NEURAL INFORMATION
PROCESSING SYSTEMS

37th Annual Conference of Neural Information Processing Systems (NeurIPS)

10-16 December 2023

Fact Sheet

2023 Location: United States, Louisiana, New Orleans Ernest N. Morial Convention Center

Registration numbers:
- Total: 16,382
- In-person registrations: 13,307
- Virtual registrations: 3,075
- In-person registration was up 35% overall registration was up 6.4% while virtual registrations were down 45%

Past attendance and location:
- 15,390 Hybrid - 9,835 in-person and 5,555 virtual - New Orleans, Louisiana, United States 2022
- 17,091 Virtual Conference 2021
- 22,823 Virtual Conference 2020
- 13,000 Vancouver, British Columbia, Canada 2019
- 8,648 Montreal, Quebec, Canada 2018
- 8,008 Long Beach, California, United States 2017
- 5,231 Barcelona, Spain 2016
- 3,852 Montreal, Quebec, Canada 2015
- 2,581 Montreal, Quebec, Canada 2014
- 1,994 Lake Tahoe, California, United States 2013
- 1,676 Lake Tahoe, California, United States 2012
- 1,452 Granada, Spain 2011
- 1,354 Vancouver, British Columbia, Canada 2010

2024 Location: Vancouver, BC, Canada

2023 Program:

- Seven conference tracks
  - 7 Invited Talks - 6 keynotes, 1 panel
  - 58 Workshops
  - 14 Tutorials
  - 20 Competitions
  - 77 Orals
  - 9 Socials
  - 9 Affinity Group Workshops

- Paper reviews

- 3,540 total accepted combined papers
    - 3,218 main conference track
    - 322 datasets and benchmarks track
- 13,330 total submissions
    - 12,343 main conference track
    - 987 datasets and benchmarks (more than double the previous year's 487 submissions)
- Paper acceptance rate:
    - 26.1% main conference track
    - 32.6% datasets and benchmark
- Reviewers
    - 968 Area Chairs
    - 98 Senior Area Chairs
    - 12,974 main conference reviewers
    - 1,503 datasets and benchmark reviewers
    - 396 Ethics reviewers - 502 papers (3.77% of all submissions) were flagged for ethics review, down from 474 papers (4.37% of all submissions) in 2022
- Papers are available in the NeurIPS Proceeding - https://proceedings.neurips.cc/

- Invited Keynote Speakers
    - Bjorn Ommer - NextGenAI: The Delusion of Scaling and the Future of Generative AI
    - Lora Aroyo - The Many Faces of Responsible AI
    - Linda Smith - Coherence statistics, self-generated experience and why young humans are much smarter than current AI
    - Jelani Nelson - Sketching: core tools, learning-augmentation, and adaptive robustness
    - Panel moderator: Alexander Rush with participants: Aakanksha Chowdhery, Angela Fan, Percy Liang discussed Beyond Scaling, an LLM panel
    - Christopher Re - Systems for Foundation Models, and Foundation Models for Systems
    - Susan Murphy - Online Reinforcement Learning in Digital Health Interventions

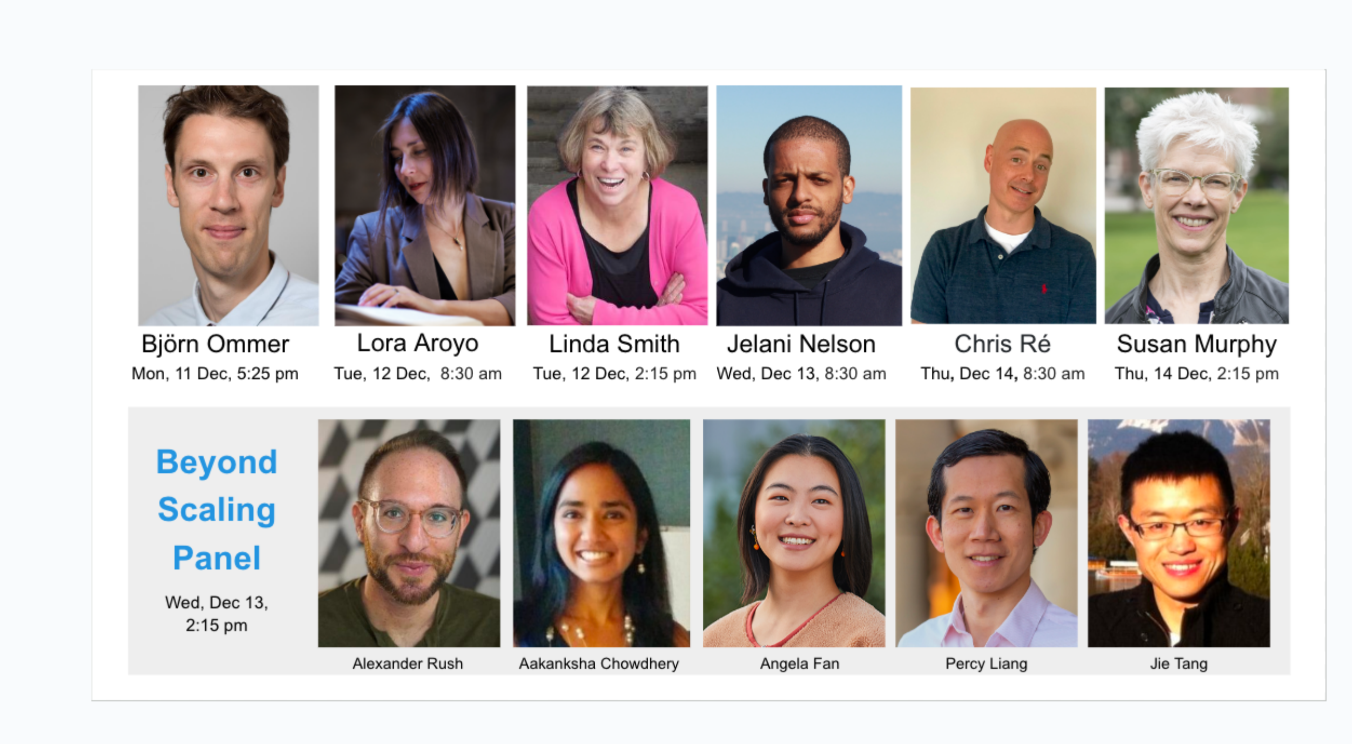
- New - Creative AI
    - Artists were invited to showcase how they use AI to create art in:
        - Two Creative AI performance
        - Three Creative AI sessions

- Educational Outreach Program
    - 94 University Students
    - 6 Local institutions - Louisiana Tech University, University of Louisiana at Lafayette, Louisiana State University, University of New Orleans, Dillard University, and Tulane University.
- Expo
    - 92 Exhibiting Sponsors

- Nine Affinity Groups Represented
    - Black in AI (https://blackinai.github.io/)
    - Global South in AI (https://businessschoolofai.teachable.com/p/globalsouthinai1)
    - Indigenous in AI (https://indigenousinai.org/)
    - LatinX in AI (https://www.latinxinai.org/events)
    - Muslims in ML (http://www.musiml.org/)
    - New in ML (https://nehzux.github.io/NewInML2023NeurIPS/)

- North Africans in ML (https://sites.google.com/view/northafricansinml)
- Queer in AI (https://www.queerinai.com/neurips-2023)
- Women in Machine Learning (https://wimlworkshop.org/)

Invited Keynote Speaker Image:



Award Recipients:

Test-of-Time Award

"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.

Two Outstanding Main Track Papers:

Privacy Auditing with One (1) Training Run

Authors: Thomas Steinke, Milad Nasr, Matthew Jagielski

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

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 to 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

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

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 Benchmark Track 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

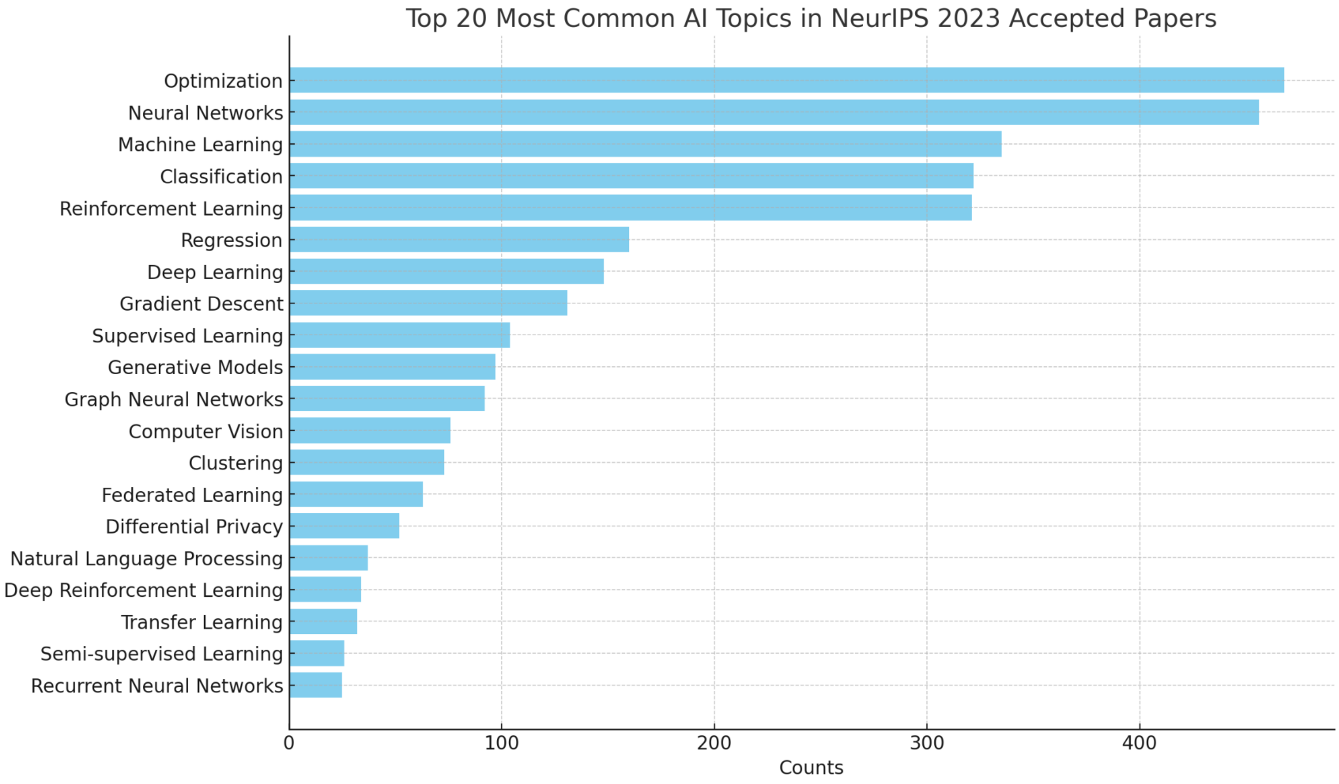
Abstract: Modern climate projections lack adequate spatial and temporal resolution due to computational constraints. A consequence is inaccurate and imprecise predictions of critical processes such as storms. Hybrid methods that combine physics with machine learning (ML) have introduced a new generation 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

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/.



Accessing NeurIPS content:
Tutorials and Invited Talks are available on the virtual site to anyone with a NeurIPS account. All virtual content will be available to the public online 30 days after the conference.

General Chairs:

- Alice Oh (KAIST)
- Tristan Naumann (Microsoft Research)

Program Chairs:

- Amir Globerson (Tel Aviv University, Google)
- Kate Saenko (Boston University, Meta)
- Moritz Hardt (Max Planck Institute for Intelligent Systems, Tübingen)
- Sergey Levine (UC Berkeley)

NeurIPS 2022 program highlights for comparison:

- 7 Invited Talks “Keynotes”
- 62 Workshops
- 10 Affinity Group sessions
- 16 Socials

- 6 Poster sessions
- 5-9 December (Week 2) virtual program
    - 13 Tutorials
    - Workshops continue
- 76 Panels
- 26 Competitions
- 2,905 accepted papers
- 9,634 full paper submissions
- 20% paper acceptance percentage
- 13 Prestigious Paper Award Winners
- 2 Datasets and Benchmark Award Winners
- Test of Time Award Winner
- 184 Main Orals
- 16 Datasets and Benchmark Orals
- 25 papers in the Competition Track
- 10,406 Reviewers
- 1,000 Top Reviewers
- Highlighted paper topics
    - neural networks
    - reinforce learning
    - language models
    - graph neural
    - federated learning
    - representation learning
    - general model
    - deep learning
    - vision transformation
    - offline reinforcement

# # #