Workshop on “Secure and Trustworthy Large Language Models”

Workshop Summary:

The striding advances of large language models (LLMs) are revolutionizing many long-standing natural language processing tasks ranging from machine translation to question-answering and dialog systems. However, as LLMs are often built upon massive amounts of text data and subsequently applied in a variety of downstream tasks, building, deploying and operating LLMs entails profound security and trustworthiness challenges, which have attracted intensive research efforts in recent years. There are documented cases in which LLMs can

- Inherit pre-existing biases and exhibit discrimination against disadvantaged or marginalized social groups;
- Be vulnerable to security and privacy attacks that deceive the models or leak sensitive information of training data;
- Generate misinformation and make hard-to-justify predictions with a lack of transparency and interpretability;
- Be unreliable and unpredictable under domain shift or other input variations in critical scenarios.

The primary aim of the proposed workshop is to identify such emerging challenges, discuss novel solutions to address them, and explore new perspectives and constructive views across the full theory/algorithm/application stack. The potential topics include but are not limited to

- Reliability assurance and assessment of LLMs
- Privacy leakage issues of LLMs
- Copyright protection
- Interpretability of LLMs
- Plagiarism detection and prevention
- Security of LLM deployment
- Backdoor attacks and defenses in LLMs
- Adversarial attacks and defenses in LLMs
- Toxic speech detection and mitigation
- Challenges in new learning paradigms of LLMs (e.g., prompt engineering)
- Fact verification (e.g. hallucinated generation)

By bringing together experts from academia and industry, we aim to:

- Create a forum for discussing and identifying novel solutions and perspectives across the full theory/algorithm/application spectrum.
- Facilitate hands-on sessions and panels that enable participants to brainstorm, share insights, and collaboratively work on real-time solutions.
- Showcase state-of-the-art research and developments in each of these areas, promoting knowledge sharing and setting the groundwork for future research collaborations.

Modality: In-person

Tentative Schedule:

The workshop will include 8 invited talks (35 minutes + 5 minutes Q/A), 1 panel (1 hour), and 3 contributed talks (10 minutes) selected from the submitted papers.

To encourage discussions, we will arrange poster sessions before breaks to ensure participants have sufficient time to interact with each other. We will also hold a panel discussion to offer the opportunity for discussions among the invited speakers who are experts in different aspects of LLMs.

Additionally, during breaks, we will introduce structured breakout sessions, where participants will be divided into smaller groups to delve deeper into specific topics, thereby ensuring that every voice has an opportunity to be heard. To encourage the participation of early-career researchers and underrepresented groups, we will host a series of 'roundtable chats', where they can engage in focused discussions with senior researchers and peers.

Through these strategies, we aim to create an environment that not only values expert opinions but also recognizes the importance of diverse perspectives in shaping the future of the field.

Time Slot	Activity/Session	Duration	Details
9:00 - 9:40	Invited Talk 1	40 mins	35 mins talk + 5 mins Q/A
9:40 - 10:20	Invited Talk 2	40 mins	35 mins talk + 5 mins Q/A
10:20 - 11:00	Poster Session 1 & Breakout Session 1	40 mins	Interaction & focused discussions
11:00 - 11:40	Invited Talk 3	40 mins	35 mins talk + 5 mins Q/A
11:40 - 12:20	Invited Talk 4	40 mins	35 mins talk + 5 mins Q/A

12:20 - 1:00	Lunch Break & Poster Session 2	40 mins	Interaction & dining
1:00 - 1:40	Invited Talk 5	40 mins	35 mins talk + 5 mins Q/A
1:40 - 2:20	Invited Talk 6	40 mins	35 mins talk + 5 mins Q/A
2:20 - 2:40	Contributed Talk 1 & 2	20 mins	Selected from submitted papers
2:40 - 3:20	Poster Session 3 & Breakout Session 2	40 mins	Interaction & focused discussions
3:20 - 4:00	Invited Talk 7	40 mins	35 mins talk + 5 mins Q/A
4:00 - 4:40	Invited Talk 8	40 mins	35 mins talk + 5 mins Q/A
4:40 - 4:50	Contributed Talk 3	10 mins	Selected from submitted papers
4:50 - 5:50	Panel Discussion	1 hour	Topic TBD

5:50 - 6:30	Roundtable Chats & Closing Remarks	40 mins	Discussions with senior researchers & peers

Invited Speakers (Alphabetical Ordering):

A group of renowned speakers with diverse backgrounds (e.g., gender, race, affiliations, seniority, and nationality) is invited (with 7 confirmed).

Nicholas Carlini, Research Scientist, Google Brain (**confirmed**)
Tatsu Hashimoto, Assistant Professor, Stanford (**confirmed**)
He He, Assistant Professor, NYU
Cho-Jui Hsieh, Associate Professor, UCLA (**confirmed**)
Robin Jia, Assistant Professor, USC (**confirmed**)
Katherine A. Keith, Assistant Professor, Williams College
Bo Li, Associate Professor, University of Chicago (**confirmed**)
Graham Neubig, Associate Professor, CMU (**confirmed**)
Sameer Singh, Associate Professor, UCI
Zhou Yu, Associate Professor, Columbia (**confirmed**)

Audience:

Based on our experiences of organizing similar events previously, we expect 50 - 100 participants to attend this workshop. We have a list of participants who are highly likely to join this workshop. We will send out email invitations to these people as well as additional researchers who recently published papers on secure and trustworthy LLMs. Additionally, we also plan to advertise our workshop on social media such as Twitter.

Call for Papers (exclusivity of published work):

We are committed to upholding the integrity and uniqueness of the content presented in our workshop. To ensure this, our call for papers and participation will explicitly state that previously published work, especially from other machine learning conferences, is not acceptable for submission or presentation. Furthermore, any work that is being presented at the main ICLR conference, including those that might be part of invited talks, will also be excluded from our workshop. To implement and monitor this, we will:

- Explicit Communication: All calls for submissions and invited speakers' guidelines will contain clear instructions about the exclusivity criteria, ensuring contributors are aware before they even begin the submission process.
- Review Process: Our review committee will be briefed on the importance of this criterion. They will be tasked with checking the originality of the submissions against existing publications and the main ICLR conference schedule.
- Pre-workshop Verification: Before finalizing the workshop schedule, we will cross-verify the list of accepted papers and talks against the main ICLR conference program to rule out any overlaps.

- Feedback Mechanism: We will have a system in place where attendees and other participants can report if they find any content that violates this policy. Such reports will be taken seriously, investigated, and addressed promptly.

By instating these measures, we aim to maintain the workshop's novelty, ensuring it serves as a platform for fresh ideas and discussions, free from any redundancy with existing published work.

Review:

The review process for the submitted papers will be double-blind. Each submission will be reviewed by at least 3 reviewers, and we have attracted a wide and diverse program committee to guarantee this. To ensure unbiased assessments, no organizer will be involved in reviewing or evaluating any submission from a contributor who belongs to the same organization as the organizer. All reviewers and organizing committee members will be required to declare any potential conflicts of interest upfront, and appropriate reallocations of review assignments will be made to avoid any improprieties. Decisions on the acceptance will be made in a fair, and transparent manner. The selection criteria will be public to the authors.

Diversity Commitment:

In the spirit of fostering an inclusive, diverse, and forward-thinking academic community, our workshop has taken concrete steps to ensure representation in every facet of its organization and content delivery. We strongly believe that diversity fuels innovation, enriches discussions, and propels the academic field forward. Therefore, our commitment to diversity is not just about fulfilling quotas but about bringing together a multiplicity of voices, perspectives, and experiences that can contribute meaningfully to the discourse.

Diverse Viewpoints and Backgrounds:
Our list of invited speakers and organizers reflects a broad array of viewpoints and backgrounds. Representing institutions from coast to coast and beyond, our speakers hail from various stages of academic seniority and from multiple racial and ethnic backgrounds. We have consciously made efforts to include voices from underrepresented communities in academia.

Gender Diversity:
Our conscious effort to balance gender representation can be seen in the list of speakers and organizers. We have ensured that both male and female scholars are actively involved, contributing their unique perspectives to the discussions.

Geographic and Institutional Representation:
Our speakers and organizers come from diverse geographic locations and represent a mix of universities and research institutions. This geographic and institutional diversity ensures that multiple educational cultures and methodologies are represented.

Range of Scientific Seniority:
From assistant professors embarking on their academic journeys to established names in the field, our workshop brings together a range of scientific seniorities. This ensures that attendees benefit from both fresh perspectives and seasoned insights.

New and Returning Organizers:
While some of our organizers have had the experience of organizing similar events in the past, such as the “Workshop on Socially Responsible Machine Learning" at ICML 2021 and ICLR 2022, we have also ensured the

inclusion of those who have not organized in past editions. This brings a balance of experience and fresh vision to the organizing process.

Continual Evaluation:
While we are proud of the steps we have taken, we also understand that diversity and inclusion are continuous processes. We pledge to regularly reassess and refine our strategies to be even more inclusive in the future.

In conclusion, we believe that this intentional and proactive approach to diversity will not only make our workshop richer in content and discussions but will also set a positive precedent for future academic gatherings. We are dedicated to maintaining this commitment to diversity and will actively seek feedback to ensure we continue to meet and exceed these standards.

Access:

All the activities will be held in person. To facilitate participation, we will curate and maintain a website dedicated to this workshop and publish the accepted papers and talk abstracts on the website before the workshop. Talk information will be published 7 days prior to the workshop to allow participants to choose the contents they are most interested in. The recorded talks and the presented posters will be made available online afterward.

Previous and Related Events:

Several pertinent events have taken place at past editions of ICML and NeurIPS. Notably, the “ES-FoMo: Efficient Systems for Foundation Models” workshop at ICML 23 centered on the capabilities of large language models. However, our workshop takes a distinctive stance by spotlighting the challenges of security and trustworthiness intrinsic to the design, deployment, and utilization of these models.

While other workshops, such as the "Workshop on New Frontiers in Adversarial Machine Learning" at ICML 23, have broached security and trustworthiness concerns in a broader machine learning context, our session hones in on the nuances associated with large language models, often referred to as foundation models.

A closely related event is the “Socially Responsible Language Modelling Research” workshop from NeurIPS 23 that provided a holistic view of responsible LMs, our proposed workshop at ICLR 24 seeks to delve deeper into the technical intricacies of ensuring that LLMs are both secure and trustworthy. Given the rapid advancements in the domain, it's crucial for the research community to have frequent, focused forums to discuss, debate, and share solutions to emergent challenges. Our workshop will offer this platform, emphasizing the specialized challenges of security and trust in LLMs and providing an updated venue for the latest findings post-NeurIPS 23, as the few months between NeurIPS 23 and ICLR 24 can bring significant advancements, challenges, or insights in the domain. Moreover, hosting multiple workshops on related themes reinforces the importance of the topic in the research community. It sends a strong message about the significance of responsible, secure, and trustworthy development and deployment of language models. It can also motivate researchers to continue their efforts in this domain, knowing there are multiple prestigious venues to present and discuss their findings.

Organizers (Alphabetical Ordering)**:**

The workshop organizers consist of a diverse group of researchers hailing from both industry and academia, and they represent the entire range of scientific seniority, from postdoctoral researchers to assistant and full professors. Notably, some of the organizers possess rich experience in organizing similar events, such as the “Workshop on Socially Responsible Machine Learning" at ICML 2021 and ICLR 2022.

Anima Anandkumar, Bren Professor, Caltech & Nvidia
Jinghui Chen, Assistant Professor, Penn State
Nanyun Peng, Assistant Professor, UCLA
Yulia Tsvetkov, Assistant Professor, UW
Chaowei Xiao, Assistant Professor, University of Wisconsin, Madison
Ting Wang, Associate Professor, Stony Brook University
Yisen Wang, Assistant Professor, Peking University
Jieyu Zhao, Assistant Professor, USC

Organizer Resumes:

Anima Anandkumar (Caltech & Nvidia) holds dual positions in academia and industry. She is a Bren professor at Caltech CMS department and a director of machine learning research at NVIDIA. At NVIDIA, she is leading the research group that develops next-generation AI algorithms. At Caltech, she is the co-director of Dolcit and co-leads the AI4Science initiative, along with Yisong Yue. She has spearheaded the development of tensor algorithms, first proposed in her seminal paper. They are central to effectively processing multidimensional and multimodal data, and for achieving massive parallelism in large-scale AI applications. Prof. Anandkumar is the youngest named chair professor at Caltech, the highest honor the university bestows on individual faculty. She is the recipient of several awards such as the Alfred. P. Sloan Fellowship, NSF Career Award, Faculty fellowships from Microsoft, Google and Adobe, and Young Investigator Awards from the Army research office and Air Force office of sponsored research. She has been featured in documentaries and articles by PBS, wired magazine, MIT Technology review, yourstory, and Forbes. Anima received her B.Tech in Electrical Engineering from IIT Madras in 2004 and her PhD from Cornell University in 2009. She was a postdoctoral researcher at MIT from 2009 to 2010, visiting researcher at Microsoft Research New England in 2012 and 2014, assistant professor at U.C. Irvine between 2010 and 2016, associate professor at U.C. Irvine between 2016 and 2017, and principal scientist at Amazon Web Services between 2016 and 2018. Workshop organization experience:

- ICLR 2020 Diversity+Inclusion Chairs.
- Steering Committee NeurIPS workshop on Machine Learning and the Physical Sciences 2020
- IPAM workshop on Efficient Tensor Representations for Learning and Computational Complexity 2021
- ICLR workshop on Integration of Deep Neural Models and Differential Equations 2020
- Workshop Chair for ICML 2017
- NeurIPS workshop on Non-convex Optimization for Machine Learning: Theory and Practice 2016

Email: anima@caltech.edu
Homepage: http://tensorlab.cms.caltech.edu/users/anima/
Google Scholar: https://scholar.google.com/citations?hl=en&user=bEcLezcAAAAJ

Jinghui Chen (Penn State) is currently an Assistant Professor in the College of Information Science and Technology at Pennsylvania State University. He received a doctorate in computer science from the University of California, Los Angeles, during which he received the Outstanding Graduate Student Research Award. He is also the recipient of the Cisco Faculty Research Award. His research interests broadly include the theory and

applications in different aspects of machine learning, such as ML robustness, optimization, trustworthy and safety issues.
Email: jzc5917@psu.edu
Homepage: https://jinghuichen.github.io/
Google Scholar: https://scholar.google.com/citations?user=mKia7Y4AAAAJ&hl=en

Nanyun Peng (UCLA) is an Assistant Professor at the Computer Science Department, University of California, Los Angeles. Her research aims to build robust and generalizable Natural Language Processing (NLP) tools that lower the communication barriers and enable AI agents to become companions for humans. Her recent work has been focusing on several research topics, including creative language generation, low-resource information extraction, and zero-shot cross-lingual transfer. She received her Ph.D. in Computer Science at Johns Hopkins University, Center for Language and Speech Processing, as well as dual Bachelor's degrees in Linguistics and Economics. Workshop organization experience:

- Workshop on Creative AI Across Modalities, AAAI 2023

Email: violetpeng@cs.ucla.edu
Homepage: https://vnpeng.net/
Google Scholar: https://scholar.google.com/citations?user=XxRXvX0AAAAJ

Yulia Tsvetkov (UW) is an Assistant Professor at the Paul G. Allen School of Computer Science & Engineering at the University of Washington. Her research group works on computational ethics, multilingual NLP, and machine learning for NLP. This research is motivated by a unified goal: to extend the capabilities of human language technology beyond individual populations and across language boundaries, thereby making it available to all users. Prior to joining UW, Yulia was an assistant professor at Carnegie Mellon University and a postdoc at Stanford. Yulia is a recipient of NSF CAREER, Sloan Fellowship, Google Faculty and Amazon Machine Learning Research awards, and Okawa Research award. Workshop organization experience:

- MLCL 2016 – Workshop on Multilingual and Cross-lingual Methods in NLP
- SCLeM 2018 – Second Workshop on Subword and Character LEvel Models in NLP
- SafeConvAI at SIGDial 2021: A Special Session on Safety for E2E Conversational AI
- mmmPIE 2022 – Workshop on Performance and Interpretability Evaluations of Multimodal, Multipurpose, Massive-Scale Models

Email: yuliats@cs.washington.edu
Homepage: https://homes.cs.washington.edu/~yuliats/
Google Scholar: https://scholar.google.com/citations?user=SEDPkrsAAAAJ&hl=en

Ting Wang (Stony Brook University) is currently an Associate Professor and Empire Innovation Scholar in the Department of Computer Science at Stony Brook University. He holds a Ph.D. in Computer Science from the Georgia Institute of Technology. His recent research focuses on tackling the challenges in security assurance, privacy preservation, and decision-making transparency to unleash the full potential of technological advances in machine learning. His work has been widely published in top computer security and machine learning venues and recognized with multiple best (or runner-up) paper awards and media coverage. Workshop organization experience:

- Workshop on Interactive Mining for Big Data, CIKM 2014
- Workshop on Deployable Machine Learning for Security Defense, KDD 2020, 2021

Email: inbox.ting@gmail.com
Homepage: https://alps-lab.github.io/
Google Scholar: https://scholar.google.com/citations?user=cwcBTegAAAAJ&hl=en

Yisen Wang (PKU) is an Assistant Professor at Peking University. His research interest is broadly the representation learning from various types of data (unlabeled or noisy or adversarial data, structured data like graphs, etc.). Specifically, he recently focuses on theoretical and algorithmic approaches for trustworthy machine learning, self-supervised/weakly-supervised learning, and graph learning. He has received the Best Paper Award of ECML-PKDD 2021, and also achieved the 1st Place in the CVPR 2021 Adversarial Competitions and the Champion in the 2020 GeekPwn CAAD Competitions. Workshop organization experience:

- Workshop on Adversarial Robustness In the Real World, ICCV 2023

Email: yisen.wang@pku.edu.cn
Homepage: https://yisenwang.github.io/
Google Scholar: https://scholar.google.com/citations?user=uMWPDboAAAAJ&hl=en

Chaowei Xiao (Nvidia & ASU) is an assistant professor at the University of Wisconsin, Madison, and also a research scientist at NVIDIA. His research lies at the intersection of machine learning, privacy, and security. Workshop organization experience:

- Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems, CVPR 2019
- Workshop on Socially Responsible Machine Learning, ICML 2021
- Workshop on Neural Architecture Search, ICCV 2021

Email: cxiao34@wisc.edu
Homepage: https://xiaocw11.github.io/
Google Scholar: https://scholar.google.com/citations?user=Juoqtj8AAAAJ&hl=en

Jieyu Zhao is is an Assistant Professor of Computer Science Department at University of Southern California. She obtained her PhD from the department of Computer Science at UCLA. Her research interest lies in the fairness of ML/NLP models. Her paper got the EMNLP Best Long Paper Award (2017). She was one of the recipients of 2020 Microsoft PhD Fellowship and has been selected to participate in 2021 Rising Stars in EECS workshop. Her research has been covered by news media such as Wires, The Daily Mail and so on. She was invited by UN-WOMEN Beijing on a panel discussion about gender equality and social responsibility. Her workshop organization experience:

- Workflow Chair, AAAI 2023
- NLP for Social Good. ACL 2021, EMNLP 2022
- Women in Machine Learning (WiML). NeurIPS 2021
- Workshop on Socially Responsible Machine Learning, ICML 2021.

Email: jieyuz@usc.edu
Homepage: https://jyzhao.net
Google Scholar: https://scholar.google.com/citations?user=9VaGBCQAAAAJ&hl=en