FOR IMMEDIATE RELEASE

13th Annual International Conference on Learning Representations (ICLR) 2025 Fact Sheet

Global Participation

- 11,039 participants spanning 85 countries
- 10,435 in-person
- 604 virtual
- List of countries with over 200 participants:
    1. United States: 2172
    2. China, Peoples Republic of China: 1946
    3. Singapore: 947
    4. United Kingdom: 583
    5. Korea, Republic of Korea: 528
    6. Germany: 398
    7. Canada: 308
    8. Switzerland: 227
    9. Japan: 213
    10. Hong Kong SAR, China: 207

Previous ICLR Locations and No. of Participants

- 2024: Vienna (Austria), 6,533 participants from 79 countries
- 2023: Kagali (Africa) 3,758 participants from 73 countries
- 2022: Virtual (Global) 5,200 participants from 81 countries
- 2021: Virtual (Global) 6,300 participants from 64 countries
- 2020: Virtual (Global) 5,600 participants from 76 countries
- 2019: New Orleans (USA) 2,600 participants from 50 countries
- 2018: Vancouver (Canada) 1,950 participants from 38 countries
- 2017: Toulon (France)
- 2016: San Juan (Puerto Rico)
- 2015: San Diego (USA)
- 2014: Banff (Canada)
- 2013: Scottsdale (USA)

Program Committee Statistics

- 11,603 submissions, 3704 accepted
    - 32% acceptance rate (last year: 31%)
    - 4,341 more submissions, 1,444 more accepted compared to last year
- 18,325 reviewers (9,375 more than 2024)
- 823 area chairs (199 more than 2024)
- 71 senior area chairs (11 more than 2024)
- 99.98% of papers received at least 3 reviews

- 2.6: avg. number of papers a reviewer reviewed
- 14.1: avg. number of papers in an AC's stack
- Average paper score before/after the discussion period: 4.78 -> 5.14

- 6 Invited Talks
- 213 Oral Presentations
- 85 Spotlight Posters
- 23 Socials
- 40 Workshops
- 6 Affinity events –
  - Muslims in ML Social
  - Women in Machine Learning Social
  - LatinX in AI Social
  - Queer in AI Social
  - Tiny Papers - integration into workshops
  - 49 Blogpost Track Posters out of 96 submissions - Announcement
- 6 Mentorship Chats with 19 senior researchers

Test of Time Winner

Adam: A Method for Stochastic Optimization
Diederik P. Kingma, Jimmy Ba
https://arxiv.org/abs/1412.6980

> As one of the most widely adopted optimization algorithms in deep learning, Adam revolutionized neural network training, enabling significantly faster convergence and more stable training across a wide variety of architectures and tasks. The algorithm automatically adjusts parameter-specific learning rates based on first and second moments of gradients, handling sparse gradients and non-stationary objectives. Adam's practical success has made it the default optimizer for countless state-of-the-art models, from computer vision and natural language processing to reinforcement learning, demonstrating remarkable versatility across problem domains and neural network architectures.

Test of Time Runner-Up

Neural Machine Translation by Jointly Learning to Align and Translate
Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio
https://arxiv.org/abs/1409.0473

> Introducing a form of attention, this paper fundamentally changed how sequence-to-sequence models process information. Before this work, encoder-decoder architectures usually compressed entire input sequences into fixed-length vectors, creating memory bottlenecks for longer sequences. The proposed approach enabled the model to "attend" to different parts of the source sentence dynamically during translation, allowing for processing of relevant contextual information. This attention mechanism has since become a cornerstone of modern deep learning, extending far beyond machine translation to form the foundation for transformers and large language models. The paper's practical impact has been immense, making it one of the most influential contributions to neural network architectures.

FOR IMMEDIATE RELEASE

Outstanding Paper Winners

Safety Alignment Should be Made More Than Just a Few Tokens Deep.
Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, Peter Henderson.

The safety alignment of current Large Language Models (LLMs) is vulnerable. Simple attacks, or even benign fine-tuning, can jailbreak aligned models. We note that many of these vulnerabilities are related to a shared underlying issue: safety alignment can take shortcuts, wherein the alignment adapts a model's generative distribution primarily over only its very first few output tokens. We unifiedly refer to this issue as shallow safety alignment. In this paper, we present case studies to explain why shallow safety alignment can exist and show how this issue universally contributes to multiple recently discovered vulnerabilities in LLMs, including the susceptibility to adversarial suffix attacks, prefilling attacks, decoding parameter attacks, and fine-tuning attacks. The key contribution of this work is that we demonstrate how this consolidated notion of shallow safety alignment sheds light on promising research directions for mitigating these vulnerabilities. We show that deepening the safety alignment beyond the first few tokens can meaningfully improve robustness against some common exploits. We also design a regularized fine-tuning objective that makes the safety alignment more persistent against fine-tuning attacks by constraining updates on initial tokens. Overall, we advocate that future safety alignment should be made more than just a few tokens deep.

Learning Dynamics of LLM Finetuning.
Yi Ren, Danica J. Sutherland.

Learning dynamics, which describes how the learning of specific training examples influences the model's predictions on other examples, gives us a powerful tool for understanding the behavior of deep learning systems. We study the learning dynamics of large language models during different types of finetuning, by analyzing the step-wise decomposition of how influence accumulates among different potential responses. Our framework allows a uniform interpretation of many interesting observations about the training of popular algorithms for both instruction tuning and preference tuning. In particular, we propose a hypothetical explanation of why specific types of hallucination are strengthened after finetuning, e.g., the model might use phrases or facts in the response for question B to answer question A, or the model might keep repeating similar simple phrases when generating responses. We also extend our framework and highlight a unique “squeezing effect" to explain a previously observed phenomenon in off-policy direct preference optimization (DPO), where running DPO for too long makes even the desired outputs less likely. This framework also provides insights into where the benefits of on-policy DPO and other variants come from. The analysis not only provides a novel perspective of understanding LLM's finetuning but also inspires a simple, effective method to improve alignment performance.

AlphaEdit: Null-Space Constrained Model Editing for Language Models.
Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, Tat-Seng Chua

Large language models (LLMs) often exhibit hallucinations, producing incorrect or outdated knowledge. Hence, model editing methods have emerged to enable targeted knowledge updates. To achieve this, a prevailing paradigm is the locating-then-editing approach, which first locates influential parameters and then edits them by introducing a perturbation. While effective, current studies have demonstrated that this perturbation inevitably disrupt the originally preserved knowledge within LLMs, especially in sequential editing scenarios. To address this, we introduce AlphaEdit, a novel solution that projects perturbation onto the null space of the preserved knowledge before applying it to the parameters. We theoretically prove that this projection ensures the output of post-edited LLMs remains unchanged when queried about the preserved knowledge, thereby mitigating the issue of disruption. Extensive experiments on various LLMs, including LLaMA3, GPT2-XL, and GPT-J, show that AlphaEdit boosts the performance of most locating-then-editing methods by an average of 36.7% with a single line of additional code for projection solely.

Honorable Mentions:

Data Shapley in One Training Run.
Jiachen T. Wang, Prateek Mittal, Dawn Song, Ruoxi Jia.

Data Shapley offers a principled framework for attributing the contribution of data within machine learning contexts. However, the traditional notion of Data Shapley requires re-training models on various data subsets, which becomes computationally infeasible for large-scale models. Additionally, this retraining-based definition cannot evaluate the contribution of data for a specific model training run, which may often be of interest in practice. This paper introduces a novel concept, In-Run Data Shapley, which eliminates the need for model retraining and is specifically designed for assessing data contribution for a particular model of interest. In-Run Data Shapley calculates the Shapley value for each gradient update iteration and accumulates these values throughout the training process. We present several techniques that allow the efficient scaling of In-Run Data Shapley to the size of foundation models. In its most optimized implementation, our method adds negligible runtime overhead compared to standard model training. This dramatic efficiency improvement makes it possible to perform data attribution for the foundation model pretraining stage. We present several case studies that offer fresh insights into pretraining data's contribution and discuss their implications for copyright in generative AI and pretraining data curation.

SAM 2: Segment Anything in Images and Videos.
Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, Christoph Feichtenhofer

We present Segment Anything Model 2 (SAM 2), a foundation model towards solving promptable visual segmentation in images and videos. We build a data engine, which improves model and data via user interaction, to collect the largest video segmentation dataset to date. Our model is a simple transformer

architecture with streaming memory for real-time video processing. SAM 2 trained on our data provides strong performance across a wide range of tasks. In video segmentation, we observe better accuracy, using 3x fewer interactions than prior approaches. In image segmentation, our model is more accurate and 6x faster than the Segment Anything Model (SAM). We believe that our data, model, and insights will serve as a significant milestone for video segmentation and related perception tasks. We are releasing our main model, the dataset, an interactive demo and code.

Faster Cascades via Speculative Decoding.
Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Seungyeon Kim, Neha Gupta, Aditya Krishna Menon, Sanjiv Kumar.

Cascades and speculative decoding are two common approaches to improving language models' inference efficiency. Both approaches interleave two models, but via fundamentally distinct mechanisms: deferral rule that invokes the larger model only for “hard” inputs, while speculative decoding uses speculative execution to primarily invoke the larger model in parallel scoring mode. These mechanisms offer different benefits: empirically, cascades offer compelling cost-quality trade-offs, often even outperforming the large model; speculative cascades offer impressive speed-ups, while guaranteeing quality-neutrality. In this paper, we leverage the best of both these approaches by designing new speculative cascading techniques that implement their deferral rule through speculative execution. We characterize the optimal deferral rule for our speculative cascades, and employ a plug-in approximation to the optimal rule. Experiments with Gemma and T5 models on a range of language benchmarks show that our approach yields better cost quality trade-offs than cascading and speculative decoding baselines.

Session recordings will be public one month after ICLR 2025 concludes.

###

Media Contact:
Jill Miley
Interprose, PR for ICLR
press@iclr.cc