Addison J. Wu

addisonwu@princeton.edu addisonwu05.github.io Citizenship: USA, Canada

Education

2023 – 2027 **Princeton University**
BSE in Computer Science, Minor in Mathematics – GPA: 3.98/4.00
Honors: Exemplary Independent Work Award, Shapiro Prize for Academic Excellence, NeurIPS Scholar Award, Tau Beta Pi, Sigma Xi (Associate)
Coursework: addisonwu05.github.io/academics

Experience

2026 – Center for AI Safety
Research Intern, advised by Mantas Mazeika, Dan Hendrycks

2024 – Princeton Language and Intelligence
Undergraduate Researcher, advised by Tom Griffiths, Zhuang Liu
Model behavior, safety, post-training, evaluation

2023 – 2025 Multi-Physics Interaction Lab, University of Waterloo
Undergraduate Researcher, advised by Jean-Pierre Hickey
Neural networks for atmospheric infrasound localization

Publications

Google Scholar ∗ → Equal contribution

Conference/Journal Papers

1. **Large Language Models Develop Novel Social Biases Through Adaptive Exploration**
Addison J. Wu*, Ryan Liu*, Xuechunzi Bai, Thomas L. Griffiths
ICML 2026 Oral. *SPSP 2026 Invited Symposia Talk*. [link]
2. **Ads in AI Chatbots? An Analysis of How Large Language Models Navigate Conflicts of Interest**
Addison J. Wu*, Ryan Liu*, Shuyue Stella Li, Yulia Tsvetkov, Thomas L. Griffiths
COLM 2026. [link]
3. **Are Large Language Models Sensitive to the Motives Behind Communication?**
Addison J. Wu*, Ryan Liu*, Kerem Oktar*, Theodore R. Sumers, Thomas L. Griffiths
NeurIPS 2025. *PragLM @ COLM 2025 Oral*. [link]
4. **Mind Your Step (by Step): Chain-of-Thought can Reduce Performance on Tasks where Thinking Makes Humans Worse**
Ryan Liu*, Jiayi Geng*, **Addison J. Wu**, Ilia Sucholutsky, Tania Lombrozo, Thomas L. Griffiths
ICML 2025. [link]
5. **Toward a Neural Network-Based Approach for Improved Atmospheric Infrasound Localization**
Addison J. Wu, Arnav Joshi, Jean-Pierre Hickey
IEEE Access, 2025 [link]

Workshop Papers

1. **Quantifying Empirical Compute-Supervision Tradeoffs in RLVR**
Ryo Mitsuhashi*, Patrick Chen*, Isabelle Tseng*, Jasin Cekinmez*, **Addison J. Wu***
Combining Theory and Benchmarks @ ICML 2026. [link]

2. **OccludeBench: Can VLMs See the Hint?**
Jasin Cekinmez*, **Addison J. Wu***, Ryo Mitsuhashi*, Linrong Cai, Xingyu Fu, Zhuang Liu
Computer Vision in the Wild @ CVPR 2026. [link]

3. **Query Timing Produces Opposite Positional Biases Between LLMs and Humans**
Jasin Cekinmez*, **Addison J. Wu***, Thomas L. Griffiths
ICBINB @ ICLR 2026 *Entropic Paper Award*. [link]

Preprints

1. **SportD: Can VLMs Physically Strategize?**
Jasin Cekinmez*, **Addison J. Wu***, Haotian Xia, Akshaya Bharadhwaj, Anay Putty, Anirudh Ravishankar, Jaewoong Lee, Jinglin Xiao, Kyumin Andrew Shim, Mishika Ahuja, Nisarga Patil, Leo Liu, Zhuohan Liu, Weining Shen
arXiv preprint, 2026 [link]

2. **Guess the Unified Model: How Much Can We Recover from Generated Images?**
Jasin Cekinmez*, Ryo Mitsuhashi*, **Addison J. Wu***, Yida Yin
arXiv preprint, 2026 [link]

Press Coverage

2026 Arizona's Family, AI can invent its own hiring bias — and it happens fast, new study finds
2026 MIT Technology Review, AI is more likely than humans to form biases when hiring
2026 Indian Express, Your AI 'assistant' is making money for someone else, you are the one paying. Here's exactly how
2026 Hacks/Hackers, New research: 18 of 23 AI models prioritize company revenue over users when ads enter the picture
2025 The Information, Where LLMs Are Falling Short [full text]

Invited and Contributed Talks

2026 **ICML 2026 Oral**
Large Language Models Develop Novel Social Biases Through Adaptive Exploration

2026 **SPSP 2026 Invited Symposia**
Large Language Models Develop Novel Social Biases Through Adaptive Exploration
(presented by co-author X. Bai)

2025 **UCLA NLP Seminar Series**
Are Large Language Models Sensitive to the Motives Behind Communication?
Large Language Models Develop Novel Social Biases Through Adaptive Exploration
(presented by co-first author R. Liu)

2025 **PragLM @ COLM 2025**
Are Large Language Models Sensitive to the Motives Behind Communication?

Teaching Experience

2024–2025 Undergraduate Course Assistant, Algorithms and Data Structures (COS 226)
Fall 2024, Spring 2025, Fall 2025

Professional Services

2026 Reviewer, NeurIPS 2026
2025 Reviewer, MTI-LLM @ NeurIPS 2025