Proposal: The 2nd Workshop on Reliable and Responsible Foundation Models at ICML 2025

Tagline: Making foundation models more reliable and responsible to be deployed in society.
Modality: Hybrid
Anticipated audience size: We expected more than 500 attendees.
Contacts: Xinyu Yang: xinyuya2@andrew.cmu.edu; Huaxiu Yao: huaxiu@cs.unc.edu

Workshop Summary

Foundation models (FMs), with their emergent and reasoning abilities, are reshaping the future of scientific research and broader human society. However, as their intelligence approaches or surpasses that of humans, concerns arise regarding their responsible use in real-world applications, such as reliability, safety, transparency, and ethics. The workshop on reliable and responsible FMs delves into the urgent need to ensure that such models align with human values. The significance of this topic cannot be overstated, as the real-world implications of these models impact everything from daily information access to critical decision-making in fields like medicine and finance, especially for embodied FMs that directly interact with the physical world. Stakeholders, including developers, practitioners, and policymakers, care deeply about this because the reliable and responsible design, deployment, and oversight of these models dictate not only the success of AI solutions but also the preservation of societal norms, order, equity, and fairness. Some of the fundamental questions that this workshop aims to address are:

- **Diagnosis:** How can we identify and characterize unreliable and irresponsible behaviors in FMs? Topics include prompt sensitivity, lack of self-consistency, and hallucinations in generation.
- **Evaluation:** How should we assess the harmful capabilities of FMs and quantify their societal impact? For example, how can we predict the consequences of misuse of highly capable FMs? In addition, how can we evaluate FMs with limited human-annotated data?
- **Sources:** How can we pinpoint and understand the known or emerging sources of FM unreliability? This involves examining training data, optimization objectives, and architectural design.
- **Generalization:** How can responsible and reliable properties be effectively adapted to increasingly advanced FMs, particularly as they incorporate new features such as more modalities or long CoT?
- **Governance:** What principles or guidelines should inform the next generation of FMs to ensure they are reliable and responsible? How can real-time monitoring of these FMs be enabled?
- **Guarantee:** Can we establish theoretical frameworks for reliably and responsibly FMs? For instance, how can we establish effective risk control frameworks for FMs?
- **Practice:** How to leverage domain-specific knowledge to guide FMs towards improved reliability and responsibility across diverse areas, such as drug discovery, education, or clinical health?

As prospective participants, we primarily target ML researchers and industry partitioners interested in the questions and foci outlined above. Our target audience includes professionals deeply involved with FMs and their applications, especially those who focus on the reliability and responsibility of these models. We also welcome submissions from researchers in the natural sciences (e.g., physics, chemistry, biology) and social sciences (e.g., pedagogy, sociology) to offer attendees a more comprehensive perspective. In summary, our topics of interest include, but are not limited to:

- Theoretical foundations of FMs and related domains, including uncertainty quantification, continual learning, and reinforcement learning
- Empirical investigations into the reliability and responsibility of various FMs
- In-depth discussions exploring new dimensions of FM reliability and responsibility
- Interventions during pre-training to enhance the reliability and responsibility of FMs
- Innovations in post-training processes to bolster the reliability and responsibility of FMs
- Advancements in improving the reliability and responsibility of FMs for test-time scaling

- Discussions on aligning models with potentially superhuman capabilities to human values
- Benchmark methodologies for assessing the reliability and responsibility of FMs
- Issues of reliability and responsibility of FMs in broad applications

Invited Speakers

We are pleased that a group of researchers have agreed to give invited talks at our workshop. Each speaker will bring a unique perspective to current developments in the reliability and responsibility of FMs. We aim to provide titles for all talks prior to the event. **Our speakers include:**

1. Manish Raghavan (Professor at MIT Sloan and EECS, male, **confirmed**)
2. Richard Zemel (Professor at Columbia CS, male, **confirmed**)
3. Sarah H. Cen (Incoming Assistant Professor at CMU ECE, female, **confirmed**)
4. Andrew Ilyas (Incoming Assistant Professor at CMU CS and ECE, male, **confirmed**)
5. René Vidal (Rachleff University Professor at University of Pennsylvania, male, **confirmed**)
6. Caiming Xiong (VP of AI Research and Applied AI at Salesforce, male, **confirmed**)
7. Philip Torr (Professor at University of Oxford, male, **confirmed**)
8. Kate Donahue (Incoming Assistant Professor at UIUC, female, **confirmed**)

Diversity Commitment

In the **selection of organizers and speakers**, we promoted diversity in all its forms. The final roster of organizers and speakers comprises individuals from varied genders, races, affiliations, and scientific backgrounds. Among the organizers and speakers, we ensured representation across the full spectrum of scientific seniority, including Ph.D. students, assistant professors, full professors, and industry researchers.

Tentative Schedule

Recognizing the challenges posed by varying time zones in a hybrid meeting format, we will incorporate live streams via Zoom and recordings from YouTube to ensure broad participation. The poster session can be attended either virtually or in person. We will utilize a channel in a chat platform like Rocket.Chat to facilitate interactions among workshop participants.

Morning:

- 08:50 – 09:00 Introduction and opening remarks
- 09:00 - 09:30 Invited Talk 1
- 09:30 - 10:00 Invited Talk 2
- 10:00 - 10:15 Contributed Talk 1
- 10:15 - 11:15 Poster Session 1
- 11:15 - 11:45 Invited Talk 3
- 11:45 - 12:15 Invited Talk 4
- 12:15 - 13:30 *Break*

Afternoon:

- 13:30 - 14:00 Invited Talk 5
- 14:00 - 14:30 Invited Talk 6
- 14:30 - 14:45 Contributed Talk 2
- 14:45 - 15:45 Poster Session 2
- 15:45 - 16:15 Invited Talk 7
- 16:15 - 16:30 Contributed Talk 3
- 16:30 - 17:00 Invited Talk 8
- 17:00 - 18:00 Panel discussion

History and Previous Related Workshops

Our workshop builds upon the "Reliable and Responsible Foundation Models" workshop at ICLR 2024. While continuing to address the reliability and responsibility of FMs, we will place additional emphasis on the related issues of more advanced FMs using test-time scaling and their broader applications. Unlike other workshops that focus on a singular aspect of FMs' reliability and responsibility, we seek to stimulate discussions on these issues from various angles, including theoretical underpinnings, model architectures, and implications in real-world applications. In the past two years, several workshops have touched upon themes related to our focus, including the "Models of Human Feedback for AI Alignment" (ICML 2024), "Next Generation of AI Safety" (ICML 2024), "Trustworthy Multi-modal Foundation Models and AI Agents" (ICML 2024), "Workshop on Responsibly Building Next Generation of Multimodal Foundation Models" (NeurIPS 2024), "Socially Responsible Language Modelling Research" (NeurIPS 2024), "Towards Safe & Trustworthy Agents" (NeurIPS 2024), "Safe Generative AI", "2nd ICML Workshop on New Frontiers in Adversarial Machine Learning" (ICML 2023), "Workshop on Spurious Correlations, Invariance and Stability" (ICML 2023), "Workshop on Distribution Shifts" (NeurIPS 2023), "Socially Responsible Language Modelling Research" (NeurIPS 2023).

Organizers and Biographies

Mohit Bansal (UNC-Chapel Hill)

- Email: mbansal@cs.unc.edu
- Webpage: https://www.cs.unc.edu/ mbansal/
- Google Scholar: https://scholar.google.com/citations?user=DN8QtscAAAAJ&hl=en
- Bio: Mohit Bansal is the John R. & Louise S. Parker Professor and the Director of the MURGe-Lab (UNC-NLP Group) in the Computer Science department at UNC-Chapel Hill. Prior to this, he was a research assistant professor (a 3-year endowed position) at TTI-Chicago. He received his Ph.D. in 2013 from the University of California at Berkeley (where he was advised by Dan Klein) and his B.Tech. from the Indian Institute of Technology at Kanpur in 2008. His research expertise is in natural language processing and multimodal machine learning, with a particular focus on grounded and embodied semantics, language generation and Q&A/dialogue, and interpretable and generalizable deep learning. He is a recipient of the IIT Kanpur Young Alumnus Award, DARPA Director's Fellowship, NSF CAREER Award, Google Focused Research Award, Microsoft Investigator Fellowship, Army Young Investigator Award (YIP), DARPA Young Faculty Award (YFA), and outstanding paper awards at ACL, CVPR, EACL, COLING, and CoNLL. He has been a keynote speaker for the AACL 2023 and INLG 2022 conferences. His service includes ACL Executive Committee, ACM Doctoral Dissertation Award Committee, CoNLL Program Co-Chair, ACL Americas Sponsorship Co-Chair, and Associate/Action Editor for TACL, CL, IEEE/ACM TASLP, and CSL journals.

Kate Donahue (MIT and UIUC)

- Email: kpd46@mit.edu
- Webpage: www.katedonahue.me
- Google Scholar: https://scholar.google.com/citations?user=c9SPOdwAAAAJ&hl=en
- Bio: Kate Donahue is a METEOR postdoc at MIT and an incoming CS professor at UIUC. She works on algorithmic problems relating to the societal impact of AI, such as fairness, human/AI collaboration, and game-theoretic models of data sharing. She has previously organized workshops at EC and TTIC and an AI for Society seminar series at MIT.

Giulia Fanti (Carnegie Mellon University)

- Email: gfanti@andrew.cmu.edu
- Webpage: https://gfanti.github.io/
- Google Scholar: https://scholar.google.com/citations?user=Rn_BmTYAAAAJ&hl=en&oi=ao
- Bio: Giulia Fanti is an Angel Jordan Associate Professor of ECE at Carnegie Mellon University, focusing on privacy-preserving technologies. She obtained her Ph.D. in EECS from U.C. Berkeley and her B.S. in ECE from Olin College of Engineering in 2010. Her work has been recognized with several awards, including best paper awards, a Sloan Fellowship, an Intel Rising Star Faculty Award, and an ACM SIGMETRICS Rising Star Award.

David Madras (Google Deepmind)

- Email: madras@cs.toronto.edu
- Webpage: https://www.cs.toronto.edu/ madras/
- Google Scholar: https://scholar.google.com/citations?user=MgnNDpkAAAAJ

- Bio: David Madras is a research scientist at Google Deepmind, contributing to developing the Gemini model series. Previously, he received his Ph.D. at the University of Toronto, advised by Richard Zemel. He is primarily interested in how to learn better and fairer algorithmic decision-making systems. My interests include fairness, causal inference, and generative modeling.

Han Shao (University of Maryland, College Park)

- Email: han@cmsa.fas.harvard.edu
- Webpage: https://sites.google.com/view/hanshao
- Google Scholar: https://scholar.google.com/citations?user=OVRxQj8AAAAJ
- Bio: Han Shao is a CMSA postdoc at Harvard, working with Cynthia Dwork and Ariel Procaccia. Starting in fall 2025, she will be an Assistant Professor in the Department of Computer Science at the University of Maryland, College Park (UMD). Before that, she did her PhD at TTIC, where she was extremely fortunate to be advised by Avrim Blum.

Hongyi Wang (Rutgers University)

- Email: hw689@rutgers.edu
- Webpage: https://hwang595.github.io/
- Google Scholar: https://scholar.google.com/citations?user=zYdZORsAAAAJ
- Bio: Hongyi Wang is the Head of Infrastructure at GenBio AI. He will join the CS department at Rutgers University as a tenure-track Assistant Professor in Fall 2025. He worked with Prof. Eric Xing for two years as a postdoctoral fellow at CMU. His research focuses on large-scale machine learning algorithms and systems. He obtained his Ph.D. from the Department of Computer Sciences at the University of Wisconsin-Madison. He has co-organized workshops at MLSys 2023 and has served as a tutorial speaker at CVPR 2023. Dr. Wang has received several accolades, including the Rising Stars Award from the Conference on Parsimony and Learning in 2024, the NAACL 2024 Best Demo Award runner-up, and the Baidu Best Paper Award at the SpicyFL workshop at NeurIPS 2020.

Steven Wu (Carnegie Mellon University)

- Email: zstevenwu@cmu.edu
- Webpage: https://zstevenwu.com/
- Google Scholar: https://scholar.google.com/citations?user=MbF6rTEAAAAJ
- Bio: Steven Wu is an Assistant Professor in the School of Computer Science at CMU, with a primary appointment in the Software and Societal Systems Department (with the Societal Computing program) and affiliated appointments with the Machine Learning Department and the Human-Computer Interaction Institute. He is also affiliated with the CyLab and the Theory Group.

Xinyu Yang (Carnegie Mellon University)

- Email: xinyuya2@andrew.cmu.edu
- Webpage: https://xinyuyang.me/
- Google Scholar: https://scholar.google.com/citations?user=Fvq2R14AAAAJ
- Bio: Xinyu Yang is a second-year PhD student at CMU ECE. His research focuses on the development of scalable and generalizable foundation model systems in the wild. Recently, he has been particularly interested in hardware-aware algorithm design with sub-linear complexity.

Zhun Deng (UNC-Chapel Hill)

- Email: zhundeng@cs.unc.edu
- Webpage: https://www.zhundeng.org/
- Google Scholar: https://scholar.google.com/citations?user=nkmi-moAAAAJ&hl=en&authuser=2
- Bio: Zhun Deng is a tenure-track Assistant Professor at the Department of Computer Science, University of North Carolina at Chapel Hill. Previously, he was a postdoctoral researcher at Columbia University. He completed his Ph.D. in the Theory of Computation group at Harvard University. He is broadly interested in problems at the frontier of machine learning, statistics, and theoretical computer science. He mostly develops practical frameworks with theoretical guarantees to address cutting-edge issues regarding reliable learning and responsible computing. His papers have won multiple honors, such as Spotlight and Oral Presentation, at flagship machine learning conferences, including ICML, NeurIPS, ICLR, and AISTATS.

Huaxiu Yao (UNC-Chapel Hill)

- Email: huaxiu@cs.unc.edu
- Webpage: https://www.huaxiuyao.io/
- Google Scholar: https://scholar.google.com/citations?hl=en&user=A20BZnQAAAAJ&view_op=list_works&sortby=pubdate
- Bio: Huaxiu Yao is a tenure-track Assistant Professor at the Department of Computer Science with a joint appointment in the School of Data Science and Society, UNC-Chapel Hill. He was a Postdoctoral Scholar in Computer Science at Stanford University. Huaxiu earned his Ph.D. degree from Pennsylvania State University. He focuses on the theoretical and applied aspects of building reliable and responsible foundation models. He is also dedicated to applying foundation models to solve real-world scientific and social applications, such as healthcare, transportation, and education. He has organized and co-organized workshops at ICML and NeurIPS and has served as a tutorial speaker at conferences such as KDD, AAAI, and IJCAI. Additionally, Huaxiu has extensive industry experience, having interned at companies such as Amazon Science, and Salesforce Research.