I CAN'T BELIEVE IT'S NOT BETTER: WHERE LARGE LANGUAGE MODELS NEED TO IMPROVE

Workshop Summary. The success of Large Language Models (LLMs) has reshaped natural language processing and, increasingly, machine learning (ML) at large. Trained on vast amounts of data, they are able to achieve impressive performances on tasks such as translation [1], multimodal understanding [2–5], reasoning [6–8], and increasingly open-ended and self-evolving agentic behaviors [9–11] that empower applications across many domains, such as tool use and code agents [12–15] or scientific discovery [16, 17]. In some cases, these performances can even match or surpass those of humans [18, 19].

But at the same time, it's becoming increasingly clear that LLMs are not without flaws. For example, it is well-known that LLMs can hallucinate [20, 21], and the mechanisms behind these inaccuracies remain an active area of research [22]. Additionally, the alignment of LLMs remains brittle in the face of shifting goals and adversarial prompting [23–27], while new benchmarks question their capability to reason and show limitations thereof [28, 29]. It is thus not surprising that recent studies have shown limitations and even risks associated with their deployment in critical settings, such as clinical decision-making [30], biosecurity [31] and factual knowledge assessment [32].

Ideally, findings about such flaws or limitations can be used to immediately improve LLMs and their capabilities, but this might not always be possible, for example due to computational limitations of academic researchers or because the approaches taken might not have been fruitful. In that case, it can be hard to share the found insights about the limitations of LLMs (or failed attempts in resolving them) since the current publication mechanism tends to prioritize positive over negative results. However, sharing and discussing limitations of LLMs and failed attempts to resolve them can be valuable for the community to find a way to ultimately overcome these limitations.

We propose to organise this workshop as a platform to investigate important limitations of current LLMs both through works that explicitly showcase a limitation (as the ones cited above), as well as through works that aim to overcome such current limitations through a promising approach but struggled to do so (negative results).

Call for Contributions. We will invite contributions and focus the discussion on two types of works:

(i) Works that showcase and investigate important limitations of current LLMs (across topics such as reasoning, alignment, efficiency and scaling, or hallucination). Such works can include, for example, studies that rethink pitfalls in established alignment and reasoning methodologies [33–35, 28, 32, 36], analyses of risks in open-ended, self-evolving agentic systems [37, 38], or demonstrations of practical downsides in applications (especially in safety-critical domains) [39, 30, 31].

(ii) Works that attempted promising ideas to overcome identified challenges in LLMs but fell short of the expected improvements. Such works are especially valuable if they clarify and analyze the reasons for the failure of the ideas. For example, [40] learned that using multi-model synthetic preference data for alignment does not work as well as expected and increases reward-hacking and jailbreak attack success rates. They traced this down to high linear separability between chosen and rejected responses in the multi-model synthetic preference data. Similarly, [41] argued that a human-level AI scientist remains out of reach based on current LLM agent systems and identified the root cause by extensive evaluations and analysis. Specifically, the authors trace the fundamental bottleneck is the limited capability of requisite verification procedures and implementation gaps.

Embracing negative results as valuable learning opportunities will help the community learn from past failures, and ultimately drive the development of better LLMs. 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 2nd, 2025, and accepting submissions of long-paper track (5 page limit) and tiny-paper track (2 page limit) until January 31st, 2026. The reviewing period will then conclude on February 25th, after which final decisions, reviews, and meta-reviews will be released on March 1st, 2026. Submissions of camera-ready papers and posters will be due on March 8th, 2026, then the completed workshop program and accepted papers will be imported to iclr.cc on March 11th, 2026. The specified due dates are set for 23:59 (11:59 pm) AoE. A tentative call for papers can be found in Appendix A.

Workshop Program. The full-day *in-person* workshop will take place on April 26th or 27th, 2026 (TBD). We will host five invited talks with moderated Q&A alongside six spotlight talks, highlighting particularly noteworthy submissions nominated by the program committee. Online participants will be able to view each talk through a live stream and will have the opportunity to ask the speaker questions during the Q&A through an online chat. There will be a 60-minute poster session where all accepted submissions will be displayed, to give plenty of space for discussion among participants. Finally, we will host a moderated panel discussion on the topic of *"I Can't Believe It's Not Better: Where Large Language Models Need To Improve"* for roughly one hour. The table below lays out a tentative schedule based on five invited talks of 40 minutes each (incl. 10 minutes for Q&A to facilitate discussion) and six spotlight talks of 10 minutes each (incl. Q&A). 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

BRT	Morning	BRT	Afternoon
08:00	Opening Remarks	13:00	Invited Talk 3 (incl. Q&A)
08:10	Invited Talk 1 (incl. Q&A)	13:40	Invited Talk 4 (incl. Q&A)
08:50	Invited Talk 2 (incl. Q&A)	14:20	Spotlight Talks 4, 5 & 6
09:30	Coffee Break	14:50	Coffee Break
10:00	Spotlight Talks 1, 2 & 3	15:20	Invited Talk 5 (incl. Q&A)
10:30	Poster Session	16:00	Panel Discussion
11:30	Lunch Break	16:55	Closing Remarks

Speakers and Panelists. This year's invited speakers and panelists will lead critical discussions on where and how LLMs need to improve, moving beyond benchmarks and the pursuit of state-of-the-art performance, highlighting important challenges and building connections to real-world applications. We currently have confirmations from five speakers and four panelists who represent a diverse range of backgrounds, seniority, research areas, and domains (for detailed biographies see Appendix B):

Confirmed Speakers

Name (alphabetically)	Institution	Country	Position	Research Area
Surbhi Goel	University of Pennsylvania	USA	Assistant Professor	Theoretical Foundations for Modern ML
Sewon Min	University of California, Berkeley	USA	Assistant Professor	Understanding and Advancing LLMs
Preslav Nakov	MBZUAI	UAE	Full Professor	Natural Language Processing and Fact Checking
Aditi Raghunathan	Carnegie Mellon University	USA	Assistant Professor	Failures, Safety and Reliability of Frontier Models
Verena Rieser	Google DeepMind	UK	Senior Staff Research Scientist	Alignment of LLMs

Confirmed Panelists

Name (alphabetically)	Institution	Country	Position	Research Area
Samy Bengio	Apple	USA	Senior Director	Reasoning capabilities of LLMs
Sewon Min	University of California, Berkeley	USA	Assistant Professor	Understanding and Advancing LLMs
Preslav Nakov	MBZUAI	UAE	Full Professor	Natural Language Processing and Fact Checking
Aditi Raghunathan	Carnegie Mellon University	USA	Assistant Professor	Failures, Safety and Reliability of Frontier Models

We require all speakers and panelists to participate in person. In case of exceptional circumstances (such as visa issues), we will allow speakers to give their invited talk remotely. However, for panelists in such circumstances, we will organise a replacement in order to guarantee the flow of discussion.

Outreach and Anticipated Audience. To solicit an audience and submissions for the *"I Can't Believe It's Not Better: Where Large Language Models Need to Improve"* 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, Bluesky, LinkedIn, and Discord.
- The ICBINB homepage will feature and link to a website dedicated to the workshop.

- Slack spaces and mailing lists across several labs at CMU, Cornell, Stanford, UMich, UPenn, University of Amsterdam, University of Oxford, Google, Meta, Microsoft, Apple, and Snap, as well as Slack spaces and mailing lists 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 experts in LLM reasoning, alignment, efficiency and scaling, and hallucinations.

Based on experience from previous editions of the ICBINB workshop series (see Section **Relation to other Workshops**) and the growth of the field, we expect to attract around 40–50 submissions and an attendance of 100–150 people with this intended outreach approach.

We are confident to attract a curious and inquisitive audience with this approach, which in combination with the distinguished invited speakers, the spotlight oral talks, the extended poster session, the panel discussion, and the online inclusion of questions will create vivid discussions and a thought-provoking atmosphere at ICLR 2026.

Relation to other Workshops. Past workshops and tutorials focusing on Large Language Models (LLMs) have explored their applications [42, 43], analyses [44, 45], or ethical considerations [46]. In contrast, our workshop offers a complementary perspective by emphasizing the often overlooked yet critical aspect of negative results. We seek to understand where and why LLMs face limitations in multiple areas, including reasoning, alignment, hallucinations, and efficiency and scaling. Our aim is to foster an environment that inspires researchers to share and learn from negative results in order to better understand and improve LLMs.

Unlike other workshops that prioritize positive findings, ours highlights work that may otherwise be overlooked due to different reviewing priorities. Continuing the ICBINB initiative and inspired by our previous ICLR and NeurIPS workshops (detailed below), we continue to champion the importance of slow science and transparent discussions of unexpected or negative findings.

Although other teams, inspired by our ICBINB initiative and previous ICLR and NeurIPS workshops, have organized ICBINB workshops focused on negative results in *specific domains* [47, 48], our workshop is domain agnostic, instead aiming to uncover the current broader limitations of LLMs. This broader scope aligns well with the general nature of ICLR, thus ICLR offers an ideal platform to engage the wider ML community in an inspiring discussion about the current limitations of LLMs.

The ICBINB workshop series has consistently sought to build a community to discuss surprising and negative results, encouraging a culture of transparency and shared learning. Previous workshops in the series include the ICLR 2025 "Challenges in Applied Deep Learning" workshop (in-person), the NeurIPS 2023 "Failure Modes in the Age of Foundation Models" workshop (in-person), and the 2022 NeurIPS "Understanding Deep Learning Through Empirical Falsification" workshop (hybrid). These past editions generally 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.

This year's proposal is distinct from previous ICBINB workshops, as it focuses specifically on negative results related to LLMs around recent topics such as reasoning, alignment, hallucinations, or efficiency and scaling. While the 2023 ICBINB workshop focused on foundation models, its focus was on understanding what classic ML problems were not yet solved by foundation models. Since then, the field's understanding of LLMs has increased drastically, and new research areas have developed. Our workshop focuses on the current state, by highlighting negative results in these newer areas of research.

Reviewer Guidelines and Conflicts of Interest. Reviewers will be asked to check that papers follow the workshop themes, in particular that they focus on challenges in LLM research in the form outlined above (Section **Call for Contributions**). Beyond purely showcasing these challenges and limitations, papers are expected to provide insight into the reasons underlying the issue. Results that are particularly unexpected should be up-weighted. Submissions that provide particular insight into open challenges 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. 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 two 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 a maximum of 3 papers per reviewer. Based on an expected number of 40–50 submissions, this will require around 55–65 reviewers. Based on an initial outreach to reliable reviewers from previous workshop editions, we already have a confirmed program committee of 50 reviewers at this stage (names included at the end of this proposal), and are thus confident to easily reach the required number of reviewers, ensuring that we can provide quality feedback on all submissions.

Diversity Commitment. Our workshop invited speakers and panelists from a diverse range of research areas. In planning this workshop, we prioritized promoting diversity of demographics, seniority, and backgrounds for both organizers, speakers and panelists. When selecting speakers and panelists, we invited researchers from a variety of levels of seniority, from newly appointed Assistant Professor to Senior Director of AI. We also aimed to invite diverse voices from both industry and academia, as well as voices across genders and ethnicities. Lastly, we made sure to not only invite North America based researchers, but also researchers based in Europe and Asia.

Similarly, the workshop organizers themselves span a variety of backgrounds, ethnicities, genders, and seniority levels, from graduate students to research scientists in industry. As part of a workshop series, we encourage diversity of organizing experiences, from first-time organizers to more experienced community members. More demographic information is provided further below in the organizer’s introduction. Organizer Yubin Xie has taken formal training on diversity and inclusion for machine learning conference organizations.

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 diverse viewpoints from authors with diverse levels of experiences, authors and reviewers will be provided with clear guidelines for the creation and evaluation of submissions. Importantly, we will feature a tiny paper track to make our workshop more accessible to underrepresented and under-resourced researchers (see below) and provide means for virtual participation to the workshop to accommodate researchers who are unable to attend in person due to financial, health, and visa constraints.

Virtual Access to Workshop Materials and Outcome. To ensure our workshop is accessible to a broad and diverse audience, we will use SlidesLive or other available tools to provide a live stream for virtual participation of ICLR-registered researchers. The real-time streaming and Q&A interactions will be available to facilitate the engagement of researchers who cannot attend in person. Our workshop sessions, including talks, panels, and paper presentations, will be recorded and made available on the ICLR site after the event. In addition, accepted papers at our workshop will be publicly accessible on OpenReview. Furthermore, the paper information, along with the posters collected from accepted authors, will be hosted on the workshop website. Lastly, similar to previous editions of the ICBINB workshop series, we plan to feature selected papers with exemplary scientific rigor, insightful findings, and excellent presentation in a special issue of PMLR [49–51].

LLM Usage Policy. Regarding the LLM usage, we primarily follow the Policies on Large Language Model Usage at ICLR 2026 and ICLR’s Code of Ethics. The use of publicly available LLMs is allowed as a general-purpose assist tool; however, the major contribution of the submitted work should be an original intellectual product of humans. We ask authors to disclose any use of LLMs in a mandatory section that does not count towards page limits. Papers that fail to disclose significant LLM usage can lead to desk rejection. The AI usage and disclosure will be evaluated by both reviewers and meta-reviewers. In addition, the authors and reviewers are ultimately responsible for their contributions. Authors and reviewers should understand that they take full responsibility for the content written under their name, including content generated by LLMs that could be construed as plagiarism or scientific misconduct (e.g., fabrication of facts). LLMs are not eligible for authorship. Following the ICLR workshop guidelines, we note that the tiny-paper track is a participatory initiative where AI assistance is permitted but submissions must be primarily human-authored with original thought and analysis.

Tiny Papers. Our workshop will feature a tiny paper track to make our workshop more accessible to underrepresented and under-resourced researchers. Tiny papers will be similar in scope to regular submissions, but only two pages in length and with lower requirements in terms of investigated aspects and depth of analysis. This track is designed to encourage the sharing of valuable but often unpublishable insights. For example, authors do not need to provide a full theoretical or empirical analysis of why a certain limitation occurs; instead, they should describe the problem that was investigated, the approach that was taken, and the negative or unexpected outcome. By lowering the barrier to contribution, the Tiny Paper Track aims to broaden participation and foster open discussion of challenges, limitations, and lessons learned in the study and improvement of LLMs.

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, Priya, Fan, Zhaoying, Jennifer, Yubin, and Rui 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 is Nikolai's first opportunity. The organizers are geographically located in: North America (6) and Europe (2). Professional affiliations are: Academia (4) and Industry (4). Genders are: Male (4), Female (4). Seniority levels range from graduate students 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", the 2024 NeurIPS workshop on "Foundation Model Interventions", and the 2025 ICLR workshop "I Can't Believe It's Not Better: Challenges in Applied Deep Learning". Google Scholar

Priya D'Costa (*pdcosta@alumni.upenn.edu*) is a Consultant at SAP America Inc., in the Intelligent HANA Team. She holds a Masters Degree in Computer Science, and spent 3 years at the Computational Social Science Lab at UPenn, as an early member of the Team Communication Toolkit team. Before her transition to Computer Science, she spent 5 years working in Finance. Additionally, she was one of the organizers of the ICBINB workshop "Challenges in Applied Deep Learning" at ICLR 2025. Google 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 learning the causal and strcutured representation for general-purpose 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, and the ICLR ICBINB workshop in 2025. Personal Site Google Scholar

Zhaoying Pan (*pan433@purdue.edu*) is a PhD student in electrical and computer engineering at Purdue University, advised by Dr. Joy Wang. Her current research focuses on trustworthy machine learning, especially distribution robustness. Before that, Zhaoying received her Master's degree from the University of Michigan, and her past research experience spans a variety of applications of computer vision and machine learning. Zhaoying was one of the organizers of the ICBINB workshop "Challenges in Applied Deep Learning" at ICLR 2025. Personal Site Google Scholar

Nikolai Rozanov (*nikolai.rozanov13@imperial.ac.uk*) is a computer science PhD student at Imperial College London and MBZUAI working with Prof. Iryna Gurevych and Dr. Marek Rei. Before that Nikolai was Co-Founder, CTO and Head of Research for over 7 years at a research based start-up in London, Wluper. During his industry experience Nikolai led a team of senior researchers, published in tier-1 venues including ACL, EMNLP, CoNLL, EACL; and led national Innovate UK Grants. More recently, Nikolai's research focus is on LLM Agents and Reasoning. 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, CVS's Generative AI Training workshops, co-organized CMU's brAIn seminar series, and organized student events as President of her Graduate Student Association. Additionally, she was one of the organizers of the ICBINB workshop "Challenges in Applied Deep Learning" at ICLR 2025. 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”, and the 2025 ICLR workshop “I Can’t Believe It’s Not Better: Challenges in Applied Deep Learning”. He also organized ICML workshops on Computational Biology from 2020-2023. Personal Site Google Scholar

Rui Yang (*ruiyang204@gmail.com*) is a postdoc researcher at the Broad Institute. Her research focuses on AI for Health, developing deep learning models to link epigenetic signatures and human diseases. Prior to that, Rui earned her Ph.D. in Computational Biology from Cornell University. 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”, and the 2025 ICLR workshop “I Can’t Believe It’s Not Better: Challenges in Applied Deep Learning”. Personal Site Google Scholar

Program Committee List

The following 50 reviewers have already accepted our call:

Adam Golinski , Adiba Proma , Andreas Kriegler , Aniq Ur Rahman , Ashutosh Kumar , Caleb Chuck , Dingling Yao , Dongsik Yoon , Dr A Mani , Drew McNutt , Fan Feng , Felix Michels , Fernando Martínez-Plumed , Francisco Ruiz , Han Wu , Hongye Cao , Jose Miguel Lara Rangel , Louis Jinrui Liu , Maike Behrendt , Manikandan Ravikiran , Mephu Nguifo Engelbert , Minghao Fu , Miro Astore , Mozhgan Saeidi , Natalie Sauerwald , Nicholas Apostoloff , Nivedha Sivakumar , Oleg Smirnov , Oliver De Candido , Pengyu Zhang , Pradeep Niroula , Qi Liu , Qianzi Li , Raviteja Sista , Rui Yang , Sarah Schneider , Selena Ge , Sevda OGUT , Shashank Yadav , Shen Zheng , Simon Damm , Tarun Raheja , Tobias Uelwer , Tuhin Sahai , Vibha Belavadi , Vikramjit Mitra , Xavi Suau , Xinran Gu , Yubin Xie , Zhaoying Pan

References

[1] SEAMLESS Communication Team. Joint speech and text machine translation for up to 100 languages. *Nature*, 637:587–593, 2025.

[2] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. URL https://arxiv.org/abs/2403.05530.

[3] Shuai Bai et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

[4] Bo Li et al. LLaVA-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. URL https://arxiv.org/abs/2408.03326.

[5] OpenAI. GPT-4v(ision) system card. https://openai.com/index/gpt-4v-system-card/, September 2023. System card describing the deployment and evaluation of GPT-4 with vision.

[6] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[7] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

[8] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[9] Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin godel machine: Open-ended evolution of self-improving agents. *arXiv preprint arXiv:2505.22954*, 2025.

[10] Jiahao Qiu, Xuan Qi, Tongcheng Zhang, Xinzhe Juan, Jiacheng Guo, Yifu Lu, Yimin Wang, Zixin Yao, Qihan Ren, Xun Jiang, et al. Alita: Generalist agent enabling scalable agentic reasoning with minimal predefinition and maximal self-evolution. *arXiv preprint arXiv:2505.20286*, 2025.

[11] Huan-ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, et al. A survey of self-evolving agents: On path to artificial super intelligence. *arXiv preprint arXiv:2507.21046*, 2025.

[12] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023.

[13] Niels Mündler, Mark Müller, Jingxuan He, and Martin Vechev. Swt-bench: Testing and validating real-world bug-fixes with code agents. *Advances in Neural Information Processing Systems*, 37:81857–81887, 2024.

[14] Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better llm agents. In *Forty-first International Conference on Machine Learning*, 2024.

[15] Hongru Wang, Cheng Qian, Manling Li, Jiahao Qiu, Boyang Xue, Mengdi Wang, Heng Ji, and Kam-Fai Wong. Toward a theory of agents as tool-use decision-makers. *arXiv preprint arXiv:2506.00886*, 2025.

[16] Ruofan Jin, Zaixi Zhang, Mengdi Wang, and Le Cong. Stella: Self-evolving llm agent for biomedical research. *arXiv preprint arXiv:2507.02004*, 2025.

[17] Pingchuan Ma, Tsun-Hsuan Wang, Minghao Guo, Zhiqing Sun, Joshua B Tenenbaum, Daniela Rus, Chuang Gan, and Wojciech Matusik. Llm and simulation as bilevel optimizers: A new paradigm to advance physical scientific discovery. *arXiv preprint arXiv:2405.09783*, 2024.

[18] Jiahao Qiu, Jingzhe Shi, Xinzhe Juan, Zelin Zhao, Jiayi Geng, Shilong Liu, Hongru Wang, Sanfeng Wu, and Mengdi Wang. Physics supernova: Ai agent matches elite gold medalists at ipho 2025. *arXiv preprint arXiv:2509.01659*, 2025.

[19] Xiaoliang Luo, Akilles Rechardt, Guangzhi Sun, Kevin K Nejad, Felipe Yáñez, Bati Yilmaz, Kangjoo Lee, Alexandra O Cohen, Valentina Borghesani, Anton Pashkov, et al. Large language models surpass human experts in predicting neuroscience results. *Nature human behaviour*, 9 (2):305–315, 2025.

[20] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46, 2025.

[21] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[22] Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

[23] Anthropic-Team. Agentic misalignment: How LLMs could be an insider threat. https://www.anthropic.com/research/agentic-misalignment, June 2025. Anthropic Research.

[24] Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *The Thirteenth International Conference on Learning Representations*, 2025.

[25] Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned llms with simple adaptive attacks. In *The Thirteenth International Conference on Learning Representations*, 2025.

[26] Yotam Wolf, Noam Wies, Oshri Avnery, Yoav Levine, and Amnon Shashua. Fundamental limitations of alignment in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 53079–53112, 2024.

[27] U Anwar, A Saparov, J Rando, D Paleka, M Turpin, P Hase, ES Lubana, E Jenner, S Casper, O Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. *Transactions on Machine Learning Research*, 2024.

[28] Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*, 2025.

[29] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

[30] Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.

[31] Zaixi Zhang, Zhenghong Zhou, Ruofan Jin, Le Cong, and Mengdi Wang. Genebreaker: Jailbreak attacks against dna language models with pathogenicity guidance. *arXiv preprint arXiv:2505.23839*, 2025.

[32] Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68, 2024.

[33] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

[34] Yiyou Sun, Georgia Zhou, Hao Wang, Dacheng Li, Nouha Dziri, and Dawn Song. Climbing the ladder of reasoning: What llms can-and still can't-solve after sft? *arXiv preprint arXiv:2504.11741*, 2025.

[35] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

[36] Abhay Sheshadri, John Hughes, Julian Michael, Alex Mallen, Arun Jose, Fabien Roger, et al. Why do some language models fake alignment while others don't? *arXiv preprint arXiv:2506.18032*, 2025.

[37] Siwei Han, Jiaqi Liu, Yaofeng Su, Wenbo Duan, Xinyuan Liu, Cihang Xie, Mohit Bansal, Mingyu Ding, Linjun Zhang, and Huaxiu Yao. Alignment tipping process: How self-evolution pushes llm agents off the rails. *arXiv preprint arXiv:2510.04860*, 2025.

[38] Shuai Shao, Qihan Ren, Chen Qian, Boyi Wei, Dadi Guo, Jingyi Yang, Xinhao Song, Linfeng Zhang, Weinan Zhang, Dongrui Liu, et al. Your agent may misevolve: Emergent risks in self-evolving llm agents. *arXiv preprint arXiv:2509.26354*, 2025.

[39] Xin Xu, Tong Xiao, Zitong Chao, Zhenya Huang, Can Yang, and Yang Wang. Can llms solve longer math word problems better? In *The Thirteenth International Conference on Learning Representations*, 2025.

[40] Yifan Wang, Runjin Chen, Bolian Li, David Cho, Yihe Deng, Ruqi Zhang, Tianlong Chen, Zhangyang Wang, Ananth Grama, and Junyuan Hong. More is less: The pitfalls of multi-model synthetic preference data in dpo safety alignment. *arXiv preprint arXiv:2504.02193*, 2025.

[41] Minjun Zhu, Qiujie Xie, Yixuan Weng, Jian Wu, Zhen Lin, Linyi Yang, and Yue Zhang. Ai scientists fail without strong implementation capability. *arXiv preprint arXiv:2506.01372*, 2025.

[42] *Building Trust in LLMs and LLM Applications: From Guardrails to Explainability to Regulation*, Singapore, 2025. ICLR. URL https://building-trust-in-llms.github.io/iclr-workshop/index.html.

[43] *Multi-modal Foundation Models and Large Language Models for Life Sciences*, Vancouver, Canada, 2025. ICML. URL https://fm4ls.github.io/.

[44] *Reasoning and Planning for Large Language Models*, Singapore, 2025. ICLR. URL https://workshop-llm-reasoning-planning.github.io/.

[45] *Evaluating the Evolving LLM Lifecycle*, San Diego, California, 2025. NeurIPS. URL https://sites.google.com/view/llm-eval-workshop.

[46] *Secure and Trustworthy Large Language Models*, Vienna, Austria, 2024. ICLR. URL https://set-llm.github.io/.

[47] *ICBINB - Failure Modes of Sequential Decision-Making in Practice*, New Orleans, Lousianna, 2024. NeurIPS. URL https://sites.google.com/view/rlc2024-icbinb.

[48] *ICBINB - COSYNE*, Cascais, Portugal, 2024. COSYNE. URL https://alexhwilliams.info/cosyne-icbinb/.

[49] Arno Blaas, Priya D'Costa, Fan Feng, Andreas Kriegler, Ian Mason, Zhaoying Pan, Tobias Uelwer, Jennifer Williams, Yubin Xie, and Rui Yang, editors. *Proceedings on "I Can't Believe It's Not Better: Challenges in Applied Deep Learning" at ICLR 2025 Workshops*, volume 296 of *Proceedings of Machine Learning Research*. PMLR.

[50] Javier Antorán, Arno Blaas, Kelly Buchanan, Fan Feng, Vincent Fortuin, Sahra Ghalebikesabi, Andreas Kriegler, Ian Mason, David Rohde, Francisco J. R. Ruiz, Tobias Uelwer, Yubin Xie, and Rui Yang, editors. *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, . PMLR.

[51] Javier Antorán, Arno Blaas, Fan Feng, Sahra Ghalebikesabi, Ian Mason, Melanie F. Pradier, David Rohde, Francisco J. R. Ruiz, and Aaron Schein, editors. *Proceedings on "I Can't Believe It's Not Better! - Understanding Deep Learning Through Empirical Falsification" at NeurIPS 2022 Workshops*, volume 187 of *Proceedings of Machine Learning Research*, . PMLR.

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 2026** in Rio de Janeiro (Brazil), dedicated to discussing **negative results and rigorous evidence of limitations in LLMs**. We invite researchers and industry practitioners to submit papers on negative results and unexpected challenges encountered in developing, aligning, scaling, evaluating, and deploying these systems. The primary goal is to create a platform for open, honest discussion about the hurdles and roadblocks in building reliable, efficient, and safe LLM systems. We believe that sharing these experiences is crucial for the field: it prevents teams from retracing unproductive paths, strengthens our understanding of failure modes and boundary conditions, and fosters a culture of transparency and learning.

To this aim, we invite submissions that

- *showcase and investigate important limitations of current LLMs*. This may include the evaluations on pitfalls in common approaches to alignment, reasoning, etc., and evaluations in real-world (especially safety-critical) applications. Example papers: [33–35, 28, 32, 37–39, 30, 31][1].
- *attempt promising ideas to overcome common challenges but fall short of the expected gains*, accompanied by analyses that clarify *failure modes* and *boundary conditions*. Example papers: [36, 40, 41].

Specifically, the submitted papers should contain the

- *Problem.* A problem in a clearly specified domain/setting (e.g., clinical decision-making tasks, long-horizon tool use, code agent, etc), with assumptions, target metrics, and desired improvements precisely stated.
- *Proposed solution.* A solution for this type of problem as proposed in prior literature, including its core mechanism, required preconditions, and the hypotheses under which it is expected to work.
- *Observed outcome.* A concise description of the negative or null outcome, including what failed to improve (and by how much), instability or regressions observed, and quantitative evidence (metrics, error bars, compute budget, data/seed details).
- *Reason for failure.* An investigation (and ideally an answer) to why it did not work as promised by the literature: e.g., dataset artifacts/leakage, mis-specified objectives (reward hacking), shortcut cues, distribution shift, optimization or scaling constraints (compute, memory, context length), fragile tool-use/memory orchestration in agents, or evaluation mismatches; supported by diagnostics/ablations and ending with boundary conditions and actionable takeaways.

Around these "negative results", a non-exhaustive, topic-wise list of LLM research problems includes (but is not limited to): (i). *Reasoning.* Works that reveal brittle logic, shallow or non-transferable chains of thought, limited systematic generalization, or domain-specific reasoning failures. (ii). *Alignment.* Misalignment between user intent and model behavior; failures in safety tuning, adversarial robustness, or goal preservation (including post-deployment drift). (iii). *Efficiency and Scaling.* Limitations in training, inference, and fine-tuning under realistic compute/latency/memory constraints, with particular emphasis on energy use and sustainability. (iv). *Agents.* Challenges in multi-step planning, tool use/selection, memory, self-monitoring/reflection, or stability of open-ended agentic systems. (v). *Hallucinations.* Studies of factual inaccuracies, fabricated/phantom citations, and calibration of model confidence/trust, alongside leak-resistant evaluation and mitigation. (vi). *Other.* Any well-supported finding that challenges prevailing assumptions, exposes boundary conditions, or provides constructive negative results.

Besides these points, papers will be assessed on:

[1] While these examples are full conference or journal papers, we are aiming for short submissions in a similar spirit (early results rather than well-established work).

- 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 2026 conference LaTeX style files.
- Submissions should be no more than 5 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 ... and a description of the (negative) outcome.

Important Dates:

- Paper Submission Deadline - **January 31st, 2026**
- Notification of Acceptance/Rejection - **March 1st, 2026**
- Camera-ready & poster submission - **March 8th, 2026**
- In-person Workshop - **April 26th or 27th, 2026** (TBA)

B Speakers and Panelists

Samy Bengio (PhD in computer science, University of Montreal, 1993) is a senior director of machine learning research at Apple since 2021. Before that, he was a distinguished scientist at Google Research since 2007 where he was heading part of the Google Brain team, and at IDIAP in the early 2000s where he co-wrote the well-known open-source Torch machine learning library. His research interests span many areas of machine learning such as deep architectures, representation learning, vision and language processing and more recently, reasoning. He is action editor of the Journal of Machine Learning Research and on the board of the NeurIPS foundation. He was on the editorial board of the Machine Learning Journal, has been program chair (2017) and general chair (2018) of NeurIPS, program chair of ICLR (2015, 2016), general chair of BayLearn (2012-2015), MLMI (2004-2006), as well as NNSP (2002), and on the program committee of several international conferences such as NeurIPS, ICML, ICLR, ECML and IJCAI. Personal Site Google Scholar

Surbhi Goel is the Magerman Term Assistant Professor of Computer and Information Science at University of Pennsylvania. Her research interests lie at the intersection of theoretical computer science and machine learning, with a focus on developing theoretical foundations for modern machine learning paradigms. Previously, she was a postdoctoral researcher at Microsoft Research NYC in the Machine Learning group. She received her Ph.D. in Computer Science from the University of Texas at Austin, where she was advised by Adam Klivans. Among her honors are the Bert Kay Dissertation award, a JP Morgan AI Fellowship, and a Simons-Berkeley Research Fellowship. She is also the co-founder of Learning Theory Alliance (LeT-All), a community building and mentorship initiative for the learning theory community. Personal Site Google Scholar

Sewon Min is an Assistant Professor in EECS at UC Berkeley, affiliated with Berkeley AI Research (BAIR), and a Research Scientist at the Allen Institute for AI. Her research lies at the intersection of natural language processing and machine learning, with a focus on large language models (LLMs). She studies the science of LLMs and develops new models and training methods for better performance, flexibility, and adaptability, such as retrieval-based LMs, mixture-of-experts, and modular systems. She also studies LLMs for information-seeking, factuality, privacy, and mathematical reasoning. She has organized tutorials and workshops at major conferences (ACL, EMNLP, NAACL, NeurIPS, ICLR), served as a Senior Area Chair, and received honors including best paper and dissertation awards (including ACM Dissertation Award Runner-up), a J.P. Morgan Fellowship, and EECS Rising Stars. She earned her Ph.D. from the University of Washington and has held research roles at Meta AI, Google, and Salesforce. Personal Site Google Scholar

Preslav Nakov is a Full Professor at MBZUAI, Abu Dhabi. Prior to joining MBZUAI, Professor Nakov worked at the Qatar Computing Research, HBKU where he was a principal scientist. Previously, he was a research fellow at the National University of Singapore (2008–2011) and a researcher at the Bulgarian Academy of Sciences (2008). He has been an honorary lecturer at Sofia University, Bulgaria since 2014. Professor Nakov authored a Morgan and Claypool book titled Semantic Relations Between Nominals (2nd edition in 2021) and two books on computer algorithms. He was also the first to receive the Bulgarian President's John Atanasoff award, named after the inventor of the first automatic electronic digital computer. Professor Nakov is one of the leading experts on "fake news", disinformation, fact checking, propaganda, and media bias detection and has published tens of research papers on solutions and stop-gaps for the ever-growing online social media infodemic. He's served on the program committees of the major conferences in computational linguistics and artificial intelligence. Most recently, he was a program committee chair of the annual conference of the Association for Computational Linguistics (ACL 2022). Personal Site Google Scholar

Aditi Raghunathan is an Assistant Professor at Carnegie Mellon University. She works broadly in machine learning, and her goal is to make machine learning more reliable and robust. Raghunathan's work spans both theory and practice, and leverages tools and concepts from statistics, convex optimization, and algorithms to improve the robustness of modern systems based on deep learning.Until recently, Raghunathan was a postdoc at Berkeley AI Research. She received her Ph.D. from Stanford University in 2021 where she was advised by Percy Liang. Her thesis won the Arthur Samuel Best Thesis Award at Stanford. Previously, she obtained her BTech in Computer Science from IIT Madras in 2016. Personal Site Google Scholar

Verena Rieser is a Senior Staff Research Scientist at Google DeepMind, where she works on Safer Conversational AI. She is also honorary professor at Heriot-Watt University in Edinburgh and a co-founder of the Conversational AI company ALANA AI. Verena holds a PhD from Saarland University in Germany and a MSc from the University of Edinburgh, where she also spent time as a postdoctoral researcher. She has 20 years of experience in developing and researching data-driven conversational systems. In the early 2000s she developed a series of breakthrough innovations that laid the groundwork for statistical dialogue control using Reinforcement Learning. More recently, Verena and her team pioneered work on identifying and addressing safety risks in neural conversational systems, which was awarded with a Leverhulme Senior Research Fellowship by the Royal Society. Personal Site Google Scholar