Top AI Conference ICLR Announces Best Paper Award Winners and Honorable Mentions

Singapore, 23 April 2025 —The International Conference on Learning Representations (ICLR), — the premier gathering of professionals dedicated to the advancement of the many branches of artificial intelligence (AI) and deep learning—has announced the prestigious Best Paper Award Winners and Honorable Mentions during its 13th annual event at the Singapore Expo.

Selection Process

The ICLR 2025 Outstanding Paper Committee went through a two-stage selection process to identify a collection of outstanding papers and honorable mentions that showcase excellent research presented at the conference. The committee began with an initial pool of 36 papers, which were either recommended by the area chairs or received exceptional scores from reviewers. Committee members first conducted preliminary reviews to select finalists. The finalists were read by all members of the committee who ranked the papers based on factors such as theoretical insights, practical impacts, exceptional writing, and experimental rigor. The program chairs confirmed the decisions.

A total of three Outstanding Paper Winners, and three Honorable Mentions were awarded. Congratulations to all the authors for their exceptional contributions to ICLR!

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.

Abstract: 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.

  Abstract: 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

Abstract: 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.

Abstract: 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

  Abstract: 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.

  Abstract: 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.

About ICLR: The International Conference on Learning Representations (ICLR) is the premier gathering of professionals dedicated to the advancement of the branch of artificial intelligence called representation learning but generally referred to as deep learning. For more information about ICLR visit: https://iclr.cc/.

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