Proposal: ICLR 2024 Workshop on Reliable and Responsible Foundation Models

Tagline: Making foundation models more reliable and responsible to be deployed in society.
Modality: Hybrid
Anticipated audience size: We expected more than 1000 attendees.

Workshop Summary

Models and methods based on large-scale foundation models (FMs) are dominating in a large variety of applications in natural language processing, computer vision and other domains. Positioned at the frontier of the technological surge, foundation models offer a plethora of benefits with their incredible capabilities, but also introduce challenges related to reliability, transparency, and ethics. The workshop on reliable and responsible FMs addresses the urgent need to ensure that such models are trustworthy, robust and aligned with human values. The significance of this topic cannot be emphasized enough, as the real-world implications of foundation models impact everything from daily information access to critical decision-making in fields ranging from medicine to finance. The responsible design, deployment, and oversight of these models affect not only the success of AI solutions but also the preservation of societal norms, equity, and fairness. Moreover, these issues will become increasingly more important in the future, as the capabilities and adoption of FMs increase. Some of the fundamental questions that this workshop aims to address are:

- How can we identify and characterize unreliable and irresponsible behaviors in FMs? Topics include susceptibility to spurious features, prompt sensitivity, lack of self-consistency, and issues of nonfactuality or "hallucinations".
- How should we assess the potentially harmful capabilities of FMs and quantify their societal impact? For example, how can we predict the consequences of misuse of highly capable large language models?
- How can we pinpoint and understand the causes behind known or emerging sources of FM unreliability? This may involve examining training data, objectives, architectural design, learned weights, or other facets.
- What principles or guidelines should inform the design of the next generation of FMs to ensure they are both reliable and responsible?
- Can we establish theoretical frameworks that guarantee the reliability and responsibility of FMs?
- In practical applications, how might we 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 machine learning 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 for 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
- Empirical investigations into the reliability and responsibility of various FMs
- In-depth discussions exploring new dimensions of foundation model reliability and responsibility
- Interventions during pre-training to enhance the reliability and responsibility of FMs
- Innovations in fine-tuning processes to bolster the reliability and responsibility of FMs
- 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 with diverse backgrounds, affiliations, and areas of expertise have agreed to give invited talks at our workshop. Each speaker will bring a unique perspective to current developments in reliable and responsible FMs and related topics. We aim to provide titles for all talks prior to the event.

1. Andrew Wilson (New York University, male), distribution shift, uncertainty estimation (**confirmed to speak**)
2. Lilian Weng (OpenAI, female), safety and alignment in FMs (**confirmed to speak**)
3. Denny Zhou (Google DeepMind, male), reasoning, LLM theory (**confirmed to speak**)
4. Weijie Su (Stanford University, male), privacy, alignment (**confirmed to speak**)
5. Been Kim (Google DeepMind, female), human-centered learning, interpretability (**confirmed to speak**)
6. Nicolas Papernot (University of Toronto, male), AI safety, adversarial learning (**confirmed to speak**)
7. Mor Geva Pipek (Tel Aviv University, female), interpretability (**confirmed to speak**)
8. James Zou (Stanford University, male), bias and fairness (**confirmed to speak**)

Diversity Commitment

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

To further enhance the accessibility of our workshop to a broader audience, we also plan to offer a **registration fee grant** for those who might otherwise be unable to register. To support these initiatives, we will seek sponsorships from leading companies such as Google Deepmind, Meta AI, OpenAI, Facebook, Amazon, and Salesforce.

Our **review process** for submitted materials will be double-blind (conducted via OpenReview) to mitigate institutional and author biases. The program will be curated to ensure a wide representation of research areas while upholding the standards of quality set by the double-blind review process. Consequently, our workshop will benefit from a diverse cohort of participants, and we will invite several contributors to speak alongside our primary invitees.

We are launching a **reviewing mentorship program** designed to foster the growth and development of junior reviewers. Within this initiative, junior reviewers will be paired with senior reviewers. These mentor-mentee relationships aim to ensure that junior reviewers receive real-time feedback, guidance, and mentorship as they navigate the complexities of crafting insightful and constructive reviews for workshop submissions. By facilitating this collaborative and educational process, we hope not only to elevate the quality of reviews but also to cultivate the next generation of expert reviewers in the field.

Recognizing the challenges posed by **varying time zones** in a hybrid meeting format, we will incorporate a blend of synchronous and asynchronous activities to ensure wide participation. Specifically, we will ask both invited speakers and authors of accepted papers to provide pre-recorded videos in advance, enabling registered attendees to access the content flexibly. Live sessions, such as panel discussions and Q&A segments for invited talks or spotlights, will be facilitated through platforms like sli.do.

Tentative Schedule

This workshop will adopt a hybrid format. Specifically, our workshop will feature **eight 30-minute invited talks (comprising a 25-minute presentation followed by a 5-minute Q&A session), three 15-minute contributed talks selected from submissions, two 1-hour poster sessions, and a 1-hour panel discussion** to delve into the future of reliable and responsible FMs. The poster session can be attended either virtually or in person. Virtual attendees have the flexibility to choose a session that aligns with their time zone. Virtual poster sessions will be hosted in dedicated virtual spaces, such as Gather.Town. To enhance the experience for remote attendees, we will utilize a dedicated channel on a chat platform like Rocket.Chat to facilitate interactions among workshop participants.

Tentative Schedule of Paper Submission. We will follow the suggested dates by ICLR.

- Workshop paper submission deadline: February 3, 2024
- Workshop paper notification date: March 3, 2022
- Final workshop program, camera-ready, videos uploaded: April 3, 2024

Tentative Workshop Schedule.

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

Previous Related Workshops

In the past two years, several workshops have touched upon themes related to our focus, including the “2nd ICML Workshop on New Frontiers in Adversarial Machine Learning” (ICML 2023), “Workshop on Spurious Correlations, Invariance and Stability” (ICML 2022, 2023), “Workshop on Distribution Shifts” (NeurIPS 2022, 2023), “Socially Responsible Language Modelling Research” (NeurIPS 2023), “Trustworthy and Socially Responsible Machine Learning” (NeurIPS 2022), “Workshop on Machine Learning

Safety" (NeurIPS 2022), and "Pitfalls of Limited Data and Computation for Trustworthy ML" (ICLR 2023). While these workshops share some thematic overlap with ours, we'd like to emphasize two distinguishing features of our upcoming workshop: (1) Our workshop zeroes in on the reliability and responsibility challenges intrinsic to foundation models. The advent of foundation models ushers in unique issues, such as hallucinations and alignment problems, which demand specialized attention; (2) Instead of narrowing our lens to a singular facet of machine learning's reliability and responsibility, our workshop promotes a comprehensive dialogue. We seek to stimulate discussions on the reliability and responsibility of foundation models from various angles, including theoretical underpinnings, model architectures, and implications in real-world applications.

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 (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 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.

Zhun Deng (Columbia University)

- Email: zhun.d@columbia.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 postdoctoral researcher at Columbia University, and also part of Simons Collaboration on the Theory of Algorithmic Fairness. Previously, he completed his Ph.D. in the Theory of Computation group at Harvard University, advised by Cynthia Dwork. He is also fortunate to work with David Parkes, Weijie Su, and James Zou on various projects. His papers have won multiple honors such as Spotlight and Oral Presentation at flagship machine learning conferences, including ICML, NeurIPS, ICLR, and AISTATS.

Chelsea Finn (Stanford University)

- Email: cbfinn@cs.stanford.edu
- Webpage: https://ai.stanford.edu/ cbfinn/
- Google Scholar: https://scholar.google.com/citations?user=vfPE6hgAAAAJ

- Bio: Chelsea Finn is an Assistant Professor in Computer Science and Electrical Engineering at Stanford University, and the William George and Ida Mary Hoover Faculty Fellow. Her research interests lie in the capability of robots and other agents to develop broadly intelligent behavior through learning and interaction. To this end, her work has pioneered end-to-end deep learning methods for vision-based robotic manipulation, meta-learning algorithms for few-shot learning, and approaches for scaling robot learning to broad datasets. Her research has been recognized by awards such as the Sloan Fellowship, the IEEE RAS Early Academic Career Award, and the ACM doctoral dissertation award, and has been covered by various media outlets including the New York Times, Wired, and Bloomberg. Prior to Stanford, she received her Bachelor's degree in Electrical Engineering and Computer Science at MIT and her PhD in Computer Science at UC Berkeley. She has organized more than 10 workshops in ICML, NeurIPS, ICLR, and RSS.

Pavel Izmailov (OpenAI & New York University)

- Email: pavel@openai.com
- Webpage: izmailovpavel.github.io
- Google Scholar: https://scholar.google.ru/citations?user=AXxTpGUAAAAJ&hl=en
- Bio: Pavel Izmailov is a research scientist at OpenAI and an incoming assistant professor at New York University. Previously, he received his Ph.D. in Computer Science at New York University, working with Andrew Gordon Wilson. Pavel works on a wide variety of topics related to robustness of large-scale deep learning models: out-of-distribution generalization, interpretability, uncertainty estimation, Bayesian deep learning, and AI alignment. His work on Bayesian model selection was recognized with an outstanding paper award at ICML 2022.

He He (New York University)

- Email: hhe@nyu.edu
- Webpage: https://hhexiy.github.io/
- Google Scholar: https://scholar.google.com/citations?hl=en&user=K-isjagAAAAJ&view_op=list_works
- Bio: He He is an Assistant Professor of Computer Science and the Center for Data Science at New York University. She is also affiliated with the CILVR Lab, the Machine Learning for Language Group, and the Alignment Research Group. Before joining NYU, she spent a year at Amazon Web Services and was a postdoc at Stanford University. She received her PhD from the University of Maryland, College Park. Her research aims to (i) understand the computational foundation of generalization in novel scenarios, and (ii) build interactive systems that align with user's goals.

Pang Wei Koh (University of Washington)

- Email: pangwei@cs.washington.edu
- Webpage: https://koh.pw/
- Google Scholar: https://scholar.google.com/citations?user=Nn990CkAAAAJ&hl=en
- Bio: Pang Wei Koh is an assistant professor in the Allen School of Computer Science and Engineering at the University of Washington. His research interests are in the theory and practice of building reliable and interactive machine learning systems. His research has been published in Nature and Cell, featured in media outlets such as The New York Times and The Washington Post, and recognized by the MIT Technology Review Innovators Under 35 Asia Pacific award and best paper awards at ICML and KDD. He received his PhD and BS in Computer Science from Stanford University. Prior to his PhD, he was the 3rd employee and Director of Partnerships at Coursera. He has organized several sessions of NeurIPS Workshops on Distribution Shifts.

Eric Mitchell (Stanford University)

- Email: eric.mitchell@cs.stanford.edu
- Webpage: https://ericmitchell.ai/
- Google Scholar: https://scholar.google.com/citations?user=q77J4fgAAAAJ&hl=en
- Bio: Eric Mitchell is a final-year PhD candidate in Computer Science at Stanford University. His research interests focus on enhancing the safety and accessibility of foundation models, especially language models, His research has been notably recognized through the Knight-Hennessy Graduate Fellowship and the Stanford Accelerator for Learning grant. He also worked as a research scientist intern at DeepMind in 2022. Before embarking on his PhD journey, Eric contributed as a research engineer at Samsung's AI Center in NYC.

Cihang Xie (University of California Santa Cruz)

- Email: cixie@ucsc.edu
- Webpage: https://cihangxie.github.io/
- Google Scholar: https://scholar.google.com/citations?user=X3vVZPcAAAAJ&hl=en
- Bio: Cihang Xie is an Assistant Professor of Computer Science and Engineering at the University of California, Santa Cruz. His research straddles the domains of computer vision and machine learning, aspiring to create human-level computer vision systems. Cihang's particular focus is on ensuring robust model performance amidst distribution shifts and pioneering deep representation learning with limited supervision. He earned his Ph.D. from Johns Hopkins University under the mentorship of Bloomberg Distinguished Professor Alan Yuille. Cihang has enriched his research experience through internships, collaborating with experts like Kaiming He and Laurens van der Maaten at Facebook AI Research (FAIR), and Quoc Le at Google Brain. In recognition of his contributions, he was honored with the 2020 Facebook Fellowship.

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. Currently, focuses on both 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.