I CAN'T BELIEVE IT'S NOT BETTER: CHALLENGES IN APPLIED DEEP LEARNING

Advertising Tagline. *Why do deep learning approaches often fail to deliver as expected in the real world? Dive deep into the pitfalls and challenges of applied deep learning.*

Workshop Summary. In recent years, we have witnessed a remarkable rise of deep learning (DL), whose impressive performance on benchmark tasks has led to increasing ambitions to deploy DL in real-world applications across all fields and disciplines (Hu et al., 2023; Wang et al., 2023b; Jumper et al., 2021; Brooks et al., 2024; Hong et al., 2024). However, despite its potential, DL still faces many challenges during deployment in dynamic, real-world conditions, exposing practical limitations that are often overlooked in controlled benchmarks. For instance, in robotics, world models or simulators (Brooks et al., 2024; Yang et al., 2024) may produce hallucinations due to a lack of grounded understanding of physics and causality; in healthcare, popular DL models for computer vision can struggle with tumor segmentation tasks for certain diseases (Yao et al., 2023); in commercial applications, the LLM-based chat agents can fail to comprehend ethics and social norms, resulting in societal, ethical, and fairness issues, or even harmful content (Anwar et al., 2024; Wang et al., 2023a). However, current publication mechanisms tend to prioritize solutions that work on standard benchmarks, lacking a platform to systematically collect real-world failure cases. Moreover, discussions about these failures are usually confined within specific domains, with limited cross-domain interaction, even though these failures may have similar underlying causes. Establishing a platform for collecting and sharing real-world challenges and failures of DL can address fundamental issues to facilitate more successful deployment of DL across domains, and enhance understanding of theoretical and empirical weaknesses in machine learning (ML) research. Building such a platform and fostering this community has been the continuous goal of our *I Can't Believe It's Not Better* (ICBINB) initiative. As DL systems have become increasingly present in everyday life also for non-scientific people, we want to put a special focus on real-world applications now. Therefore, in this proposed ICBINB workshop, we aim to explore the challenges, unexpected outcomes, and common principles underlying similar issues or failure modes encountered across various fields and disciplines when deploying DL models in real-world scenarios.

Call for Contributions. We will focus the discussion on: (i) **Challenges & failure modes**: We will invite papers from diverse fields including but not limited to healthcare, scientific discovery, robotics, education, equality & fairness, and social sciences to discuss the challenges and failure modes when deploying DL models for domain-specific applications as well as the underlying reasons. The failure modes may include *suboptimal performance*, concerns with the *safety and reliability* of applying DL models in unpredictable real-world applications, as well as *ethical and societal challenges*. (ii) **Common challenges across domains & underlying reasons**: We aim to discuss common reasons or patterns in challenges and failure modes across disciplines, which may include, but are not limited to, *data-related issues* (e.g., distribution shift, bias, label quality), *model limitations* (e.g., ethics, fairness, interpretability, scalability, domain alignment), and *deployment challenges* (e.g., computational demands, hardware constraints). Identifying these problems will create opportunities for researchers from different domains to interact and share insights, accelerating research by translating findings from one field to another. It will also deepen DL researchers' understanding of the universal fundamental issues that should be addressed within the current theoretical and empirical research paradigms. Embracing negative results as valuable learning opportunities will help the community learn from past failures, and drive the development of more robust, reliable, and applicable AI models. Our call will be open to novel, ongoing, and unpublished research. Our established reviewer guidelines and network of reviewers will enable us to follow the suggested ICLR timeline, releasing a *Call for Papers* by December 6th, 2024, and accepting submissions until February 3rd, 2025. The reviewing period will run from February 4th until February 28th, after which final decisions, reviews, and meta-reviews will be released on March 5th, 2025. The camera ready and poster submission will be due on March 19th, 2025. The specified due dates are set for 23:59 (11:59 pm) AoE time. A tentative call for papers can be found in Appendix A.

Workshop Program. The full-day *in-person* workshop will take place on April 27th or 28th, 2025 (TBD). We will host five invited talks with moderated Q&A alongside six spotlight talks,

highlighting particularly noteworthy submissions nominated by the program committee. All talks will be live-streamed for online participants to view with the ability to have questions asked to the speaker via an online chat. There will be a 90-minute poster session leading into the lunch break where all accepted submissions will be displayed. Finally, we will host a moderated panel discussion on the topic of *"Recognizing and Mitigating Challenges of When Applying DL in Real World"* for roughly one hour. The table below lays out a tentative schedule based on five invited talks of 30 minutes each and six spotlight talks of 10 minutes each. All oral presentations will include a Q&A segment for members of the audience. We believe the diverse mix of speakers as well as contributions will lead to thought-provoking and broad-ranging discussions, as it has in past editions of ICBINB workshops.

Tentative schedule

PST	Morning	PST	Afternoon
08:15	Opening Remarks	13:30	Invited Talk 3 (incl. Q&A)
08:30	Invited Talk 1 (incl. Q&A)	14:00	Invited Talk 4 (incl. Q&A)
09:00	Invited Talk 2 (incl. Q&A)	14:30	Spotlight Talks 4, 5 & 6
09:30	Coffee Break	15:00	Coffee Break
10:00	Spotlight Talks 1, 2 & 3	15:30	Invited Talk 5 (incl. Q&A)
10:30	Poster Session	16:00	Panel Discussion
12:00	Lunch Break	16:55	Closing Remarks

Speakers and Panelists. This year's invited speakers and panelists will promote critical discussions to dig into the challenges and failures of deploying DL in real-world applications, moving beyond benchmarks and the pursuit of state-of-the-art performance. We currently have confirmations from five speakers and four panelists who represent a diverse range of research directions and domains (for detailed biographies see Appendix B):

Confirmed Speakers

Name (alphabetically)	Institution	Country	Position	Area
Nick Haber	Stanford	USA	Assistant Professor	Psychology: Cognition, and Education
John Kalantari	University of Minnesota	USA	Assistant Professor	AI for Healthcare
Roberta Raileanu	Meta	UK	Research Scientist	Robotics
Sunayana Sitaram	Microsoft	India	Principal Researcher	Natural Language Processing
Otilia Stretcu	Google	USA	Senior Research Scientist	AI for Science

Confirmed Panelists

Name (alphabetically)	Institution	Country	Position	Area
Nick Haber	Stanford	USA	Assistant Professor	Psychology: Cognition, and Education
Zachary Lipton	CMU/Abridge	USA	Associate Professor	AI for Healthcare and Natural Language Processing
Liyue Shen	UMich	USA	Assistant Professor	AI for Biomedical Imaging
Sunayana Sitaram	Microsoft	India	Principal Researcher	Natural Language Processing

As the conference is still more than half a year in the future, some confirmations are tentative, since funding for some invited speakers is not 100% secured yet. We require all panelists to participate in person and will allow no more than one speaker remotely (ideally zero). Speakers have assured us to reconfirm their commitment over the course of the next few months.

Outreach and Access. To solicit an audience and submissions for the "I Can't Believe It's Not Better: Challenges in Applied Deep Learning" workshop, we will use the following channels:

- We will ask the invited speakers/panelists if they would additionally advertise the workshop to their students and/or collaborators and provide them with materials.
- Posting to social media sites like X, Discord and Mastodon.
- Posting to Google groups affiliated with DL applications, should their posting rules allow it.
- The ICBINB homepage will feature and link to a website dedicated to the workshop.
- Promote the workshop locally at conferences in different disciplines. Specifically, we plan to promote it at conferences this winter provided some organizers or our colleagues will be attending, including those on general ML/AI (NeurIPS, AAAI), computer vision (BMVC, ACCV, WACV), natural language processing (EMNLP), healthcare (ML4H), robotics (CoRL).

- Slack spaces across several labs at CMU, Cornell, Stanford, UMich, UPenn, University of Amsterdam, Google, Meta, Microsoft, Apple, and Snap, as well as Slack spaces of international research initiatives such as channels on Embodied AI, Therapeutics Data Commons, Computational Behavior, and AI for Science.
- Our ICBINB initiative and advisor team also includes great researchers at universities and in the industry who will help spread the word about the workshop to their institutions and collaborators, including domain-specific experts in robotics, healthcare, psychology, fairness, and social aspects, AI for science.
- Mailing lists across multiple departments at the same institutions or companies.

In addition to the call for papers and general workshop information, all workshop-related content, such as talk titles, recorded talks, published papers, and posters as well as links to follow-up publications in a journal special issue (such as PMLR), will be made available on the workshop-specific webpage to allow people who cannot attend physically to fully catch up with the ICBINB activities. To enable this, we plan to use SlidesLive and RocketChat technologies, as we are experienced with those technologies.

We are confident that a thought-provoking atmosphere at ICLR 2025 will be achieved through the combination of distinguished invited speakers, spotlight oral talks, an extended poster session, the panel discussion, online inclusion of questions and finally the curious, inquisitive and reflective nature of everyone involved in the workshop, as we have seen from previous iterations.

Relation to other Workshops. There have been several past workshops and tutorials focusing on the real-world and domain applications of DL models Workshop (2023b;a; 2021; 2023c; 2024d;c). While these workshops have explored the integration of AI with scientific domains, our workshop offers a complementary perspective by emphasizing the often overlooked yet critical aspect of negative results. We seek to understand where and why DL models fall short in applications in these domains and identify potential common reasons or patterns across them. Inspired by our ICBINB initiative and past NeurIPS workshops (see below), there are also ICBINB workshops on negative results *in specific domains* organized by other teams Workshop (2024b;a), yet our workshop aims to uncover the common failure modes *across domains* in real-world settings. Importantly, the general nature of ICLR and its larger audience can help us achieve this goal more effectively, as it serves as an influential platform for the broader ML community working on developing application-agnostic DL and foundation models. The goal of the ICBINB workshop series is to promote slow science and build a community to discuss surprising and negative results, encouraging a culture of transparency and shared learning. In 2023, ICBINB organized the in-person workshop "Failure Modes in the Age of Foundation Models" at NeurIPS. The year before, ICBINB held a hybrid workshop titled "Understanding Deep Learning Through Empirical Falsification". These past editions achieved $\sim$150 peak physical attendees, over $\sim$2.5k unique views virtually, and had around 40 submissions out of which roughly 35 were accepted. We similarly expect attendance numbers of 100–150 people for our workshop at ICLR 2025. This year's proposal emphasizing the challenges and practical limitations of deep models in real-world applications, has not been the focus of any previously held ICBINB workshops.

Internal Guidelines and Conflicts of Interest. Reviewers will be asked to check that papers follow the workshop themes, in particular highlighting a specific challenge, failure mode, or negative result of DL-based systems in real-life applications. Beyond purely showcasing these challenges and failure modes, papers are expected to provide insight into the reasons underlying the issue. Results that are particularly unexpected should be up-weighted. Additionally, we invite reflections on the challenges of popular models in real-life applications. Regarding paper assignments, we will use OpenReview's bidding system to make sure reviewers have the domain knowledge necessary for reviewing papers.

Submissions that provide particular insight into open challenges for DL applications should be highlighted as potential spotlight talks. Reviewers are additionally asked to nominate papers for the "Entropic Award" for the most surprising negative result, and the "Didactic Award" for the most pedagogical and well-explained paper. Papers with exemplary scientific rigor and insightful findings will be nominated for publication in PMLR. We will not accept works that have been previously published. To avoid direct conflicts of interest we will use submitted Advisors, Relations & Conflicts on author OpenReview profiles. Reviewers will not review submissions from their affiliated

institutions. Workshop organizers/advisors, in the process of writing meta-reviews and final decisions, will not be asked to assess a contribution from their institution. The organizers/advisors will not submit any contribution (talk or paper) to this workshop. Organizers reserve 2 slots for opening and closing remarks, as shown in the workshop schedule.

In keeping with previous workshop editions, we aim to assign 4 reviewers per paper and 2-3 papers per reviewer. Based on an initial outreach to reliable reviewers from previous workshop editions, we already have a confirmed program committee of 56 reviewers at this stage (names included at the end of this proposal). Our pool of reviewers ensures we can provide quality feedback on all submissions.

Diversity. Our workshop invited speakers and panelists from a diversity of research fields. In planning this workshop, we prioritized promoting diversity of demographics and academic backgrounds for both organizers and speakers. When selecting speakers, we aimed to invite people from a variety of seniority levels. We also aimed to invite a diversity of voices from industry and academia as well as voices across genders and ethnicities. Similarly, the workshop organizers themselves span a variety of international ethnicities, genders, and seniority levels from recent Master's graduates to research scientists in industry. We include first-time organizers alongside more experienced community members. Organizer Yubin Xie has taken formal training on diversity and inclusion for machine learning conference organizations.

Providing a venue to share negative results regardless of domain encourages diversity of thinking. We aim to promote greater openness and transparency in the field and reduce gaps in sharing and distributing new knowledge.

Our workshop aims to foster an inclusive environment and to provide a venue for the publication of diverse viewpoints that may be difficult to publish in traditional venues. To foster an inclusive environment we will provide and enforce a workshop code of conduct. To encourage the acceptance of diverse viewpoints, authors and reviewers will be provided with clear guidelines for the creation and evaluation of submissions. Particular efforts will be made to encourage submissions from members of ML affinity groups through their mailing lists and in the language of our call for papers. Importantly, we will feature a tiny paper track to make our workshop more accessible to under-represented and under-resourced researchers (see below).

Lastly, we will make use of available online tooling to encourage virtual participation from researchers who are unable to attend in person due to financial, health, and visa constraints.

Tiny Papers. Our workshop will feature a tiny paper track, in order to make our workshop more accessible to under-represented and under-resourced researchers. Tiny papers will be similar to normal papers, but of a 2-page length maximum and with lower requirements in terms of investigated aspects and depth of analysis. For example, they will not be required to investigate underlying reasons for failed applications of DL in the real world, but just need to describe the application that was tackled with deep learning, how it was tackled, and the (negative) outcome of the experiment.

ORGANIZING TEAM

This section includes the biographies of the organization team showing their organizational experience, skills, and background. Technical expertise spans core ML/DL, computer vision, robotics, NLP, AI for healthcare, AI for computational biology, computational social science, finance, and more. This diversity allows us to engage people from different domains to participate and submit papers, and gain valuable insights during the organization, review, and award selection phases. Regarding organizational experience, Arno, Fan, Andreas, Tobias, Yubin, and Rui all helped organize previous in-person iterations of ICBINB workshops. We also strive to allow new members the opportunity to participate in the organization of our workshops, this being Jennifer, Priya, and Zhaoying's first such opportunity. The organizers are geographically located in: North America (5), Europe (3), and Asia (1). Professional affiliations are: Academia (5) and Industry (4). Genders are: Male (5), Female (4). Seniority levels range from recent Master's graduates to full-time ML research scientists.

ORGANIZERS

Arno Blaas (*ablaas@apple.com*) is a Machine Learning Research Scientist at Apple (Barcelona, Spain). He obtained his PhD from the University of Oxford. His research interests include ML robustness and AI safety in general. He was one of the organizers of the 2022 NeurIPS workshop, "I Can't Believe It's Not Better: Understanding Deep Learning Through Empirical Falsification", and is one of the co-organizers of this year's NeurIPS workshop on "Foundation Model Interventions". Google Scholar

Priya D'Costa (*pdcosta@alumni.upenn.edu*) recently graduated with a Master's degree in Computer Science from the University of Pennsylvania. She spent 2 years at the Computational Social Science Lab at the University of Pennsylvania, working on developing an NLP-based conversation analysis framework for computationally measuring concepts in social science, such as conflict and negotiations. Before her transition to Computer Science, she spent 5 years working in Finance. Personal SiteGoogle Scholar

Fan Feng (*ffeng1017@gmail.com*) is an incoming postdoctoral researcher with the CMU-CLeaR group, also working jointly at MBZUAI and UCSD. He completed his PhD at the City University of Hong Kong and was a visiting PhD student at the University of Amsterdam. His research focuses on the intersection of causal discovery and decision-making, with the goal of improving the interpretability, efficiency, and robustness of ML/RL systems by uncovering their underlying causal world models. He has organized the NeurIPS ICBINB workshops in 2022 and 2023, as well as the Reinforcement Learning Beyond Rewards workshop at RLC 2024. Personal Site Google Scholar

Andreas Kriegler (*andreas.kriegler@tuwien.ac.at*) is a PhD student at the Technical University of Vienna supervised by Margrit Gelautz. He is further funded by the Austrian Institute of Technology (AIT). His research is centered around geometrical approaches to enable 3D perception for robotics systems. Personal interests include philosophy of science and working towards a more rigorous approach for computational sciences. He was one of the organizers of the 2022 NeurIPS workshop, "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models". Personal Site Google Scholar

Zhaoying Pan (*pan433@purdue.edu*) is a PhD student at Purdue University, advised by Dr. Joy Wang. Zhaoying received her master's degree from the University of Michigan and her bachelor's degree from the University of Chinese Academy of Sciences. Her current research focuses on trustworthy machine learning and its application in healthcare. Her past research experience spans a variety of applications of computer vision and machine learning. Personal Site Google Scholar

Tobias Uelwer (*tobias.uelwer@iais.fraunhofer.de*) is a Data Scientist at Fraunhofer IAIS and is also affiliated with the Lamarr Institute for Machine Learning and Artificial Intelligence. He obtained his PhD from the Technical University of Dortmund. His research interests are inverse problems, adversarial robustness of deep neural networks, and applications of machine learning. He was one of the organizers of the 2023 NeurIPS workshop, "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models". Personal Site Google Scholar

Jennifer Williams (*jlw1@alumni.cmu.edu*) is a Machine Learning Scientist at CVS Health. She obtained her PhD at Carnegie Mellon University. Her research focuses on machine learning for

healthcare, natural language processing, and causality. She previously co-founded CVS's ML Lunch and Learn Series, co-organized CMU's brAIn seminar series, and organized student events as President of her Graduate Student Association. Personal Site Google Scholar

Yubin Xie (*yx443@cornell.edu*) is a machine learning scientist from noetik.ai. He obtained his Ph.D. from Cornell University and Memorial Sloan Kettering Cancer Center in computational biology and medicine. His research focuses on machine learning and statistical methods on single-cell cancer tissue imaging and genomics data. He was one of the organizers of the 2023 NeurIPS workshop, "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models". He also organized ICML workshops on Computational Biology from 2020-2023. Personal Site Google Scholar

Rui Yang (*ruy4001@med.cornell.edu*) is a PhD candidate at Cornell University under the supervision of Dr. Christina Leslie. Her research focuses on developing neural network to study the chromatin 3D folding and regulatory genomics. She was one of the organizers of the 2023 NeurIPS workshop, "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models". Personal Site Google Scholar

ADVISORS

I Can't Believe It's Not Better (ICBINB) is an initiative whose members extend beyond the organizing committee. In particular, we note below some of the community members who will not be involved in the direct day-to-day organization of the workshop but have provided expertise and guidance and will continue to do so during the organization process.

Ian Mason (*imason@fujitsu.com*) is a senior researcher at Fujitsu Research of America. He previously organized 2022 and 2023 NeurIPS workshops with I Can't Believe It's Not Better. Previously he was a postdoctoral associate in the Brain & Cognitive Sciences Department at MIT and completed his PhD in the Institute of Perception, Action and Behavior at the University of Edinburgh. His research interests include domain adaptation, robustness, few-shot learning and continual learning. Currently he works on self-improving systems and foundation models. Personal Site Google Scholar

Melanie F. Pradier (*melanief@microsoft.com*) is a senior researcher at Microsoft Research Cambridge in the UK, working on probabilistic models and causal representation learning for healthcare applications. Previously, she was a postdoc at Harvard University where she developed human-centric ML models to personalize antidepressant prescriptions, and designed meaningful priors for deep Bayesian models. Melanie has also worked at Sony European Research Center, Sony Corporation R&D, and the Memorial Sloan-Kettering Cancer Center. She has been an organizer of several previous workshops, including: I Can't Believe It's Not Better at NeurIPS 2020-2022 (3 years), Machine Learning for Drug Discovery at ICLR 2022, "Bridging the Gap: from Machine Learning Research to Clinical Practice" at NeurIPS 2021, and "Deep Generative Models for Health (DGM4H) at NeurIPS 2023. Personal Site Google Scholar

Francisco J. R. Ruiz (*franrruiz@google.com*) is a Research Scientist at DeepMind (London). He has co-organized the Advances in Approximate Inference Symposium (AABI, formerly a NeurIPS workshop) for three consecutive years and was General Chair for AISTATS 2023. His research is focused on statistical machine learning; in particular, his interests include: approximate Bayesian inference, probabilistic modeling for discrete data, generative models, applications of Bayesian nonparametrics, and time series models. Personal Site Google Scholar

PROGRAM COMMITTEE LIST

The following 56 reviewers have already accepted our call:

Niv Sivakumar, Shriarulmozhivaram Gobichettipalayam-Chandrasekaran, Mozhgan Saeidi, Nicholas Apostoloff, Sahil Khose, Sangwon Ha, Tuhin Sahai, Zeren Shui, Engelbert Nguifo, Erin Grant, Fan Feng, Fernando Martínez-Plumed, Florian Maire, Francisco Rodriguez, Haytham Fayek, Jan Ramon, Mani A, Tammo Rukat, Yubin Xie, Hadar Shavit, Felix Michels, Skyler Wu, Andreas Kriegler, Tatiana Likhomanenko, Evan Ott, Effy Li, Zeyu Zhang, Canyu Chen, Zheng Shen, Lora Aroyo, Thomas Mensink, Lisa Alazraki, Mingwei Shen, Rui Yang, Louis Liu, Weijia Zhang, Priya DCosta, Zhaoying Pan, Sebastian Caldas, Caleb Ellington, Oliver De Candido, Qi Liu, Jie Liu, Caleb Chuck, Cong Liu, Xavier Suau, Miguel Sarabia del Castillo, Wilfried Wöber, Thomas Germer, Sarah Schneider, Pengyu Zhang, Adriano Hernandez, Junyi Chai, Miro Astor, Aviya Litman, Natalie Sauerwald.

REFERENCES

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VtmBAGCN7o.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *Advances in Neural Information Processing Systems*, 36, 2023a.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023b.

Workshop. Robust and reliable machine learning in the real world, 2021. URL https://iclr.cc/virtual/2021/workshop/2129.

Workshop. Modern experimental design and active learning in the real world, 2023a. URL https://realworldml.github.io/.

Workshop. Ai for science, 2023b. URL https://ai4sciencecommunity.github.io/.

Workshop. Machine learning and the physical sciences (mlps), 2023c. URL https://ml4physicalsciences.github.io/2023/.

Workshop. Icbinb - cosyne, 2024a. URL https://alexhwilliams.info/cosyne-icbinb/.

Workshop. Icbinb - failure modes of sequential decision-making in practice, 2024b. URL https://sites.google.com/view/rlc2024-icbinb.

Workshop. Workshop on computer vision in the wild, 2024c. URL https://computer-vision-in-the-wild.github.io/cvpr-2024/.

Workshop. Machine learning for healthcare, 2024d. URL https://www.mlforhc.org.

Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Pack Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=sFyTZEqmUY.

Wenjian Yao, Jiajun Bai, Wei Liao, Yuheng Chen, Mengjuan Liu, and Yao Xie. From cnn to transformer: A review of medical image segmentation models. (arXiv:2308.05305), August 2023. doi: 10.48550/arXiv.2308.05305. URL http://arxiv.org/abs/2308.05305. arXiv:2308.05305 [cs, eess].

APPENDIX

A CALL FOR PAPERS (TENTATIVE)

The **I Can't Believe It's Not Better (ICBINB)** initiative is excited to announce its upcoming workshop at **ICLR 2025** in Singapore, dedicated to discussing **negative results of deep learning in real-world applications**. We invite researchers and industry professionals to submit their papers on negative results, failed experiments, and unexpected challenges encountered in the application of deep learning to real-world problems across industry and science. The primary goal of this workshop is to create a platform for open and honest discussion about the hurdles and roadblocks in applying deep learning. We believe that sharing these experiences is crucial for the advancement of the field, providing valuable insights that can prevent others from repeating the same mistakes, and fostering a culture of transparency and learning. We invite submissions that apply deep learning to various domains including, but not limited to, social sciences, biology, physics, chemistry, engineering, robotics, psychology, healthcare, neuroscience, marketing, economics, or finance.

Submitted papers should contain the following four elements:

- A problem in an application that was tackled with deep learning.
- A solution for this type of problem was proposed in the deep learning literature.
- A description of the (negative) outcome.
- An investigation (and ideally an answer) to the question of why it did not work as promised by the deep learning literature.

The potential reasons for failure may include but are not limited to data-related issues (e.g., distribution shift, bias, label quality, noisy measurement, quality of simulated data), model limitations (e.g., robustness, interpretability, scalability, domain alignment), and deployment challenges (e.g., computational demands, hardware constraints). Besides these four points, papers will be assessed on:

- Clarity of writing.
- Rigor and transparency in the scientific methodologies employed.
- Novelty and significance of insights.
- Quality of discussion of limitations.
- Reproducibility of results.

Selected papers with exemplary scientific rigor, insightful findings, and excellent presentation will be nominated by reviewers for optional inclusion in a **special issue of PMLR**. Alternatively, some authors may prefer their paper to be in the **non-archival** track which is to share preliminary findings that will later go to full review at another venue. Furthermore, reviewers will nominate papers for the spotlight and contributed talks as well as two awards: the "Entropic Award" for the most surprising negative result, and the "Didactic Award" for the most well-explained and pedagogical papers.

Formatting Instructions & Guidelines:

- For the camera-ready submissions please use the ICLR 2025 conference LaTeX style files.
- Submissions should be no more than 4 pages long (excluding references), and authors should consider the following:
  - Authors may include unlimited appendices but reviewers will not be required to take them into account in their assessment of the submission.
  - If relevant, it is strongly encouraged to include the checklist from the LaTeX template and a broader impact statement, neither of which are included in the page limit.
  - We welcome first-time authors to submit to this workshop. The workshop will be run in person.

Additionally, we welcome contributions of **tiny papers** to our workshop. These are papers with the same structure and formatting instructions as seen in full workshop submissions, but with at most *2* pages of the main text. They are not required to contain all four elements mentioned above,

but should at least highlight a problem in an application that was tackled with deep learning and a description of the (negative) outcome.

Important Dates:

- Paper Submission Deadline - **February 4th, 2025**
- Notification of Acceptance/Rejection - **March 5th, 2025**
- Camera-ready & poster submission - **March 19th, 2025**
- In-person Workshop - **April 27/28th, 2025** (TBA)

B SPEAKERS AND PANELISTS

Nick Haber is an Assistant Professor at the Stanford Graduate School of Education, and by courtesy, Computer Science. After receiving his PhD in mathematics on Partial Differential Equation theory, he worked as a postdoctoral fellow at Stanford in both the Wall Lab (working chiefly on the Autism Glass Project) and the NeuroAI Lab (on building curiosity within artificial intelligence, as well as cognitive models). Personal Site Google Scholar

John Kalantari is the Chief Technology Officer of YRIKKA and an Assistant Professor at the University of Minnesota. He previously served as Director of AI at the Mayo Clinic, holding appointments in the Department of Surgery, the Department of Quantitative Health Sciences, and the Center for Individualized Medicine. He is also the founder of the Biomedical Artificial General Intelligence Lab (BAGIL) at Mayo Clinic, an interdisciplinary group focused on developing digital health tools and predictive models to improve patient care and expand healthcare access through causal machine learning and reinforcement learning. At YRIKKA, he leads pioneering advancements in multi-modal generative AI, emphasizing the quantification of model uncertainty and robustness in high-stakes applications such as national defense and healthcare. His work bridges the gap between AI research and critical real-world implementations, pushing the boundaries of generative models to handle diverse data modalities and enhance decision-making in complex environments. Personal Site Google Scholar

Zachary Lipton is an Associate Professor of Machine Learning at Carnegie Mellon University (CMU). He holds appointments in the Machine Learning Department in the School of Computer Science, the Heinz School of Public Policy (courtesy) and Societal Computing (courtesy). His research spans core ML methods and theory, their applications in healthcare and natural language processing, and critical concerns, both about the mode of inquiry itself, and the impact of the technology it produces on social systems. He is also the CTO and Chief Scientist of Abridge, a healthcare AI company defining the cutting edge of technology in the emerging ambient listening space. The industry-leading product turns raw audio of doctor-patient conversations into high-quality drafts of after-visit documentation, freeing up doctors to focus on the patient (during the visit) and to focus mostly on last-mile edits (after the visit). Personal Site Google Scholar

Roberta Raileanu is a Research Scientist at Meta and an Honorary Lecturer at UCL. She earned her PhD in Computer Science from NYU where she worked on generalization in deep reinforcement learning. Roberta also holds a degree in Astrophysics from Princeton University. Currently, she works on augmenting foundation models with planning, reasoning and decision making abilities by training them from feedback and interaction with external tools, environments, humans, and other AI agents. Personal Site Google Scholar

Liyue Shen is an assistant professor in the EECS department at the University of Michigan. Prior to that, she received her B.E. degree in Electronic Engineering from Tsinghua University in 2016, and obtained her PhD degree from the Department of Electrical Engineering, Stanford University in 2022. She also spent one year as a postdoctoral research fellow at the Department of Biomedical Informatics, Harvard Medical School. Her research interest is in Biomedical AI, which lies in the interdisciplinary areas of machine learning, computer vision, signal and image processing, biomedical imaging, medical image analysis, and data science. She is particularly interested in developing efficient and reliable computational methods for medical imaging and informatics to tackle real-world health problems. She recently focuses on the generative diffusion models, implicit neural representation learning and multimodal foundation models. She is the recipient of Stanford Bio-X Bowes Graduate

Student Fellowship (2019-2022), and was selected as the Rising Star in EECS by MIT and the Rising Star in Data Science by the University of Chicago in 2021. Personal Site Google Scholar

Sunayana Sitaram is a Principal Researcher at Microsoft Research India in Bangalore, where she has been working for the last 8 years since completing her PhD at Carnegie Mellon University in 2015. She is passionate about making AI inclusive to everyone and her current focus is on improving the evaluation and performance of Large Language Models on non-English languages. In addition to her research, Sunayana also served for the last two years as the director of the MSR India Research Fellow program, which hosts 65 young researchers to prepare them for careers in research, engineering, and entrepreneurship. Sunayana is an active contributor to the field and regularly publishes her findings, as well as serves on the organizing committee for NLP conferences such as ACL, EMNLP, and CoLM. Personal Site Google Scholar

Otilia Stretcu is a Senior Research Scientist at Google Research in Mountain View, California, working on methods and tools that enable non-AI practitioners to efficiently train and deploy AI models for specialized applications using only domain knowledge. This work spans across multiple areas including large language models, active learning, few-shot learning, and knowledge distillation. Previously, she was a PhD student at Carnegie Mellon University, co-advised by Prof. Tom Mitchell and Prof. Barnabàs Pòczos. Her PhD research focused on developing algorithms for curriculum learning, semi-supervised learning, and graph-based learning, and applying them on problems related to health and neuroscience. Prior to her PhD, Otilia received an MPhil from the University of Cambridge, UK, where she was a Gates Cambridge scholar, and a BEng from Politehnica University of Timisoara, Romania. Personal Site Google Scholar