Proposal: ICLR 2025 Workshop on Building Trust in Language Models and Applications

Abstract

As Large Language Models (LLMs) are rapidly adopted across diverse industries, concerns around their trustworthiness, safety, and ethical implications increasingly motivate academic research, industrial development, and legal innovation. LLMs are increasingly integrated into complex applications, where they must navigate challenges related to data privacy, regulatory compliance, and dynamic user interactions. These complex applications amplify the potential of LLMs to violate the trust of humans. Ensuring the trustworthiness of LLMs is paramount as they transition from standalone tools to integral components of real-world applications used by millions.

This workshop addresses the unique challenges posed by the deployment of LLMs, ranging from guardrails to explainability to regulation and beyond. The proposed workshop will bring together researchers and practitioners from academia and industry to explore cutting-edge solutions for improving the trustworthiness of LLMs and LLM-driven applications. The workshop will feature invited talks, a panel discussion, interactive breakout discussion sessions, and poster presentations, fostering rich dialogue and knowledge exchange. We aim to bridge the gap between foundational research and the practical challenges of deploying LLMs in trustworthy, use-centric systems.

https://soumyabratap.github.io/iclr-workshop/

Workshop motivation: Large language models (LLMs) are extensively and increasingly used in complex and transformative applications across industries such as healthcare, finance, legal advisory, education, and entertainment. The growing reliance on these models has surfaced significant challenges related to trustworthiness, safety, and ethical use, making the theme of building trust in language models both timely as well as vital. In practice, LLMs are often part of large pipelines with many moving parts, perhaps including retrieval, user interactions, code execution, API-use, or guardrails. Integrating LLMs into complex pipelines poses numerous additional challenges to trustworthiness. For example, retrieval capabilities may enable models to violate copyright laws or user interactions may expose models to new and unexpected data. As LLMs go from standalone use to full-fledged applications, ensuring trustworthiness is an urgent priority.

Workshop focus: While previous workshops have addressed an extensive list of LLM properties, including theory, security and privacy, as well as technical problems such as model optimization and fairness, there is a notable gap when it comes to the trustworthiness of LLM applications − such as AI assistants and co-pilots. Building such AI systems has become a widespread topic of academic and industrial research and also a central business strategy for organizations across numerous domains. Meanwhile, there has been little effort to bring together researchers and practitioners working on the unique scientific challenges and systems aspects of developing such LLM-driven systems. This practical orientation sets our proposed workshop apart by focusing on:

- **Deployed systems.** We focus on research that addresses the types of challenges faced when deploying LLMs in real-world systems, for example ones used by millions of users. Such challenges include ensuring reliability and scalability, as well as maintaining performance under diverse and sometimes unpredictable conditions.
- **User interactions.** The way users interact with LLM applications significantly impacts trust. We explore research on feedback loops and interaction models that enhance user trust and satisfaction.

Workshop Proposal for International Conference on Learning Representations (ICLR 2025).

- **Context-specific challenges.** Often applications must navigate industry specific regulations and ensure compliance with laws and standards such as HIPAA in healthcare, GDPR in data privacy, and financial regulations.

While research on the above topics may have historically been limited to industry, a large academic community now develops full-fledged AI agents equipped with retrieval or tool-use capabilities, studies human-LLM interactions, or proposes legal or regulatory frameworks for ensuring safety. Therefore, a workshop on LLM trustworthiness that extends beyond the models themselves is timely and will engage a large audience at ICLR. Moreover, we encourage research not just on full-fledged systems but also on individual LLM behaviors, for example bias or unreliability, which may surface in applications.

Workshop scope and example topics: We consider a broad range of topics covering all aspects of safely deploying Large Language Models (LLMs). A non-comprehensive list of relevant topics includes:

- Error detection and correction during content and code generation,
- Improving reliability and truthfulness of LLMs (mitigating hallucination, miscalibration of responses, distribution shifts, etc.)
- Explainability and interpretability of LLM responses,
- Robustness of LLMs (including prompt attacks, interventional effects, etc.)
- Unlearning techniques for LLMs
- Fairness of LLMs (including social bias, stereotype bias, preference bias, etc.)
- Guardrails and regulations for LLMs (including toxicity detection, avoiding offensive/insensitive language, emotional awareness, etc.)
- Metrics, benchmarks, and evaluation of trustworthy LLMs
- Practical challenges to build trust in real-world AI applications

Workshop Format: The workshop will feature five main components, introducing new ideas and engaging the community in discourse about trust in LLM models and applications:

- **Invited talks** from a diverse group of junior and senior invited speakers, followed by Q&A sessions to better engage the audience.
- **Break-out discussion session (sponsored by Adobe)** where small groups of attendees will discuss current and future directions in trustworthy LLMs. We will hand out initial prompts to kick off discussion, and we will shuffle discussion groups halfway through the discussion period. We will also provide coffee and pastries during this period, sponsored by Adobe, to attract a broad audience for lively discussions and also to help bolster attendance for the panel session afterwards.
- **Panel discussion** where panelists from diverse backgrounds and expertise will discuss the scope, future directions, and limitations of trustworthiness of LLMs.
- **Poster session** will enable researchers to present their work interactively, allowing attendees to discuss the posters with the authors and fostering networking and idea exchange.
- **Contributed talks** where exceptional submissions and competition winners will present their work in showcasing cutting-edge research and practical solutions.

Confirmed invited speakers:

- Yejin Choi (University of Washington)
- Andy Zou (Carnegie Mellon University)
- Sandra Wachter (University of Oxford)
- Avijit Ghosh (Hugging Face)
- Reza Shokri (National University of Singapore)
- Tom Goldstein (University of Maryland)
- John Dickerson (Arthur AI)

Confirmed panelists:

- Nicholas Carlini (Google DeepMind)
- Elham Tabassi (National Institute of Standards and Technology (NIST))
- Zack Lipton (Carnegie Mellon University)

Workshop Schedule:

8:30am-8:50am	Informal Coffee Social
8:50am-9:00am	Introduction and opening remarks
9:00am-9:40am	Invited Talk 1 (Andy Zou) and Q&A
9:40am-10:00am	Invited Talk 2 (Sandra Wachter) and Q&A
10:00am-10:40am	Invited Talk 3 (Tom Goldstein) and Q&A
10:40am-11:00am	Coffee Break
11:00am-11:45am	Contributed Talks
11:45am-12:30pm	Break-Out Discussion Session
12:30am-1:30pm	Lunch Break
1:30am-3:00pm	Poster Session
3:00pm-3:40pm	Invited Talk 4 (Avijit Ghosh) and Q&A
3:40pm-4:00pm	Invited Talk 5 (Reza Shokri) and Q&A
4:00pm-4:40pm	Panel Discussion
4:40pm-5:00pm	Invited Talk 6 (John Dickerson) and Q&A
5:00pm-5:40pm	Invited Talk 7 (Yejin Choi) and Q&A
5:40pm-6:00pm	Closing Notes

Anticipated audience size. Attendance at the ICLR 2024 workshops on *Reliable and Responsible Foundation Models* and *Secure and Trustworthy Large Language Models* ranged from approximately 250 to over 300 attendees. Based on these figures, we conservatively estimate a minimum of 300 participants for our upcoming workshop. This projection is supported by the fact that ICLR is growing each year. Additionally, we expect around 130 paper submissions to our workshop, which is aligned with the numbers for similar workshops held at ICLR 2024.

Technical requirements. To accommodate our workshop participants, we need a spacious room equipped with a projector, adequate seating, and areas for displaying posters, including poster boards.

Funding. Refreshments will be provided courtesy of Adobe. Upon acceptance, we will solicit sponsorships from companies such as OpenAI, Microsoft, and Google DeepMind. These funds will be used for awards (best poster and talk) and to create student travel grants.

Plan to get an audience for a workshop (advertising, reaching out, etc.) Several of the organizers, speakers, and panelists have large social media followings, so we will design a catchy graphical flyer and then make a concerted effort to post the flyer on social media in three coordinated pushes: (1) when we first announce the workshop; (2) promoting the call for papers and deadline; (3) leading up to the workshop itself. Those of us on the organizing committee as well as speakers and panelists who are in academia will also circulate these announcements on our university listservs and Slack workspaces, especially ones related to minority and underrepresented groups in AI, encouraging students to submit papers and attend.

Submission timeline and logistics. Calls for (non-archival) submissions will be made in December, with a submission deadline on February 3, 2025 as suggested by the conference. Papers should discuss research or deployed work. Submissions will be solicited in two tracks: **(a)** *full papers* will be at most 8 pages long; **(b)** *short papers* will be at most 4 pages. Full paper submissions should

discuss research or deployed work, while short papers may contain proposals or plans for research projects or small-scale follow-ups and interpretations of existing research. The short paper track will encourage disadvantaged and under-resourced groups to participate. We will only consider previously unpublished works that will not appear at the main ICLR conference, and will specify this in the call for papers. Each submission will undergo double-blind review from multiple reviewers, and one meta-reviewer, using CMT; reviewer conflicts of interest will be appropriately managed via CMT. We will notify authors by March 5, 2025, and abide by any centralized ICLR deadlines.

Managing conflicts of interest. Workshop program committee members will not assess submissions from individuals within the same organization, collaborators, or those they personally know. We will use CMT conflict management functionality to mitigate this risk. Organizers will not act as speakers or panelists.

The review process. Our review process for submitted materials will be double-blind (conducted via CMT) 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 will implement a reviewing mentorship program, inspired by several previous workshops, 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.

Diversity commitment. A commitment to diversity is especially crucial in the field of AI safety, where the fears and AI use-cases of different communities vary. Our workshop is committed to creating a diverse and inclusive environment, as evidenced by our lists of organizers, speakers, and panelists which are diverse in terms of gender, ethnicity, geographic location, seniority, and affiliation. By bringing together a wide range of perspectives, we aim to address the multifaceted challenges in trustworthy AI more effectively. Our organizer, speaker, and panelist rosters include 5 women, researchers from and living in three continents, those at a variety of seniority levels, and many people from both academia and industry. We will further promote these kinds of diversity when advertising our workshop and implementing accessibility features.

Previous related workshops. Building trust in language models is a critical and complex challenge that spans a wide range of topics, from safety and reliability to fairness and regulation, across applications. While many workshops have focused on security, privacy, fairness, or trustworthiness in machine learning models, previous workshops either focus broadly on AI, on specific applications, or on models themselves rather than challenges with incorporating models into bigger pipelines. Our proposed workshop stands out by concentrating on language models and pipelines that incorporate language models, which have their own specific pitfalls that may not apply to other types of models and applications, as well as bringing together diverse perspectives on building trust. We will foster discussions on methodological innovations, practical challenges, and evaluation metrics across domains, covering key aspects such as safety, interpretability, unlearning, fairness, and the implementation of guardrails. By uniting researchers and practitioners from different fields, this workshop will generate a clearer understanding of how to build trustworthy LLMs and LLM applications. The following discusses related workshops.

1. Responsible LMs: What's left to TEACH (Trustworthy, Enhanced, Adaptable, Capable and Human-centric) chatbots? (ICML 2023), Secure and Trustworthy Large Language Models (ICLR 2024), Socially Responsible Language Modelling Research (NeurIPS 2024). These are the most relevant workshops. Unlike these workshops, we consider aspects of trust that extend beyond the model itself and which may be a part of applications or product pipelines of which the language model only plays a limited role. For example, privacy concerns could arise when incorporating retrieval capabilities, or security breaches may be enabled by allowing an LLM to execute code.
2. Trustworthy Machine Learning: Trustworthy and Socially Responsible Machine Learning (NeurIPS 2022), Socially Responsible Machine Learning (ICLR 2022), Trustworthy and Reliable Large-Scale Machine Learning Models (ICLR 2023), Reliable and Responsible Foundation

Models (ICLR 2024). While present in the topics of these workshops, large language models are not the primary focus. Our workshop centers on LLMs and LLM applications specifically, addressing their unique challenges in building trust.

3. Topics adjacent to trust: Several previous workshops focus on an adjacent topic or a subcategory of trust. For example, AI Safety: Next Generation of AI Safety (ICML 2024), Safe Generative AI Workshop (NeurIPS 2024), Red Teaming GenAI: What Can We Learn from Adversaries? (NeurIPS 2024); Interpretability: Mechanistic Interpretability Workshop (ICML 2024), XAI in Action: Past, Present, and Future Applications (NeurIPS 2023), Interpretable AI: Past, Present and Future (NeurIPS 2024); Fairness: Algorithmic Fairness through the Lens of Metrics and Evaluation (NeurIPS 2024). Our proposed workshop brings together a comprehensive view, addressing multiple facets of trust in language models and fostering cross-disciplinary discussions to tackle these challenges holistically, and again extends to applications that leverage LLMs but where trust concerns may creep in from other components of a pipeline.
4. A specific application: Some workshops focus on trustworthy ML in just one narrow application area such as healthcare and law. For example, trustworthy ML in healthcare: Interpretable Machine Learning in Healthcare (ICML 2023), Trustworthy Machine Learning for Healthcare (ICLR 2023). GenAI in Law: Generative AI + Law (ICML 2024). However, in this proposal, our focus is general by considering any application area. Thus, our workshop stands out by bringing together insights from diverse fields and fostering discussions across different application areas.

Correspondence: Ramasuri Narayanam rnarayanam@adobe.com, Martin Pawelczyk martin.pawelczyk.1@gmail.com

Organizers

MICAH GOLDBLUM, COLUMBIA UNIVERSITY (Google Scholar) mig2132@columbia.edu
Micah is an assistant professor at Columbia University Department of Electrical Engineering. His research focuses on AI safety and privacy, including LLM watermarking, uncertainty estimation for language models, and on detecting material that generative models copy from their training sets. In 2022, Micah received the ICML Outstanding Paper Award for his work on the marginal likelihood and model selection. Before his current position, he was a postdoctoral research fellow at New York University with Yann LeCun and Andrew Gordon Wilson. Micah was the chair of the organizing committee for the NeurIPS 2020 Workshop on Dataset Curation and Security and served on the organizing committee for the NeurIPS 2023 Workshop on Backdoors in Deep Learning, the NeurIPS 2024 Workshop Red Teaming GenAI, and the NeurIPS 2024 workshop on Scientific Methods for Understanding Deep Learning.

RAMASURI NARAYANAM, ADOBE RESEARCH (Google Scholar) rnarayanam@adobe.com
Ramasuri is working as a Senior Research Scientist at Adobe Research, India since 2021. His current research focus is on building deployable AI systems, computational game theory, and data analytics. Prior to this, he worked at IBM Research - India as a Senior Research Scientist and an IBM Master Inventor for about 10 years. Overall, he has about 14 years of experience in developing and delivering Artificial Intelligence (AI) driven novel analytics solutions and technologies to tackle business problems arising in sectors such as telcos, retail, and in technology domains such as social media, business workflow analytics, software log analytics, etc. During this tenure, Ramasuri has built an IP portfolio of over 35 peer-reviewed research articles in several premier CS venues and over 40 granted patents in USPTO. He is an IEEE Senior Member, ACM Distinguished Speaker, and recipient of several awards including IBM's prestigious global Outstanding Technical Achievement Award for 4 times and Best Ph.D. Thesis award from Computer Society of India. He serves as Senior Program Committee (SPC) Member in IJCAI and SPC-AreaChair in AAMAS. He offered about 5 tutorials at premier CS conference venues such as WWW, AAAI, etc. He also organized about 3 workshops on Machine learning and Compurational social networks.

MARTIN PAWELCZYK, HARVARD UNIVERSITY (Google Scholar) martin.pawelczyk.1@gmail.com
He is a Postdoc at Harvard University where he is working with Himabindu Lakkaraju and Seth Neel. His research is focused on machine learning safety, including topics such as unlearning, robustness, privacy and interpretability. He completed his PhD at the University of Tübingen with Gjergji Kasneci, studied Statistics at the London School of Economics (LSE) and Econometrics (Applied Statistics) at the University of Edinburgh.

BANG AN, UNIVERSITY OF MARYLAND (Google Scholar) bangan@umd.edu
Bang is a fifth-year PhD candidate in Computer Science at the University of Maryland, College Park, advised by Dr. Furong Huang. Prior to joining UMD, she worked as a research scientist at IBM Research, China. Her research centers on Responsible AI, particularly enhancing the safety, alignment, robustness, fairness, and interpretability of Generative AI systems. Recently, her work has focused on automatic red-teaming for LLM safety and false refusals, detecting AI-generated content, copyright issues, and improving test-time alignment. Bang has contributed to program committees for workshops, including the Trustworthy and Socially Responsible Machine Learning Workshop at NeurIPS'22 and the NextGenAISafety Workshop at ICML'24. She is also leading the organization of NeurIPS'24 competition, Erasing the Invisible: A Stress-Test Challenge for Image Watermarks.

SOUMYABRATA PAL, ADOBE RESEARCH (Google Scholar) soumyabratap@adobe.com
Soumyabrata Pal is a Research Scientist at Adobe Research, Bengaluru, India. Prior to this, he was a postdoctoral researcher in Google Research, Bengaluru, India. He graduated with a PhD from the College of Information and Computer Sciences at the University of Massachusetts Amherst, advised by Professor Arya Mazumdar. Before coming to Amherst, he obtained his B.Tech from Indian Institute of Technology Kharagpur in India in 2016. His research interests are LLM Efficiency and Theoretical Machine Learning focused on Non-convex Optimization and Online Learning. He has published widely in top ML conferences such as NeurIPS, ICML, AISTATS and in reputed journals such as IEEE Transactions on Information Theory (TIT) and Journal of Machine Learning

Research (JMLR).

HIMABINDU LAKKARAJU, HARVARD UNIVERSITY (Google Scholar) hlakkaraju@hbs.edu
She is an Assistant Professor at Harvard University with appointments in the Business School and the Department of Computer Science. Prior to her stint at Harvard, she received her PhD in Computer Science from Stanford University. Her research interests lie within the broad area of trustworthy machine learning. More specifically, she focuses on improving the interpretability, fairness, privacy, robustness, and reasoning capabilities of different kinds of ML models, including large language models and other pre-trained models.

SHIV KUMAR SAINI, ADOBE RESEARCH (Google Scholar) shsaini@adobe.com
Shiv is a Principal Research Scientist at Adobe Research - India. He works in the areas of time series modelling and causal inference in observational data. He has applied these tools for modelling user behavior, marketing attribution root cause analysis, and anomaly detection. His recent research focuses on developing techniques to improve reliability of micro services by forecasting future outages and reducing time to detect root cause of ongoing outages. Prior to joining industry research, he obtained his Ph.D. from University of Wisconsin-Madison in 2008.

Program Committee Members

Aarshvi Gajjar, Abhimanyu Hans, Aditya Kumar, Alex Stein, Alicia Sagae, Ameya Joshi, Aniruddha Saha, Anna Sotnikova, Anne Josiane Kouam, Antoni Kowalczuk, Arpit Bansal, Avi Schwarzschild, Ayesha Ansar, Ben Feuer, Bihe Zhao, Charvi Rastogi, Chejian Xu, Chirag Nagpal, Chulin Xie, CJ Lee, Daniel Arp, Dariush Wahdany, David Beste, David Pape, Dongxian Wu, Emily Black, Emily Wenger, Felix Weißberg, Fengqing Jiang, Feyza Duman, Gauri Jagatap, Govind Mittal, Grace Kim, Hamid Kazemi, Hojjat Aghakhani, Hossein Hajipour, Hossein Souri, Jan Dubiński, Jing Xu, Joel Frank, John Kitchenbauer, Jonas Möller, Jonas Ricker, Jonathan Evertz, Kelly Marshall, Kezhi Kong, Khalid Saifullah, Lea Schönher, Liu Leqi, Manli Shu, Martin Bertran Lopez, Mathew Monfort, Mayuka Jayawardhana, Michael Feffer, Michel Meintz, Minh Pham, Minsu Cho, Mintong Kang, Mohammadreza Soltani, Naren Dhyani, Nari Johnson, Neel Jain, Nupur Kulkarni, Olatunji Iyiola Emmanuel, Patrick Yubeaton, Pranamesh Chakraborty, Pratyush Maini, Pruthuvi Maheshakya Wijewardena, Qi Zhang, Renkun Ni, Riccardo Fogliato, Roman Levin, Rongting Zhang, Sahar Abdelnabi, Sahil Verma, Sai Pranaswi, Sandeep Avula, Shahrzad Kianidehkordi, Shantanu Gupta, Shawn Shan, Shreya Agarwal, Sina Däubener, Srishti Gupta, Sudipta Banerjee, Thanh Nguyen, Thorsten Eisenhofer, Tianqi Du, Tom Blanchard, Vasu Singla, Vedant Nanda, Vijay Keswani, Vincent Hanke, Wenhao Wang, Xun Wang, Zeming Wei, Yi Zheng, Yichuan Mo, Zhangchen Xu, Zhijian Zhuo, Zuowen Yuan