Yukai Yang

Website | Email | LinkedIn | GitHub

Education

University of Chicago — Chicago, IL
M.S. in Statistics (GPA: 3.7/4.0) — *Sep 2023 - Mar 2025*
Courses taken: Topics in High-Dimensional Probability with Applications in Data Science, Topics in Random Matrix Theory, Machine Learning on Graphs, Groups and Manifolds, Inverse Problem and Data Assimilation, etc.

New York University — New York, NY
B.A. in Data Science and Mathematics, (magna cum laude, GPA: 3.9/4.0) — *Aug 2019 - Dec 2022*
Courses taken: Foundations of Machine Learning, Mathematical Statistics, ML for Functional Genomics (Columbia), Advacned Techniques in DL, Convex and Non-smooth Optimization, etc.

Peer-Reviewed Papers

* indicates equal contributions.

[1] Yukai Yang, Chris Zhuang, Ruqi Zhang. **C2C: Clustering-to-Confidence for Reliable Molecule Generation**.
In preparation for ICML 2026

- Proposed a new method that clusters generated molecules using a learned similarity classifier, measures cluster entropy, and retains the lowest-entropy cluster as the reliable set.
- Demonstrated substantial reliability gains on various molecular generation benchmarks with generic LLMs.
- Provides a model-agnostic, scalable plug-in that boosts generator trustworthiness via similarity clustering.

[2] Xiaofeng Lin*, Yukai Yang*, Daniel Guo, Sahil Arun Nale, Charles Fleming, Guang Cheng. **Divide-and-Conquer Attacks on LLM Agents: Orchestrating Multi-Step Jailbreaks in Tool-Enabled Systems**.
Under review at ARR 2025

- Showed that multi-agent tool-using systems are uniquely vulnerable: malicious goals can be decomposed into benign steps that evade per-turn safety checks.
- Introduced a divide-and-conquer framework showing agent pipelines amplify jailbreak success.
- Showed vulnerabilities across both close and open-source models, motivating holistic, multi-step defenses.

[3] Yukai Yang*, Tracy Zhu*, Marco Morucci, Tim G. J. Rudner. **Weak-to-strong Confidence Prediction for Large Language Models**.
Under review at TMLR 2025. Previously presented at: Workshop on Statistical Foundations of Large Language Models, Attributing Model Behavior at Scale, Safe Generative AI, and Regulatable ML. (**NeurIPS Workshop**), 2024

- Introduced a method where a lightweight probe over frozen backbone embeddings predicts whether a stronger black-box LLM will be correct, without accessing the model itself.
- Provided empirical evidence that uncertainty transfers across scales, with strong QA and reasoning results.
- Showed that selective prediction via weak-to-strong transfer improves black-box reliability, informing future safety and calibration methods.

Grants & Awards

Summer Research Grant - $5000 — Jul 2024
Center for Data Science, New York University — *New York, NY*

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

NYU Dean's Undergraduate Research Fund - $1000 — Sep 2022
College of Arts and Science, New York University — *New York, NY*

- Research grant to support undergraduate research
- Studied active learning with entropy-based heuristic for vision models

Wasserman Center Internship Grant - $1000 — Jun 2022
Wasserman Center for Career Development, New York University — *New York, NY*

- Summer internship for students to conduct research
- Research on machine learning computational social science datasets

Services

Reviewer

- 2025: AISTATS 2026; ACL Rolling Review (May round); NeurIPS Workshops: Regulatable ML, SafeGenAI
- 2024: NeurIPS Workshops: Regulatable ML (RegML), Attributing Model Behavior at Scale (ATTRIB), Statistical Foundations of Large Language Models (SFLLM), SafeGenAI

Competitions

ICPC in North American Competition May 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
- Top 30% in the North American Competition

Bud Challenge 2021 – Data Science 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.

Research & Teaching Experience

Independent Researcher (Remote) 2024
Department of Statistics and Data Science, Yale University *New Haven, CT*
- Analyze representation learning behavior of attention-based models with low-rank adapation
- Provide theoretical understanding of how learning rate and chosen rank impact the model's generalization

Student Researcher 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 to predict strong models' performance (weak-to-strong generalization)

Visiting Researcher (Remote) 2024
Trustworthy AI Lab, University of California, Los Angeles *Los Angeles, CA*
- Align Large Language Models with safety standards using in-context learning
- Train a prompt attacker/defender system in an adversarial manner to optimize the prompts for safety

Visiting Academics: Research Assistant 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 acquisition functions to improve uncertainty quantification

Teaching Assistant and Graders 2022, 2023, 2024
- University of Chicago: Clinical Data Science (DS, TA)
- New York University: Linear Algebra (Math, TA), Advanced Topics in Data Science: Deep Learning (DS, Grader)

Other Researach 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 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 pretrained AL model to help annotate protest images more efficiently

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

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: GCP, HPC, AWS, Postgres, MongoDB