Yukai Yang

929-599-8878 | Email | LinkedIn | GitHub

Education

University of Chicago Chicago, IL
M.S. in Statistics *Sep 2023 - Mar 2025*

- GPA: 3.7
- Research Interest: Uncertainty Quantification, Statistical Learning Theory, High-dimensional Statistics, Random Matrix Theory

New York University New York, NY
B.A. in Data Science and Mathematics *Aug 2019 - Dec 2022*

- GPA: 3.9
- Courses taken: Foundations of Machine Learning, Deep Learning, Mathematical Statistics, ML for Functional Genomics, Advacned Techniques in DL, Convex and Non-smooth Optimization, Algorithm, Real Analysis, etc.

Technical Skills

Programming Languages: Python, R, SQL, Bash, MATLAB, C.
Libraries: PyTorch, TensorFlow, Jax, transformers, opencv, H2O, ray, nltk, scipy, pandas, matplotlib, seaborn, numpy
Other Tools: Postgres, MongoDB, GCP, HPC, AWS.

Work Experience

Student Researcher (Remote) Feb 2024 –
Center for Data Science, New York University *New York, NY*

- Study hallucinating behaviors of large language models from their answer uncertainty
- Train weaker models to generalize and predict strong models on various tasks (weak-to-strong generalization)

Visiting Researcher (Remote) Mar 2024 -
Trustworthy AI Lab, University of California, Los Angeles *Los Angeles, CA*

- Align Large Language Models with safety standards using only in-context learning
- Train a prompt generator/classifier system in an adversarial manner to optimize the prompts for safety

Visiting Academics: Research Assistant Feb 2022 – Oct 2023
Center for Data Science, New York University *New York, NY*

- Use deep bayesian active learning to select the most informative images from a unlabeled pool
- Integrate function-space regularization with AL acquisition functions to improve uncertainty quantification

Teaching Assistant and Graders 2022, 2023, 2024
Courant Institute of Mathematical Science, New York University *New York, NY*

- University of Chicago: Clinical Data Science (DS, TA)
- New York University: Linear Algebra (Math, TA), Advanced Topics in Data Science: Deep Learning (DS, Grader)

Publications

Weak-to-strong Generalization, 2024 | *AI Safety, Large Language Models*

- Submitted to NeurIPS Workshops
- Use simple probe heads to learn representations of weaker models and predict strong models' uncetainty

Active Learning with Risk-aware Acquisition Functions, 2023 | *Bayesian Deep Learning*

- Submitted to 5th Symposium on Advances in Approximate Bayesian Inference (AABI)
- Use Hyperbolic Absolute Risk Averse (HARA) to better evaluate the uncertainty of unlabeled data

Projects

HAAT: Improve Adversarial Robustness via Hessian Approximation | *Adversarial Training*

- Improve Projected Gradient Descent with our new algorithm that includes a second-order approximation term
- Analyze and explain different PGD-based algorithms improvements with learning theory techniques

Uniform Convergence for Double-descent Curve in Different Models | *Statistics, DL Theory*

- Give a survey-style summary of recent relevant work on theoretical analysis of double-descent curve
- Provide a proof for the bound of convergence and experiment it with random feature models and DNNs

Benchmarking GPT-3: how LLMs Learn New Words | *NLP, Few-shot Learning with LLMs*

- Examine whether LLMs manage to reason in a similar way as an intellectual agent to understand new words
- Help models to learn complicated compound words faster | Prevent inadvertent entry from confusing the model

Using Function-Space-VI-based Active Learning to Label Protest Images | *CV, Active Learning*

- Apply and improve function space variational inference as the heuristic functions adaptable to the dataset
- Develop the pretrained AL model into an API that can quickly help annotate protest images in massive data flow

Kernel Approximation for Gradient Descent based Algorithms | *Convex Optimization*

- Explore how kernel methods can be used as an equivalence or approximation of various gradient descent algorithms
- Improve the error bounds of the approximations to better interpret how deep learning models learn features

Grants

Summer Research Grant - $5000 Jul 2024

Center for Data Science, New York University *New York, NY*

- Award kindly provided by Tim Rudner
- Work on weak-to-strong prediction of Large Language Models' uncertainty

Extra-curricular Activities

ICPC in Greater New York Region Mar 2022

International Collegiate Programming Contest *New York, NY*

- Programming Competition that relies on programming and math
- 1st place in all NYU team. 3rd place in the Greater New York Region
- Heading to the North American Competition

Bud Challenge 2021 – Data Analysis track Jun 2021

Business competition held by the beer company Budweiser *Shanghai*

- Build a B2B, ML-based recommendation system for retailers.
- 3rd Place in the Final Round.