ramya Korlakai Vinayak, "SinguLab: A Testbed for Recursive ML Discovery", *Under Submission*, 2025.

[P3] **Changho Shin**, Xinya Yan, Suenggwan Jo, Sungjun Cho, Shourjo Aditya Chaudhuri, Frederic Sala, "TARDIS: Mitigating Temporal Misalignment via Representation Steering", *arxiv*, 2025.

[P2] Dyah Adila, **Changho Shin**, Yijing Zhang, Frederic Sala, "Alignment, Simplified: Steering LLMs with Self-Generated Preferences", *arxiv*, 2025.

[P1] Amanda Dsouza, Christopher Glaze, **Changho Shin**, Frederic Sala, "Evaluating Language Model Context Windows: A 'Working Memory' Test and Inference-time Correction", *arxiv*, 2024.

CONFERENCE PUBLICATIONS

[C7] **Changho Shin**, John Cooper, Frederic Sala, "Weak-to-Strong Generalization Through the Data-Centric Lens", *International Conference on Learning Representations (ICLR)*, 2025.

[C6] Yijing Zhang, Dyah Adila, **Changho Shin**, Frederic Sala, "Personalize Your LLM: Fake it then Align it", *North American Chapter of the Association for Computational Linguistics (NAACL) Findings*, 2025.

[C5] **Changho Shin**, Jitian Zhao, Sonia Cromp, Harit Vishwakarma, Frederic Sala, "OTTER: Improving Zero-Shot Classification via Optimal Transport", *Neural Information Processing Systems (NeurIPS)*, 2024.

[C4] Dyah Adila*, **Changho Shin***, Linrong Cai, Frederic Sala, "Zero-Shot Robustification of Zero-Shot Models With Auxiliary Foundation Models", *International Conference on Learning Representations (ICLR)*, 2024.

Best Paper Award Honorable Mention, Oral Presentation at *NeurIPS 2023 R0-FoMo Workshop.*

[C3] **Changho Shin**, Sonia Cromp, Dyah Adila, Frederic Sala, "Mitigating Source Bias for Fairer Weak Supervision", *Neural Information Processing Systems (NeurIPS)*, 2023.

[C2] **Changho Shin**, Winfred Li, Harit Vishwakarma, Nicholas Roberts, Frederic Sala, "Universalizing Weak Supervision", *International Conference on Learning Representations (ICLR)*, 2022.

[C1] **Changho Shin**, Sunghwan Joo, Jaeryun Yim, Hyoseop Lee, Taesup Moon, Wonjong Rhee, "Subtask Gated Networks for Non-Intrusive Load Monitoring", *AAAI Conference on Artificial Intelligence*, 2019.

JOURNAL PUBLICATIONS

[J2] **Changho Shin**, Eunjung Lee, Jeongyun Han, Jaeryun Yim, Hyoseop Lee, Wonjong Rhee, "The ENERTALK Dataset, 15 Hz Electricity Consumption Data from 22 Houses in Korea", *Nature Scientific Data*, 2019 (Impact Factor = 5.929).

[J1] **Changho Shin**, Seungeun Rho, Hyoseop Lee, Wonjong Rhee, "Data Requirements for Applying Machine Learning to Energy Disaggregation", *Energies*, May 2019 (Impact Factor = 2.707).

WORKSHOP PUBLICATIONS

[W4] Dyah Adila, **Changho Shin**, Yijing Zhang, Frederic Sala, "Is Free Self-alignment Possible?", *NeurIPS 2024 Workshop on Foundation Model Interventions (MINT).*

[W3] **Changho Shin***, Joon Suk Huh*, Elina Choi, "Pool-Search-Demonstrate: Improving Data-wrangling LLMs via better in-context examples", *NeurIPS 2023 Table Representation Learning (TRL) Workshop.* **Oral Presentation**.

[W2] **Changho Shin***, Tzu-heng Huang*, Sui Jiet Tay, Dyah Adila, Frederic Sala, "Multimodal Data Curation via Object Detection and Filter Ensembles", *ICCV 2023 Datacomp Workshop* (Rank #1 in DataComp competition filtering track (small)).

[W1] **Changho Shin**, Alice Schoenauer-Sebag, "Can we get smarter than majority vote? Efficient use of individual rater's labels for content moderation", *NeurIPS 2022 Efficient Natural Language and Speech Processing (ENLSP) Workshop.*

JOB EXPERIENCE

Microsoft Research, Cambridge, USA — Jun. 2025 – Aug. 2025

(Incoming) Research Intern

- Mentor: David Alvarez-Melis

Snorkel AI, California, USA — Jun. 2024 – Aug. 2024

Research Intern

- Mentor: Christopher Glaze, Paroma Varma

Twitter, San Francisco, USA — Jun. 2022 – Aug. 2022

ML Engineer Intern

- Mentor: Alice Schoenauer Sebag • Manager: Milind Ganjoo
- Improving toxicity classification via weak supervision [W1]

Encored Technologies, Seoul, Korea — Jan. 2018 – Jul. 2020
Data Scientist
- Manager: Hyoseop Lee
- Non-intrusive load monitoring [C1, J1, J2], Energy forecasting

Korea Institute for Defense Analyses, Seoul, Korea — Jan. 2017 – Dec. 2017
Researcher

TEACHING EXPERIENCE

University of Wisconsin-Madison
- Teaching assistant for CS 839 (Foundation Models) — Fall 2023
- Teaching assistant for CS 300 (Programming II) — Fall 2022, Spring 2023
- Teaching assistant for CS 760 (Machine Learning) — Fall 2021, Spring 2022
- Teaching assistant for CS 320 (Data Programming II) — Spring 2021
- Teaching assistant for CS 220 (Data Programming I) — Fall 2020

GRADUATE COURSEWORK

- M2680.001300 Machine Learning for Information Studies @ SNU
- M2680.001400 Social Computing @ SNU
- 493.613 Mathematics for Intelligent Systems (Numerical Linear Algebra) @ SNU
- 493.701 Learning and Applications of Deep Neural Networks @ SNU
- M0000.005400 Convex Optimization @ SNU
- M0000.005400 Neural Networks @ SNU
- CS537 Introduction to Operating Systems @ UW-Madison
- CS639.004 Introduction to Computational Learning Theory @ UW-Madison
- CS726 Nonlinear Optimization 1 @ UW-Madison
- CS744 Big Data Systems @ UW-Madison
- CS761 Mathematical Foundations of Machine Learning @ UW-Madison
- CS784 Foundations of Data Management @ UW-Madison
- CS787 Advanced Algorithms @ UW-Madison
- CS839 Probability and Learning in High Dimension @ UW-Madison
- CS880 Advanced Topics in Learning Theory @ UW-Madison
- Math521 Analysis I @ UW-Madison
- Math522 Analysis II @ UW-Madison
- Math551 Elementary Topology @ UW-Madison
- Math621 Analysis III (Analysis on Manifolds) @ UW-Madison
- Math629 Introduction to Measure and Integration @ UW-Madison
- Math721 A First Course in Real Analysis @ UW-Madison
- Math733 Theory of Probability I @ UW-Madison
- Math734 Theory of Probability II @ UW-Madison
- Math761 Differentiable Manifolds @ UW-Madison
- Math833 Modern Discrete Probability @ UW-Madison
- Math888 Randomized Linear Algebra @ UW-Madison
- Stat992 Optimal Transport and Applications to Machine Learning @ UW-Madison

TECHNICAL SKILLS

Machine Learning / Deep Learning / Data Science
PyTorch, TensorFlow, Keras, scikit-learn, NumPy, Pandas, SciPy

DBMS
MySQL, MongoDB, PySpark

Research & Development Tools
Visual Studio Code, Jupyter, PyCharm, Docker, GitHub, CircleCI, Shell, AWS

Programming Languages
Python, R, MATLAB, Java, Go, C, LaTeX