NEURAL INFORMATION
PROCESSING SYSTEMS

NeurIPS Announces Outstanding Paper Awards for Main Conference, Datasets and Benchmarks and Test of Time

3,540 papers were accepted out of 13,330 submitted papers reviewed by 968 Area Chairs, 98 Senior Area Chairs and 396 Ethics Reviewers

New Orleans, LA, December 11, 2023 — The 37th annual conference on Neural Information Processing Systems (NeurIPS), a premier conference in artificial intelligence (AI) and machine learning (ML), announces two Outstanding Main Track Paper Awards, two Outstanding Main Track Runner-Ups, plus two Outstanding Datasets and Benchmark Track Papers and the annual Test of Time Award. NeurIPS will be held 10-16 Dec at the New Orleans at the Ernest N. Morial Convention Center.

This year's organizers received a record number of paper submissions. Of the 13,330 submitted papers that were reviewed by 968 Area Chairs, 98 senior area chairs, and 396 Ethics reviewers, 3,540 were accepted after 502 papers were flagged for ethics reviews.

The award winning authors will present their papers during the main conference, as follows:

Outstanding Main Track Papers:

Privacy Auditing with One (1) Training Run
Authors: Thomas Steinke, Milad Nasr, Matthew Jagielski
Oral 2D Privacy: Tuesday, 12 Dec, 3:40 - 4:40 pm CST, RO6-RO9 (Level 2)
Poster session 2: Tue 12 Dec 5:15 p.m. — 7:15 p.m. CST, #1523

Abstract: We propose a scheme for auditing differentially private machine learning systems with a single training run. This exploits the parallelism of being able to add or remove multiple training examples independently. We analyze this using the connection between differential privacy and statistical generalization, which avoids the cost of group privacy. Our auditing scheme requires minimal assumptions about the algorithm and can be applied in the black-box or white-box setting. We demonstrate the effectiveness of our framework by applying it to DP-SGD, where we can achieve meaningful empirical privacy lower bounds by training only one model. In contrast, standard methods would require training hundreds of models.

Are Emergent Abilities of Large Language Models a Mirage?
Authors: Rylan Schaeffer, Brando Miranda, Sanmi Koyejo
Oral 6A LLMs: Thursday, 14 Dec, 3:20 - 4:20 pm CST, Hall C2 (Level 1)
Poster session 6: Thu 14 Dec 5:00 p.m. — 7:00 p.m. CST, #1108

Abstract: Recent work claims that large language models display emergent abilities, abilities not present in smaller-scale models that are present in larger-scale models.What makes emergent abilities intriguing is two-fold: their sharpness, transitioning seemingly instantaneously from not present to present, and their \textit{unpredictability}, appearing at seemingly unforeseeable model scales.Here, we present an alternative explanation for emergent abilities: that for a particular task and model family, when analyzing fixed model outputs, emergent abilities appear due the researcher's choice of metric rather than due to fundamental changes in model behavior with scale. Specifically, nonlinear or discontinuous metrics produce apparent emergent abilities, whereas linear or continuous metrics produce smooth, continuous, predictable changes in model performance.We present our alternative explanation in a simple mathematical model, then test it in three complementary ways: we (1) make, test and confirm three predictions on the effect of metric choice using the InstructGPT/GPT-3 family on tasks with claimed emergent abilities, (2) make, test and confirm two predictions about metric choices in a meta-analysis of emergent abilities on BIG-Bench; and (3) show how to choose metrics to produce never-before-seen seemingly emergent abilities in multiple vision tasks across diverse deep networks.Via all three analyses, we provide evidence that alleged emergent abilities evaporate with different metrics or with better statistics, and may not be a fundamental property of scaling AI models.

Outstanding Main Track Runner-Ups:

Scaling Data-Constrained Language Models

Authors: Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, Colin Raffel

Oral 2A Efficient Learning: Tuesday, 12 Dec, 3:40 - 4:40 pm CST, Hall C2 (Level 1)
Poster session 2: Tue 12 Dec 5:15 p.m. — 7:15 p.m. CST, #813

Abstract: The current trend of scaling language models involves increasing both parameter count and training dataset size. Extrapolating this trend suggests that training dataset size may soon be limited by the amount of text data available on the internet. Motivated by this limit, we investigate scaling language models in data-constrained regimes. Specifically, we run a large set of experiments varying the extent of data repetition and compute budget, ranging up to 900 billion training tokens and 9 billion parameter models. We find that with constrained data for a fixed compute budget, training with up to 4 epochs of repeated data yields negligible changes to loss compared to having unique data. However, with more repetition, the value of adding compute eventually decays to zero. We propose and empirically validate a scaling law for compute optimality that accounts for the decreasing value of repeated tokens and excess parameters. Finally, we experiment with approaches mitigating data scarcity, including augmenting the training dataset with code data or removing commonly used filters. Models and datasets from our 400 training runs are freely available at https://github.com/huggingface/datablations.

Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Authors: Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, Chelsea Finn
Oral 6B RLL: Thursday, 14 Dec, 3:20 - 4:20 pm CST, Ballroom A-C (Level 2)
Poster session 6: Thu 14 Dec 5:00 p.m. — 7:00 p.m. CST, #625

Abstract: While large-scale unsupervised language models (LMs) learn broad world knowledge and some reasoning skills, achieving precise control of their behavior is difficult due to the completely unsupervised nature of their training. Existing methods for gaining such steerability collect human labels of the relative quality of model generations and fine-tune the unsupervised LM to align with these preferences, often with reinforcement learning from human feedback (RLHF). However, RLHF is a complex and often unstable procedure, first fitting a reward model that reflects the human preferences, and then fine-tuning the large unsupervised LM using reinforcement learning to maximize this estimated reward without drifting too far from the original model. In this paper, we leverage a mapping between reward functions and optimal policies to show that this constrained reward maximization problem can be optimized exactly with a single stage of policy training, essentially solving a classification problem on the human preference data. The resulting algorithm, which we call Direct Preference Optimization (DPO), is stable, performant, and computationally lightweight, eliminating the need for fitting a reward model, sampling from the LM during fine-tuning, or performing significant hyperparameter tuning. Our experiments show that DPO can fine-tune LMs to align with human preferences as well as or better than existing methods. Notably, fine-tuning with DPO exceeds RLHF's ability to control sentiment of generations and improves response quality in summarization and single-turn dialogue while being substantially simpler to implement and train.

Outstanding Datasets and Benchmarks Papers:

In the dataset category:
ClimSim: A large multi-scale dataset for hybrid physics-ML climate emulation
Authors: Sungduk Yu, Walter Hannah, Liran Peng, Jerry Lin, Mohamed Aziz Bhouri, Ritwik Gupta, Björn Lütjens, Justus C. Will, Gunnar Behrens, Julius Busecke, Nora Loose, Charles Stern, Tom Beucler, Bryce Harrop, Benjamin Hillman, Andrea Jenney, Savannah L. Ferretti, Nana Liu, Animashree Anandkumar, Noah Brenowitz, Veronika Eyring, Nicholas Geneva, Pierre Gentine, Stephan Mandt, Jaideep Pathak, Akshay Subramaniam, Carl Vondrick, Rose Yu, Laure Zanna, Tian Zheng, Ryan Abernathey, Fiaz Ahmed, David Bader, Pierre Baldi, Elizabeth Barnes, Christopher Bretherton, Peter Caldwell, Wayne Chuang, Yilun Han, Yu Huang, Fernando Iglesias-Suarez, Sanket Jantre, Karthik Kashinath, Marat Khairoutdinov, Thorsten Kurth, Nicholas Lutsko, Po-Lun Ma, Griffin Mooers, J. David Neelin, David Randall, Sara Shamekh, Mark Taylor, Nathan Urban, Janni Yuval, Guang Zhang, Mike Pritchard
Oral 4B Dataset & Benchmarks: Wednesday, 13 Dec, 3:30 - 4:30 pm CST, Ballroom A-C (Level 2)

Poster session 4: Wed 13 Dec 5:00 p.m. — 7:00 p.m. CST, #105

Abstract: Modern climate projections lack adequate spatial and temporal resolution due to computational constraints. A consequence is inaccurate and imprecise predictions of critical processes such as storms. Hybrid methods that combine physics with machine learning (ML) have introduced a new generation of higher fidelity climate simulators that can sidestep Moore's Law by outsourcing compute-hungry, short, high-resolution simulations to ML emulators. However, this hybrid ML-physics simulation approach requires domain-specific treatment and has been inaccessible to ML experts because of lack of training data and relevant, easy-to-use workflows. We present ClimSim, the largest-ever dataset designed for hybrid ML-physics research. It comprises multi-scale climate simulations, developed by a consortium of climate scientists and ML researchers. It consists of 5.7 billion pairs of multivariate input and output vectors that isolate the influence of locally-nested, high-resolution, high-fidelity physics on a host climate simulator's macro-scale physical state.The dataset is global in coverage, spans multiple years at high sampling frequency, and is designed such that resulting emulators are compatible with downstream coupling into operational climate simulators. We implement a range of deterministic and stochastic regression baselines to highlight the ML challenges and their scoring. The data (https://huggingface.co/datasets/LEAP/ClimSim_high-res) and code (https://leap-stc.github.io/ClimSim) are released openly to support the development of hybrid ML-physics and high-fidelity climate simulations for the benefit of science and society.

In the benchmark category:

DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models

Authors: Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, Bo Li

Oral 1B Datasets & Benchmarks: Tuesday, 12 Dec, 10:00 -10:45 am CST, Ballroom A-C (Level 2)

Poster session 1: Tue 12 Dec 10:45 a.m. — 12:45 p.m. CST, #1618

Abstract: Generative Pre-trained Transformer (GPT) models have exhibited exciting progress in capabilities, capturing the interest of practitioners and the public alike. Yet, while the literature on the trustworthiness of GPT models remains limited, practitioners have proposed employing capable GPT models for sensitive applications to healthcare and finance – where mistakes can be costly. To this end, this work proposes a comprehensive trustworthiness evaluation for large language models with a focus on GPT-4 and GPT-3.5, considering diverse perspectives – including toxicity, stereotype bias, adversarial robustness, out-of-distribution robustness, robustness on adversarial demonstrations, privacy, machine ethics, and fairness. Based on our evaluations, we discover previously unpublished vulnerabilities to trustworthiness threats. For instance, we find that GPT models can be easily misled to generate toxic and biased outputs and leak private information in both training data and conversation history. We also find that although GPT-4 is usually more trustworthy than GPT-3.5 on standard benchmarks, GPT-4 is more vulnerable given jailbreaking system or user prompts, potentially due to the reason that

GPT-4 follows the (misleading) instructions more precisely. Our work illustrates a comprehensive trustworthiness evaluation of GPT models and sheds light on the trustworthiness gaps. Our benchmark is publicly available at https://decodingtrust.github.io/.

Test of Time Award

This year, following the usual practice, the organizing committee chose a NeurIPS paper from 10 years ago to receive the Test of Time Award, and "Distributed Representations of Words and Phrases and their Compositionality" by Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean, won.

Published at NeurIPS 2013 and cited over 40,000 times, the work introduced the seminal word embedding technique *word2vec*. Demonstrating the power of learning from large amounts of unstructured text, the work catalyzed progress that marked the beginning of a new era in natural language processing.

Greg Corrado and Jeffrey Dean will be giving a talk about this work and related research on Tuesday, 12 Dec at 3:05 - 3:25 pm CST.

Authors will present their accepted papers during the many poster sessions and the 77 orals. All accepted papers will have talks available to watch on demand.

The NeurIPS program was created to foster the exchange of research advancements in AI and ML, principally by hosting an annual interdisciplinary academic conference with the highest ethical standards for a diverse and inclusive community. Over the course of the seven-day conference, the schedule will include seven Invited Talks, 77 Orals, 14 Tutorials, 20 Competitions, 15 Demonstrations, nine Socials, 58 Workshops and nine Affinity Workshops.

This year's organizing committee is led by General Chairs Alice Oh (KAIST) and Tristan Naumann (Microsoft Research); Program Chairs are Amir Globerson (Tel Aviv University, Google), Kate Saenko (Boston University, Meta), Moritz Hardt (Max Planck Institute for Intelligent Systems, Tübingen) and Sergey Levine (UC Berkeley). See the full list of organizing committee members.

Events that stream and are available virtually:

Affinity Workshops, Competitions, Invited Talks, Orals, Test Of Time, Town Hall, Tutorials, Workshops

Events that do not stream:

Creative AI, Expo Day, Mentorship, Poster Sessions, Socials.

NeurIPS 2023 program will be online and available to all in late January 2024.

NEURAL INFORMATION
PROCESSING SYSTEMS

About the conference of Neural Information Processing Systems (NeurIPS)

The conference is organized by the Neural Information Processing Systems Foundation, a non-profit corporation whose purpose is to foster insights into solving difficult problems by bringing together researchers from biological, psychological, technological, mathematical and theoretical areas of science and engineering. For more information, please visit NeurIPS.cc and the NeurIPS blog for event updates.

Press Contacts:

Becky Obbema
Interprose for NeurIPS 2023
becky.obbema@interprosepr.com

NeurIPS Communication Chairs
press@neurips.cc