Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI

uncertainty-foundation-models.github.io

Modality: In-person.
Expected Audience: 300 participants.
Related topics: Uncertainty in representation learning, Transparency, Generative AI and large language models.

1 Summary

How can we trust large language models (LLMs) when they generate text with confidence, but sometimes hallucinate or fail to recognize their own limitations? As foundation models like LLMs and multimodal systems become pervasive across high-stakes domains—from healthcare and law to autonomous systems—the need for uncertainty quantification (UQ) is more critical than ever. Uncertainty quantification provides a measure of how much confidence a model has in its predictions or generations, allowing users to assess when to trust the outputs and when human oversight may be needed.

Importance and novelty of the problem. Existing UQ methods have been developed primarily for discriminative models—those used for classification or regression tasks. These methods, while effective for tasks such as image classification or binary decision-making, do not translate well to the autoregressive models that underpin LLMs (Brown et al., 2020; Thoppilan et al., 2022; Touvron et al., 2023). Autoregressive models generate tokens sequentially, where each output depends on the previous one. This structure introduces unique challenges for uncertainty quantification. First, in autoregressive models, uncertainty compounds as tokens are generated, making it difficult to capture where the model's confidence falters over the sequence. Moreover, LLMs dynamically adapt to preceding context, leading to shifts in uncertainty as the model progresses through a text or multimodal sequence. For LLMs handling both text and image modalities, such as GPT-4, uncertainty quantification becomes even more complex due to the multimodal nature of the input and output spaces. Traditional UQ methods struggle to account for cross-modal interactions where uncertainty in one modality (e.g., image understanding) affects the other (e.g., text generation). Another distinct challenge with LLMs is the limited precision of human feedback. Existing techniques for calibrating discriminative models rely on access to ground truth labels. In contrast, with LLMs there are no true labels and one only has access to relative preferences of a handful of generations. Furthermore, there is a growing need for theoretical frameworks that can effectively analyze and predict model behavior in out-of-distribution (OOD) scenarios, where uncertainty is critical for identifying when a model's generation are likely to be unreliable. *To address these challenges, UQ methods must shift from traditional approaches tailored to discriminative models and embrace new techniques that account for the complex dependencies and dynamic nature of autoregressive models*.

Topics and key questions. This workshop aims to close this gap by defining, evaluating, and understanding the implications of uncertainty quantification for autoregressive models and large-scale foundation models. We invite researchers across machine learning, statistics, cognitive science, and human-computer interaction to contribute through invited talks, contributed papers, and structured discussions on key questions and topics:

- How can we develop scalable and computationally efficient methods for uncertainty estimation in LLMs?
- What are the theoretical foundations for understanding and quantifying different types of uncertainty (epistemic, aleatoric, and out-of-distribution) in generative models?
- How can we effectively detect and mitigate hallucination in generative models while maintaining their creative capabilities?
- Can we develop UQ methods to assess the reliability of multimodal models where uncertainty in one modality affects others (e.g., text and image in LLMs)?

- What are the best practices for communicating model uncertainty to different stakeholders, from technical experts to end users?
- What practical and realistic benchmarks and datasets can be established to evaluate uncertainty quantification techniques for LLMs and multi-modal foundation models?
- How can uncertainty estimates guide decision-making under risk, and what frameworks can be developed to incorporate UQ into real-world AI systems, ensuring safer and more reliable deployment?

Societal impact. The impact of uncertainty estimation extends far beyond scientific curiosity, playing a crucial role in the development of trustworthy and reliable AI systems. As foundation models like LLMs become embedded in high-stakes decision-making processes, policymakers worldwide are increasingly focused on transparency and reliability, crafting frameworks for governance that emphasize these principles. Reliable uncertainty quantification is becoming essential for compliance with emerging regulations, particularly in domains like healthcare, law, and autonomous systems, where the stakes are highest. Accurate uncertainty estimation is key not only to making foundation models more trustworthy but also to building systems that are aligned with societal expectations. By addressing these challenges, AI practitioners will be better equipped to develop safer, more reliable models while improving risk management in critical applications. Ultimately, the advancement of uncertainty quantification will provide the tools necessary to meet regulatory demands and ensure that AI systems are aligned with both ethical standards and practical requirements in real-world environments.

2 Invited Speakers

The list of invited speakers was selected based on their prior work on the topic. All speakers have published high-impact works on machine learning. All the speakers have **confirmed** their interest and ability to give the invited talk in person at the workshop.

Mihaela van der Schaar **(University of Cambridge)** is the John Humphrey Plummer Professor at the University of Cambridge. Mihaela is founder and director of the Cambridge Centre for AI in Medicine (CCAIM). Mihaela was elected IEEE Fellow in 2009 and Fellow of the Royal Society in 2024. She has received numerous awards, including the Johann Anton Merck Award (2024), the Oon Prize on Preventative Medicine from the University of Cambridge (2018), a National Science Foundation CAREER Award (2004), 3 IBM Faculty Awards, the IBM Exploratory Stream Analytics Innovation Award, the Philips Make a Difference Award and several best paper awards, including the IEEE Darlington Award. She was a Turing Fellow at The Alan Turing Institute in London between 2016 and 2024.

Yisong Yue **(Caltech)** is a professor of Computing and Mathematical Sciences at the California Institute of Technology. Yisong's research interests lie primarily in machine learning, and span the entire theory-to-application spectrum from foundational advances all the way to deployment in real systems. Yison has worked extensively on uncertainty prediction, including his recent work using conformal calibration. Yisong serves as the General Chair at ICLR 2025. Previous, Yisong was the Senior Program Chair of ICLR 2024.

Jie Ren **(Google DeepMind)** is a Senior Research Scientist at Google DeepMind (formerly Google Brain). She holds a PhD in Computational Biology and Bioinformatics and an MSc in Statistics, from the University of Southern California. Jie's research centers on developing trustworthy AI solutions that can be safely deployed in real-world scenarios, aiming to advance scientific discoveries and enhance human well-being. Her work spans key areas including uncertainty estimation and robustness in large foundation models, out-of-distribution detection and robustness in deep learning, and the development of reliable machine learning for real-world application, with a special focus on biological and medical research.

Stefano Ermon **(Stanford University)** an Associate Professor at Stanford University. Stefano is a fellow of the Woods Institute for the Environment. His research lies in machine learning and generative AI. Stefano has received outstanding or best paper awards (in ICML'24, ICLR'21) or best paper award nominations (CVPR'23) for his work on generative models.

Besides the confirmed speakers, there is a list of tentative speakers that we can invite:

- Geoff Pleiss (University of British Columbia, Bayesian optimization).
- Csaba Szepesvári (University of Alberta and team lead at Deepmind, sequential decision making, uncertainty).

start	duration	event	theme
9:00	0:10	opening remarks	
9:10	0:45	invited talk: Yisong Yue	Uncertainty in generative models
9:55	0:10	discussion + coffee	
10:05	0:45	invited talk: Stefano Ermon	
10:50	0:10	discussion + coffee	
11:00	1:45	poster session I	
12:45	1:30	lunch	
13:30	0:45	invited talk: Mihaela van der Shaar	Uncertainty and hallucination in critical applications
14:15	0:10	discussion + coffee	
14:25	0:45	invited talk: Jie Ren	
15:10	0:10	discussion + refreshments	
15:20	1:45	poster session II	
17:05	0:10	discussion + coffee	
17:15	0:45	panel	
16:00	0:10	closing remarks	
18:10		End of workshop	

- Graham Neubig (Carnegie Mellon University, question answering, code generation, hallucination of LLMs).
- Fabio Cuzzolin (Oxford Brookes University, uncertainty theory and belief functions).

3 Schedule

This will be a **full-day workshop** featuring a combination of invited talks, spotlight presentations of selected accepted papers, and poster sessions for all accepted works. Each accepted paper will include a brief recorded talk (5 minutes) that will be introduced during the breaks. We will organize two poster sessions, giving all participants with accepted papers the opportunity to discuss their work in one or both sessions, depending on the number of submissions. The extensive time allocated to poster sessions will hopefully enable the young, aspiring researchers communicate their work to a broader audience, while more established researchers will gain fresh insights on new ideas. The tentative schedule is detailed below:

Following each invited talk, we will have a 10-minute break to encourage interaction and idea exchange among workshop participants. The final schedule will be adjusted according to the ICLR timeline to align with the conference's schedule, including coffee breaks.

Panel discussion: The panel discussion in the afternoon will include both experienced researchers and rising stars, who along with the organizers will summarize the insights from the workshop and discuss open questions on uncertainty and hallucination in the era of foundation models. Some critical questions related to the theme are the following:

- How can young researchers approach UQ on models that are expensive to even make prediction on?
- Are there essential mathematical tools for analyzing UQ in current foundation models that have not yet been created?
- How is out-of-distribution quantified in an era that we do not have access to the training data and how does this impact UQ?
- Are there some scaling laws for UQ as the model becomes bigger? Is there any theoretical reasoning for those laws?

4 Audience and access

Audience: The topic of uncertainty estimation and hallucinations in foundation models has received increasing attention, with numerous papers accepted at ICLR, ICML, and NeurIPS over the past two years. Consequently, we anticipate substantial interest from the community in a dedicated workshop on this subject. We expect participation from both early career researchers and students, as well as established researchers seeking fresh insights into this emerging field. Therefore, we estimate approximately 300 participants for the workshop.

Website: We have established a website at https://uncertainty-foundation-models.github.io. This site will provide all necessary information regarding submissions and workshop attendance. Additionally, we will upload all accepted papers at least one week prior to the workshop, allowing participants to review the papers in advance.

Access: Following ICLR guidelines, our primary focus will be on in-person attendance. However, to accommodate extenuating circumstances such as visa issues or other rare exceptions, we will allow online participation for people that have a valid reason. We aim to upload all presentations, accepted papers and other materials in the aforementioned website that will enable timely access to all participants.

5 Related Workshops

There has been **no related workshop thematically in ICLR the last 5 years**. The closest workshops have been on the topic of domain generalization, i.e., *'What do we need for successful domain generalization?'* (ICLR'23) and *'Generalization beyond the training distribution in brains and machines'* (ICLR'21). However, both of these workshops, do not focus on uncertainty, or even more precisely uncertainty on the era of large models.

Beyond ICLR, in NeurIPS and ICML there have been no thematic workshops on uncertainty in the era of (large) foundation models. The closest workshops are the workshops on uncertainty on ICML'22 (*'Workshop on Distribution-Free Uncertainty Quantification'*) and ICML'21 (*'Workshop on Distribution-Free Uncertainty Quantification'*). The difference is that these workshop focused on theoretical perspectives, e.g., conformal prediction, and smaller scale models, while new ideas are required for uncertainty prediction in the era of foundation models. In addition, the aforementioned workshop did not focus on hallucinations as observed in foundation models, which pose a challenge for any application.

In NeurIPS'24, the workshop titled *'Workshop on Statistical Frontiers in LLMs and Foundation Models'* aims to provide a framework for the statistical analysis of LLMs, including conformal prediction. This workshop focuses on the broader topic of statistical analysis and does not specifically focus in neither UQ or hallucination of foundation models. As such, we do not consider there is a significant overlap with the proposed content.

6 Diversity Statement

Diversity of organizing committee. The organizing committee is made up of researchers from various backgrounds, aiming to promote diversity in gender, seniority, organizational experience, and affiliation. The core organizers are located across three continents and include one professor, three assistant professors, one senior research scientist, and one senior PhD student. We have recruited four volunteers to assist with the workshop. The committee includes four women researchers, underscoring our commitment to gender equity. Our organization team consists of researchers from diverse nationalities, including the United States, Europe, and Japan. The organizers represent a broad spectrum of academic backgrounds, including universities, research institutes, and industry. We are committed to ensuring that the workshop is inclusive and accessible to participants from diverse communities and backgrounds, fostering an environment where various perspectives are represented.

Diversity of invited speakers. The invited speakers span both experienced researchers and rising stars, while they have theoretical backgrounds but make contributions in both theoretical and more empirical domains. At the core of their research interests lie building trustworthy systems that will be reliable in real-world applications.

Diversity of the topics. The workshop aims to encourage interdisciplinary collaboration, facilitating interaction between applied mathematics, statistics, computer science, signal processing, and machine learning, particularly in their applications. This approach will foster new connections and insights across these fields.

7 Papers and review

Call for papers: There are three types of submissions that we will primarily rely on:

- Research papers presenting new findings on the specified topics.

- Research papers that have been admitted the last few months to top-tier ML conferences.
- Tiny papers track that prioritize shorter submissions that can benefit from feedback during the workshop.

The format for the first two types of submission will be limited to a maximum of 4 pages, not including references. On exceptional circumstances (e.g., the findings cannot be summarized in 4 pages and the paper is already accepted in a top-tier conference) we will enable more pages of content. In addition, the papers will be non-archival, enabling the participants to submit the accepted papers to conferences or journals. This is consistent with the policy in the majority of ML workshops we are aware of.

We will actively encourage new works to be presented in the workshop. We will impose a max of 25% (over all accepted contributions) to have been previously accepted in top-tier conferences. That is, we will limit the number of papers that are already accepted in conferences.

In the *tiny paper track*, we aim to encourage under-represented or under-resourced researchers to submit papers there. Our main goal would be to pair these researchers with experienced mentors before the workshop in order to receive feedback for their submission. The experienced mentors should have demonstrated track record of accepted papers on ML conferences.

We will present a *Best Paper Award* to the most outstanding paper, following a discussion among the PC members and core organizers.

The timeline below follows the official instructions in the ICLR site. However, the timeline can be amended as required if new conditions emerge:

Table 1: Timeline for contributed work submissions

Friday, December 6th, 2024	Portal opens for submissions
Monday, February 3rd, 2025	Submission deadline
Friday, February 14th, 2025	Reviewing deadline
Friday, February 28th, 2025	Notification date
Friday, April 18th, 2025	Camera-ready copy deadline
Sunday, April 27th, 2025	Workshop date

Plan to get an audience for the workshop. Given the prior organizing experience of the organizers, we expect that we will be able to reach a significant part of the ML community interested in the specific theme of uncertainty estimation in the era of foundation models. Specifically, we plan to promote the workshop as follows:

- We will start advertizing well in advance of the workshop date. Ideally, begin 2-3 months ahead to build anticipation.
- Our dedicated website for the workshop will be populated with details like the workshop theme, schedule, speakers, and important deadlines. This will be the first thing potential attendees see, so it should pique their interest.
- We will create social media profiles for the workshop on platforms like Twitter and regularly share updates, important announcements, and engaging content related to the workshop
- We will leverage our personal and professional networks to spread the word and encourage colleagues and contacts to share information about the workshop. We plan to build an email list of potential participants, send out announcements, and then highlight the workshop's unique selling points and how attendees will benefit.

Review process: The workshop will implement a rigorous and transparent peer review process for all submitted papers, utilizing the OpenReview platform. This workshop will implement a *double blind format of reviewing*. To ensure fairness and objectivity, we will require reviewers to declare any potential conflicts of interest before the review process begins. This information will be used to guide the assignment of papers to reviewers, preventing any compromises to the integrity of the evaluation process.

Each submitted paper will be thoroughly evaluated by a minimum of two reviews. The organizing team comprises individuals with diverse areas of expertise in the field. This breadth of knowledge allows us to step in as emergency reviewers if necessary, ensuring that all papers receive at least three comprehensive reviews by the established deadline. This contingency plan helps maintain the quality and timeliness of the review process. In the role of program chairs, the organizing committee will be responsible for making the final acceptance decisions. These determinations will be based on careful consideration of the evaluations provided by the reviewers. To uphold the highest standards of ethical conduct, any committee member with a conflict of interest related to a specific submission - such as a collaborative relationship or shared institutional affiliation - will be excluded from the decision-making process for that particular paper. This structured approach aims to provide a fair, thorough, and transparent evaluation process for all submitted works, contributing to the overall quality and credibility of the workshop.

As workshop organizers, we will not give any talk in the workshop, but will instead facilitate the invited speakers and the contributed talks.

We are introducing a *peer review mentorship* initiative aimed at nurturing early-career reviewers. This program will match novice reviewers with experienced peers in the field. These mentoring partnerships are designed to provide emerging reviewers with immediate feedback, direction, and support as they learn to navigate the intricacies of producing perceptive and valuable critiques for workshop submissions.

8 Organizers

Below, we provide further details on the core team and the program committee of the workshop.

8.1 Core team

The organizers have made substantial contributions to various facets of out-of-distribution prediction research. They have also shown extensive experience by **organizing over a dozen workshops and tutorials over the previous decade**, effectively engaging the diverse sub-communities. Members of the team have also been Program Chairs of ICLR or serve in major roles, such as the board of ACL.

Grigorios Chrysos (UW-Madison) is an Assistant Professor at the University of Wisconsin-Madison. Before that, Grigorios was a postdoctoral fellow at EPFL following the completion of his PhD at Imperial College London. Previously, he graduated from the National Technical University of Athens with a Diploma/MEng in Electrical and Computer Engineering. Grigorios has co-organized workshops in top conferences (ICCV'15, CVPR'17, ICCV'17, NeurIPS'24, AAAI'25). The most recent workshops on NeurIPS'24 ('*Fine-Tuning in Modern Machine Learning: Principles and Scalability*') and AAAI'25 ('*CoLoRAI - Connecting Low-Rank Representations in AI*') are complimentary to this workshop and focus on different perspectives of foundation models. Grigorios also organized tutorials on polynomial nets, tensors and architecture design (CVPR'22, AAAI'23, DSAA'24) and deep learning theory (CVPR'23, ISIT'24). His research interests lie in generative models, and designing models robust to noise and out-of-distribution samples. Grigorios has published several papers on the topic of both deep learning theory, and extrapolation in top-tier conferences (CVPR, NeurIPS, ICLR, ICML). Grigorios serves as an Associate Editor for TMLR and an Area Chair for top-tier ML conferences (NeurIPS'24, ICLR'25). [Google Scholar][Email]

Sharon Li (UW-Madison) is an Assistant Professor in the Department of Computer Sciences at the University of Wisconsin-Madison. She received a Ph.D. from Cornell University in 2017, advised by John E. Hopcroft. Subsequently, she was a postdoctoral fellow in the Computer Science department at Stanford University. Her research focuses on the algorithmic and theoretical foundations of reliable machine learning in the open world, as well as developing responsible foundation models including large language models and vision-language models. Sharon has served as the founding organizer and Program Chair for the *ICML Workshop on Uncertainty and Robustness in Deep Learning* (2019 and 2020), co-organized multiple other workshops including the *ICML Workshop on Distribution-free Uncertainty Quantification* in 2021 and 2022, *NeurIPS'22 Workshop on Robustness in Sequence Modeling*, and *ICCV'23 Tutorial on Reliability of Deep Learning for Real-World Deployment*. She has served as Area Chair and Senior Program Committee for top-tier ML conferences including ICLR, ICML and NeurIPS between 2020 and 2024. [Google Scholar][Email]

Anastasios Angelopoulos (UC-Berkeley) is a sixth-year Ph.D. student at UC Berkeley advised by Michael I. Jordan and Jitendra Malik. He works on theoretical machine learning with applications in vision and healthcare. Anastasios' goal is to apply modern statistical ideas to endow black-box models like deep neural networks with rigorous statistical uncertainty guarantees. Anastasios co-organized the ICML workshops on distribution-free uncertainty quantification in 2021 and 2022 and the NeurIPS workshop on statistical frontiers in foundation models in 2024. [Google Scholar][Email]

Stephen Bates **(MIT)** is an Assistant Professor of AI and Decision-making in the Massachusetts Institute of Technology Electrical Engineering and Computer Science Department. He received his Ph.D. in statistics from Stanford University in 2020, advised by Emmanuel Candès, and subsequently was a postdoctoral researcher at UC Berkeley, hosted by Michael I. Jordan. He works on uncertainty quantification with machine learning models, focusing on statistical inference, calibration, and rigorous control of error rates such as the false discovery rate, as well as applications in life science and earth science. He co-organized the *ICML Workshop on Distribution-free Uncertainty Quantification* in 2021 and 2022. Tutorial materials from that workshop were subsequently developed into a widely-read survey and YouTube video series on conformal prediction Angelopoulos & Bates (2023), highlighting how workshops he has co-organized have had a lasting impact on the research community. [Google Scholar][Email]

Barbara Plank **(LMU Munich)** is a Full Professor and Chair of AI & Computational Linguistics, LMU Munich, Germany. She completed her PhD in University of Groningen. She works on NLP, including language models and uncertainty, featuring many keynote talks on the topic. She has been a workshop organizer since over a decade (including *ACL conferences and ICLR, with the most relevant *First Workshop on Uncertainty-Aware NLP* at EACL 2024), and a keynote speaker for many conferences and workshops on Natural Language Processing (NLP), including (selected): ACL 2024, GCPR 2023, CLEF 2023, a Workshop on NLP for Human Resources in EACL 2024 and many summer schools. She currently serves as Vice President of Association for Computational Linguistics (ACL), elected in late 2023. [Google Scholar][Email]

Emtiyaz Khan **(RIKEN)** is a team leader (tenured) at the RIKEN center for Advanced Intelligence Project (AIP) in Tokyo leading the Approximate Bayesian Inference Team. Emtiyaz is an Action Editor for the Journal of Machine Learning (JMLR), and has served in organization and reviewing of most major Machine Learning conferences. From 2014 to 2016, he was a scientist at EPFL in Matthias Grossglausser's lab. During his time at EPFL, he taught two large machine learning courses for which he received a teaching award. Emtiyaz finished his PhD at UBC in 2012 under the supervision of Kevin Murphy. Emtiyaz serves as Program Chair in Artificial Intelligence and Statistics (AISTATS) 2025 and served as a Program Chair of ICLR 2024.

Volunteers: Four volunteers (2 undergrad and 2 PhD students) will help the organizers hosting the workshop. The students are James Oldfield (Queen Mary University London), Andrea Tseng (University of Wisconsin-Madison), Wuzhen Li (University of Wisconsin-Madison), and Sean Xuefeng Du (University of Wisconsin-Madison).

8.2 Program Committee

The following people have agreed to serve in the program committee:

Elias Rocamora (EPFL), Blerina Gkotse (University of Wisconsin-Madison), Justin Deschenaux (EPFL), Andrea Tseng (University of Wisconsin-Madison), James Oldfield (Queen Mary University London), Thomas Pethick (EPFL), Stratis Skoulakis (Aarhus university), Wuzhen Li (University of Wisconsin-Madison), Kimon Antonakopoulos (EPFL), Dimitris Halatsis (Imperial College London), Muhammad Ashiq (University of Wisconsin-Madison), Zhiyuan Wu (University of Oslo), Vasilis Papageorgiou (University of Wisconsin-Madison), Aggelina Chatziagapi (Stonybrook University), Jiankang Deng (Imperial College London).

Based on the volume of submissions, it may be necessary to expand the program committee. At this stage, the core team's extensive experience in organizing scientific events will be crucial for leveraging our broad network.

References

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. ISSN 1935-8237. doi: 10.1561/2200000101. URL http://dx.doi.org/10.1561/2200000101.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in neural information processing systems*, pp. 1877–1901, 2020.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.