ICLR 2024 2nd Workshop on Mathematical and Empirical Understanding of Foundation Models (ME-FoMo)

Motivation: Foundation models (FMs) have revolutionized machine learning research across domains. These models are trained on extensive, highly varied datasets and can be quickly adapted to solve many tasks of interest. FMs are extremely effective on language (e.g., GPT-3 [1], BERT [2], PaLM [3], LLaMa [17]), vision (e.g., SimCLR [4]), speech (e.g., Whisper), and multi-modal (e.g., CLIP [5], DALL-E [6]) inputs.

However, understanding of FMs lags far behind their extraordinary performance. FMs are known for their surprising emergent capabilities, such as in-context learning [1], but rigorous characterization of such phenomena is sorely lacking. Recently, substantially smaller models (e.g., LLaMA [17]) have demonstrated performance comparable to or better than huge FMs from the previous generation (e.g, OPT [19]). These findings suggest that careful selection of data, training objectives, and adaptation methods can more effectively induce desirable properties in FMs. Development of such techniques can be accelerated through better understanding.

This workshop aims to bring together researchers who work on developing an understanding of FMs, through either careful experimentation or theoretical work. Rigorous characterization of FMs can also contribute to the broader goal of mitigating undesirable behaviors. FMs are now broadly available to users, so misaligned models present real-world risk. We thus also welcome submissions of previously unpublished works that investigate how to better characterize biases in models and align them.

Topics of the workshop

The workshop will focus on three main aspects of FMs: pretraining, adaptation, and emergent capabilities. These components may include, but are not limited to, the following topics.

- **Pre-Training**: How do FMs learn useful representations?
    - Key conceptual challenge: supervised downstream tasks (e.g., solving math word problems) are often markedly different from the self-supervised pre-training objective. When and how does pre-training improve performance on a diverse set of downstream tasks?
    - Possible subtopics:
        - **Understanding the data**
            - How does the quality of the dataset impact the power of the learned representation? [26]
            - Fundamental scaling and limits: how much data do we need? Given a fixed compute budget, is it better to increase the model size or the dataset size?
            - What subsets of the data are most important for the performance and capabilities of foundation models [25, 30]?

        - **Loss Functions**
            - Vision: contrastive [4] vs. generative [32] vs. masked autoencoding [31]
            - Language: masked language modeling [2], autoregressive modeling, auxiliary objectives; tokenization methods
            - Multi-modal: contrastive objectives, translation-driven objectives
        - **Model Architecture**
            - Effect of model scale
            - Attention vs recurrence (e.g., structured state-space models [33])
            - Nonparametric or semi-parametric models: retrieval-augmented models [34]
            - Diffusion models vs autoregressive models
            - Mixture-of-experts [35]
        - **Generalization, transfer, and representation learning**
            - Role of optimization on representation learning and transfer [38]
            - Analyzing learned representations
            - Theory in simplified models
            - Training dynamics and hyperparameters at scale [35, 36, 37]
- **Adaptation**: How can we quickly adapt FMs?
    - Key conceptual challenge: FMs are trained using unlabelled data with general-purpose objectives, so how can we effectively adapt them to meaningful downstream use cases?
    - Possible subtopics:
        - **Fine-tuning, prompting, in-context learning**
            - How does fine-tuning modify the pre-trained representation?
            - Representation-based: Multimodal representation learners admit straightforward adaptation to downstream tasks through direct manipulation of the representation space (e.g., DINO [14]). How and when does this work?
            - Investigations into different prompting and decoding methods
            - Which examples should be inserted during in-context learning?
        - **Instruction Tuning**
            - What does instruction tuning do to the base model? How do models learn to generalize in this setting?
            - How can instruction tuning be made more effective [21]?
        - **Model Un-Learning and Watermarking**
            - Given data copyright concerns, there is growing interest in ensuring that a model can “un-learn” (i.e., forget) a datapoint it was pre-trained on [20]. What are effective methods for this?
            - Watermarking outputs can ensure that model generations are identifiable [22]. What types of watermarks are effective while preserving quality?
        - **Safety and Alignment**

            - Pre-trained language models are often fine-tuned to align with human preferences [18]. How does an aligned model differ from the base model?
            - How does reinforcement learning from human feedback (RLHF) work? In what cases can supervised fine-tuning achieve the same goals?
            - What are the safety deficiencies of current FMs? How can we effectively understand the internal works of FMs in order to better align them? [29]
        - **Robustness, Calibration, and Biases**
            - In what cases do FMs generalize to out-of-distribution examples? Why? How can we encourage this behavior?
            - What kinds of biases are accumulated in FMs during pre-training? How can we later remove or mitigate these biases?
        - **Efficient methods**
            - Fine-tuning often modifies a small subspace of the model parameters [27]. Do we really need scale during fine-tuning? Can fine-tuning be made more efficient?
            - Task-aware pruning and distillation methods may yield smaller, more efficient models that preserve downstream performance [28]. How do these methods work? Can we make them more effective?
- **Emergent phenomena**: Scale appears to drive qualitatively different behavior in models (e.g., in-context learning, reasoning, chain-of-thought) that can emerge suddenly during training (e.g., grokking).
    - Key conceptual challenge: We lack a rigorous understanding of what increasing the scale does to the training procedure and how these desirable emergent capabilities come about.
    - Possible subtopics:
        - **Scale-driven capabilities**
            - Chain of Thought, reasoning, in-context learning capabilities
            - Improved robustness and calibration [23]
            - Improved characterization of emergent capabilities [24]
        - **Scaling laws**
            - How and why does performance scale with data, compute, and model size? [39]
            - Grokking: how do new capabilities suddenly emerge during FM training?

Timeline for paper reviewing and acceptance: We expect the paper submission deadline to be Feb 3rd, 2024, with a review period between Feb 10th to Feb 24th. We will release all decisions to the authors by the global notification deadline of March 3, 2024.

Invited Speakers

We invited 6 speakers who span important areas of foundation models, and cover a wide range of useful perspectives to think about foundation models:

1. Mathematical / theoretical analysis: Gintare Karolina Dziugaite, Amir Globerson
2. Empirical understanding: Hannaneh Hajishirzi, Yuandong Tian, Sasha Rush
3. Alignment and safety: Sam Bowman

2 out of 6 of our speakers are women. 4 of the speakers are researchers in industry (including junior and senior researchers e.g., last authors on highly influential papers) and 5 are professors (including both assistant, associate, and full professors). Our speakers are from three different continents (North America, Asia, Europe).

- **Hannaneh Hajishirzi (Professor at University of Washington & Allen Institute for AI) [confirmed]** develops general-purpose algorithms for NLP and language models, including retrieval augmentation, instruction tuning and in-context learning, and multi-modal learning.
- **Alexander "Sasha" Rush (Associate Professor at Cornell University) [confirmed]** studies and develops data- and compute-efficient NLP systems for summarization, translation, and other generative tasks.
- **Gintare Karolina Dziugaite (Adjunct Professor at Mila, Google) [confirmed]** studies neural network pruning, generalization of neural networks, and optimization theory for deep learning.
- **Yuandong Tian (Meta) [confirmed]** studies representation learning, reinforcement learning, and optimization, including theory for contrastive learning, attention and context in language models, and long form generation.
- **Amir Globerson (Professor at Tel Aviv) [confirmed]** studies the theory of deep learning, optimization, and develops methods for language models, multimodal representation learning, and computer vision.
- **Sam Bowman (Anthropic, Associate Professor at NYU)** leads research in Anthropic and NYU about language model alignment, safety, and evaluation, including the faithfulness of language model reasoning and in-context learning, red-teaming language models, and Constitutional AI.

Organizers

Sang Michael Xie (PhD student, Stanford), Ananya Kumar (Research Scientist, OpenAI), Sewon Min (PhD student, University of Washington), Sadhika Malladi (PhD student, Princeton University), Lucio Dery (PhD student, Carnegie Mellon University), Aditi Raghunathan (Assistant Professor, CMU),Tengyu Ma (Assistant Professor, Stanford University), Percy Liang (Associate Professor, Stanford University)

Mix of established and junior researchers: We have a mix of experienced organizers, and younger researchers responsible for logistics, to ensure that the workshop is high quality. In terms of prior experience: Tengyu Ma and Aditi Raghunathan are professors at Stanford and CMU respectively, who have organized many workshops before. Percy Liang, was program committee co-chair of NeurIPS 2021 and has organized multiple workshops. Sang Michael Xie and Ananya Kumar were the main organizers of the ME-FoMo ICLR workshop in 2023. Sewon Min, Sadhika Malladi, and Lucio Dery have not organized ME-FoMo before.

Organizer background: The organizers have done significant work on foundation models, and some of their PhD theses focus on this topic. On the theory side, we've done some of the first work on theoretically analyzing fine-tuning, in-context learning, self-supervised learning, prompt tuning, and more, and have also worked on transfer learning for real world sustainability problems. Our theories and methods have also led to state-of-the-art accuracies on many datasets including ImageNet and WILDS. On the empirical understanding side, we've done work on understanding and improving language model pretraining, contrastive learning, masked language modeling, nonparameteric language models, in-context learning, and fine-tuning methods. We have organized many workshops, including the RobustML Workshop at ICLR 2021, Workshop on Formal Verification of Machine Learning at ICML 2022, ME-FoMo Workshop at ICLR 2023, NeurIPS DistShift Workshop 2021, ICML Pre-Training Workshop 2022, ACL 2023 Tutorial on retrieval-based language models, and ACL 2022 Workshop on Semiparametric NLP and Workshop on Representation Learning for NLP.

Diversity: See diversity section below.

Organizer biographies:

Tengyu Ma [website] [scholar] is an Assistant Professor of Computer Science and Statistics at Stanford University. He received his Ph.D. from Princeton University and B.E. from Tsinghua University. His research interests include topics in machine learning and algorithms, such as deep learning and its theory, non-convex optimization, deep reinforcement learning, representation learning, and high-dimensional statistics. He is a recipient of the ACM Doctoral Dissertation Award Honorable Mention, the Sloan Fellowship, and NSF CAREER Award. He has organized a workshop at the Simons Institute in Berkeley.

Ananya Kumar [website] [scholar] is a Research Scientist at OpenAI who works on understanding and improving foundation models. He has done some of the first work on theoretically analyzing fine-tuning, pretraining for robustness, and self-training. His theories and methods have led to state-of-the-art accuracies on some of the most popular benchmarks such as ImageNet and WILDS. He received his Ph.D. from Stanford University and is a recipient of the Stanford Graduate Fellowship. He co-organized the ME-FoMo Workshop at ICLR 2023.

Aditi Raghunathan [website] [scholar] is an Assistant Professor at Carnegie Mellon University. She is interested in building robust ML systems with guarantees for trustworthy real-world deployment. Previously, she was a postdoctoral researcher at Berkeley AI Research, and

received her PhD from Stanford University in 2021. Her research has been recognized by the Arthur Samuel Best Thesis Award at Stanford, a Google PhD fellowship in machine learning, and an Open Philanthropy AI fellowship. She co-organized the RobustML Workshop at ICLR 2021 and the Workshop on Formal Verification of Machine Learning at ICML 2022.

Sang Michael Xie [website] [scholar] is a sixth year PhD student at Stanford University. He has worked on data-centric methods for language model pretraining, theory and methods for understanding how pretraining improves transfer to downstream tasks, and pretraining and self-training on unlabeled data for reliable machine learning. He is a recipient of the NDSEG Fellowship. He co-organized the ME-FoMo Workshop at ICLR 2023.

Percy Liang [website] [scholar] is an Associate Professor of Computer Science at Stanford University (B.S. from MIT, 2004; Ph.D. from UC Berkeley, 2011) and the director of the Center for Research on Foundation Models. His research spans many topics in machine learning and natural language processing, including robustness, interpretability, semantics, and reasoning. He is also a strong proponent of reproducibility through the creation of CodaLab Worksheets. His awards include the Presidential Early Career Award for Scientists and Engineers (2019), IJCAI Computers and Thought Award (2016), an NSF CAREER Award (2016), a Sloan Research Fellowship (2015), a Microsoft Research Faculty Fellowship (2014), and multiple paper awards at ACL, EMNLP, ICML, and COLT. He was the program committee co-chair at NeurIPS 2021, and has organized numerous workshops including the NeurIPS DistShift Workshop 2021 and the ICML Pre-Training Workshop 2022.

Sewon Min [website] [scholar] is a sixth year Ph.D. student at the University of Washington, advised by Luke Zettlemoyer and Hannaneh Hajishirzi. Her research is in language modeling, focusing on new modeling for extensibility, better scaling and efficiency, new evaluation, and adaptations for information-seeking, legality and privacy. She co-instructed and co-organized multiple tutorials and workshops at ACL, EMNLP, NAACL and NeurIPS, including ACL 2023 Tutorial on retrieval-based language models, ACL 2022 Workshop on Semiparametric NLP and Workshop on Representation Learning for NLP. She is a recipient of the J.P. Morgan Fellowship.

Sadhika Malladi [website][scholar] is a fifth year PhD student at Princeton University, advised by Sanjeev Arora. Her research focuses on developing a rigorous understanding of optimization and representation learning in foundation models, especially in how these differ from standard learning setups.

Lucio Dery [website][scholar] is a fifth year PhD student at Carnegie Mellon University co-advised by Graham Neubig and Ameet Talwalkar. His research focuses on resource efficient methods specifically along the lines of developing compute efficient (structured pruning of pre-trained language models) and data efficient (targeted transfer learning) machine learning algorithms. Before starting his PhD, he worked as a research engineer under Luke Zettlemoyer at FAIR.

Tentative Schedule

Morning

9:15 - 9:45 Invited Talk + Q&A
9:45 - 10:15 Invited Talk + Q&A
10:15 - 10:30 Coffee Break
10:30 - 11:00 Invited Talk + Q&A
11:00 - 12:00 Spotlight Talks
12:00 - 13:00 Lunch Break

Afternoon

13:00 - 14:00 Poster Session
14:00 - 14:15 Coffee Break
14:15 - 14:45 Invited Talk + Q&A
14:45 - 15:15 Invited Talk + Q&A
15:15 - 15:45 Invited Talk + Q&A
15:45 - 16:30 Panel discussion

Format: Hybrid. We will have a mix of invited speakers (above), contributed talks, poster session, and a panel. The panel and all the talks will be livestreamed and we will allow virtual participation for the Q&A's. A potential panel moderator is Percy Liang, who is an experienced moderator and a leader in the field.

Anticipated Audience Size: 300 people (based on our estimates from organizing workshops in the past)

Funding: Upon acceptance, we will solicit sponsorships from companies such as OpenAI, Microsoft and Google DeepMind. The funds will be used for accessibility (record and stream the workshop talks), awards (best poster and talk), and to create student travel grants.

Plan to get an audience for the workshop: We will promote the workshop website through social media platforms and email lists, especially lists related to minority and underrepresented groups in AI. We will also promote the workshop and the call-for-papers in related research communities and institutional channels.

Access: We will have a website for the workshop and use it in the promotion of the workshop. The website will have links to the accepted papers. We will post the talk titles, abstracts and bios for the speakers when they are confirmed. The event will be hybrid to allow for virtual attendees to participate via livestream. After the event, we will post the slides and recordings of the talks.

Previous related workshops: This is the second ME-FoMo workshop, following the ICLR 2023 iteration. Understanding foundation models has become an even more pressing problem, and

we aim to bring together the research community towards further progress. While there have recently been a number of workshops on foundation models, including ES-FoMo: Efficient Systems for Foundation Models at ICML'23, Foundation Models for Decision Making at NeurIPS'23 & NeurIPS'22, and Trustworthy and Reliable Large-Scale Machine Learning Models Workshop at ICLR'23, none are focused on the fundamental understanding of foundation models, including pretraining, adaptation, and emergent phenomena. We believe that fundamental work on understanding foundation models can spur entirely new lines of research and is key for driving further progress for foundation models.

Conflicts of interest: No reviewer will be involved in the assessment of a paper from the same institution or organization.

Diversity Commitment

Diversity of organizers: The workshop organizers are a team of faculty, PhD students, and industry researchers, thus offering various levels of seniority. Furthermore, the organizers and speakers have varied backgrounds, life experiences, and research directions ranging from mathematical and empirical understanding to building foundation models. 3 out of 8 of the organizers are women, and various ethnic backgrounds are represented.

Diversity of speakers: We chose speakers to span important areas of foundation models: 1. Mathematical / theoretical analysis (Gintare Karolina Dzugiate, Amir Globerson), 2. Empirical understanding (Hannaneh Hajisherzi, Sasha Rush, Yuandong Tian) 3. Alignment and safety (Sam Bowman). 2 out of 6 of our speakers are women. 4 of the speakers are affiliated in industry and 5 are academic professors (3 have joint affiliations).

Getting diverse submissions: We will actively encourage diverse submissions. We will advertise the workshop to mailing lists of ML affinity groups and carefully phrase our call for participation to encourage submissions from a diverse range of researchers. We will allow non-archival submissions, follow a double-blind review process, and will not make reviews or rejected papers public. We strive to ensure diversity in our program committee and will raise a call for reviewers on ML affinity group platforms. We are understanding of the various (personal, health, financial, etc.) constraints authors may have, so we do not expect authors of accepted papers to attend in-person, although we will aim to make travel awards available for early-career researchers (see below). We will require authors who cannot attend in-person to upload a poster to the workshop website and will encourage them to optionally record a 3-min spotlight talk. We will livestream the panel and all the talks and allow virtual participation for the Q&A's.

Supporting travel for underrepresented groups: Upon acceptance, we aim to raise funds that will allow us to create student travel grants and poster printing funds. In particular we are keen to support African early career researchers and other underrepresented groups in ML research.

Friendly and welcoming environment: In the workshop we strive to create an environment that allows attendees to openly share ideas through the use of break-out, panel, and poster sessions.

Preliminary List of Potential Reviewers

We have contacted the reviewers and have gotten numerous positive responses. In last year's ME-FoMo workshop, we had over 110 reviewers.

Tomek Korbak <t.korbak@sussex.ac.uk>
Shulei Wang <shuleiw@illinois.edu>
Xindi Wu <xindiw@princeton.edu>
Yiyuan Li <bill.lyy.nisioptimum@gmail.com>
Sameera Horawalavithana <yasanka.horawalavithana@pnnl.gov>
Jianan Zhou <jianan004@e.ntu.edu.sg>
Sarah Masud <sarahmasud02@gmail.com>
Shashank Shekhar <sshkhr@fb.com>
Bilal Alsallakh <bilalsal@gmail.com>
Tianwei Yue <thithershore@gmail.com>
Saleh Elmohamed <elmohamed@cornell.edu>
Avinash Madasu <avinashmadasu17@gmail.com>
Vanya Bannihatti Kumar <vanyabk999@gmail.com>
Ashish Hooda <ahooda@wisc.edu>
Yangjun Ruan <yjruan@cs.toronto.edu>
Mattia Rigotti <mrg@zurich.ibm.vom>
Akhilesh Deepak Gotmare <dg.akhilesh@gmail.com>
Vamsi Aribandi <aribandiv@gmail.com>
Sadhika Malladi <smalladi@princeton.edu>
Putra Manggala <putra.manggala@gmail.com>
Tejasri N <ai19resch11002@iith.ac.in>
Yara Rizk <yara.rizk@ibm.com>
Lavinia F. Pieptea <laviniaflorentinapieptea@my.unt.edu>
Bingchen Zhao <zhaobc.gm@gmail.com>
Minyechil Alehegn Tefera <minyechil21@gmail.com>
Ahana Gangopadhyay <ahana@wustl.edu>
Shubham Shukla <shubhamshkl@gmail.com>
Aishwarya Balwani <abalwani6@gatech.edu>
Raj Ratn Pranesh <raj.ratn18@gmail.com>
Ellyn Ayton <ellynayton@gmail.com>
Udita Patel <uditaspatel@gmail.com>
Jinqi Luo <peter.jinqiluo@gmail.com>
Hieu Pham <hyhieu@google.com>
Adams Wei Yu <adamsyuwei@gmail.com>
Hanxiao Liu <6.hanxiao@gmail.com>
Xuanyi Dong <xuanyi.dxy@gmail.com>
Jason Wei <jason.weng.wei@gmail.com>
Xiangning Chen <xiangning@cs.ucla.edu>
Kenton Lee <kentonl@google.com>
Kelvin Guu <kguu@google.com>
Aakanksha Chowdhery <chowdhery@google.com>
Xuezhi Wang <xuezhiw@google.com>

Freda Shi <freda@ttic.edu>
Mirac Suzgun <msuzgun@cs.stanford.edu>
Yi Tay <ytay017@gmail.com>
Hyung Won Chung <h.w.chung27@gmail.com>
Jordan Hoffmann <jordanhoffmann@me.com>
Xavier Garcia <xgarcia@google.com>
Gal Kaplun <Galkaplun@g.harvard.edu>
Nikhil Vyas <vyasnikhil96@gmail.com>
Preetum Nakkiran <preetum@nakkiran.org>
Vaishaal Shankar <vs@vaishaal.com>
Kai Xiao <kaix@mit.edu>
Rui Shu <rui@openai.com>
Daniel Selsam <daniel.selsam@protonmail.com>
Iz Betagy <beltagy@allenai.org>
Pradeep Dasigi <pradeepd@allenai.org>
Kiana Ehsani <ehsanik@gmail.com>
Ben Newman <blnewman@cs.stanford.edu>
Tanmay Gupta <tanmayg@allenai.org>
Ronan Le Bras <ronanlb@allenai.org>
Kyle Lo <kylel@allenai.org>
Adam Dziedzic <adam.dziedzic@utoronto.ca>
Haonan Duan <haonand@cs.toronto.edu>
Denny Wu <dennywu@cs.toronto.edu>
Keyulu Xu <keyulux@csail.mit.edu>
Jingjing Li <jingling@cs.umd.edu>
Abraham Frandsen <abef@cs.duke.edu>
Adam Fisch <fisch@mit.edu>
Aidan Clark <aidanclark@google.com>
Aishwarya Kamath <aish@nyu.edu>
Akari Asai <akari@cs.washington.edu>
Alaa Khaddaja <alaakh@mit.edu>
Albert Gu <albertfgu@gmail.com>
Albuhair Saparov <as17582@nyu.edu>
Alex Tamkin <atamkin@stanford.edu>
Alexander Sax <sax@berkeley.edu>
Alexander Wettig <awettig@cs.princeton.edu>
Alisa Liu <alisaliu@cs.washington.edu>
Allan Jabri <ajabri@gmail.com>
Alnur Ali <alnurali@stanford.edu>
Aman Madaan <amadaan@cs.cmu.edu>
Amir Bar <amirb4r@gmail.com>
Andrew Ilyas <ailyas@mit.edu>
Annie Chen <asc8@stanford.edu>
Ari Holtzman <ahai@cs.washington.edu>
Asher Trockman <ashert@cs.cmu.edu>
Ashwini Pokle <apokle@andrew.cmu.edu>
Behnam Neyshabur <neyshabur@google.com>
Behrooz Ghorbani <ghorbani@google.com>
Ben Edelman <bedelman@g.harvard.edu>
Bhargavi Paranjape <bhargavi22294@gmail.com>
Brian Lester <blester125@gmail.com>
Charlie Snell <csnell22@berkeley.edu>
Chen Zhao <cz1285@nyu.edu>
Christina Baek <kbaek@andrew.cmu.edu>
Colin Wei <colin.y.wei@gmail.com>
Collin Burns <collinburns@berkeley.edu>

Colorado Reed <cjrd@berkeley.edu>
Dan Fu <danfu@stanford.edu>
Dan Hendrycks <hendrycks@berkeley.edu>
Daniel Levy <danilevy@cs.stanford.edu>
David Adelani <d.adelani@ucl.ac.uk>
David Wadden <dwadden@cs.washington.edu>
Derek Tam <dt.derek.tam@gmail.com>
Devin Guillory <dguillory@berkeley.edu>
Dimitris Tspiras <tsipras@stanford.edu>
Dingli Yu <dingliy@cs.princeton.edu>
Divyam Madaan <divyam.madaan@nyu.edu>
Dmitris Tsipras <tsipras@stanford.edu>
Elan Rosenfeld <elan@cmu.edu>
Ellen Wu <zeqiuwu1@uw.edu>
Eric Mitchell <eric.mitchell@cs.stanford.edu>
Eric Wallace <ericwallace@berkeley.edu>
Eric Wong <exwong@cis.upenn.edu>
Erik Jones <erjones@berkeley.edu>
Esin Durmus <esindurmus@cs.stanford.edu>
Ethan Dyer <edyer@google.com>
Frank Xu <fangzhex@cs.cmu.edu>
Gabriel Ilharco <gamaga@cs.washington.edu>
Hadi Salman <hady@mit.edu>
Hao Peng <hapeng@cs.washington.edu>
Hao Tan <haotan@adobe.com>
Haokun Liu <haokunl@cs.unc.edu>
Harold Li <liunian.harold.li@cs.ucla.edu>
Hong Liu <hliu99@stanford.edu>
Huaxiu Yao <huaxiu@cs.stanford.edu>
Ilija Radosavovic <ilija@berkeley.edu>
Jacob Springer <jspringer@cmu.edu>
Jeff Z. HaoChen <jhaochen@cs.stanford.edu>
Jeremy Cohen <jeremycohen@cmu.edu>
Jesse Dodge <jessed@allenai.org>
Jiacheng Liu <liujc@cs.washington.edu>
John Hewitt <johnhew@stanford.edu>
John Miller <miller_john@berkeley.edu>
John Thickstun <jthickstun@stanford.edu>
Jonathan Uesato <juesato@google.com>
Jong Wook Kim <jongwook@nyu.edu>
Jungo Kasai <jkasai@cs.washington.edu>
Karan Goel <kgoel@cs.stanford.edu>
Kawin Ethayarajh <kawin@stanford.edu>
Kayo Yin <kayoyin@berkeley.edu>
Kefan Dong <kefandong@stanford.edu>
Kevin Lin <kevinlin@eecs.berkeley.edu>
Kevin Miao <kevinmiao@cs.berkeley.edu>
Kurtland Chua <kchua@princeton.edu>
Lisa Dunlap <lisabdunlap@berkeley.edu>
Lucio Dery <ldery@andrew.cmu.edu>
Luyu Gao <luyug@cs.cmu.edu>
Margaret Li <margsli@cs.washington.edu>
Mayee Chen <mfchen@stanford.edu>
Mengzhou Xia <mengzhou@princeton.edu>

Michael Zhang <mzhang@cs.stanford.edu>
Michi Yasanuga <myasu@cs.stanford.edu>
Michihiro Yasunaga <myasu@cs.stanford.edu>
Mikel Artetxe <artetxe@fb.com>
Minae Kwon <minae@cs.stanford.edu>
Mitchell Wortsman <mitchellwortsman@gmail.com>
Mozes van de Kar <mozesvandekar@gmail.com>
Nikhil Kandpal <nkandpa2@cs.unc.edu>
Nikunj Saunshi <nsaunshi@cs.princeton.edu>
Niladri Chatterji <niladri@cs.stanford.edu>
Nitish Joshi <nitish@nyu.edu>
Ofir Press <ofirp@cs.washington.edu>
Pang Wei Koh <pangwei@cs.washington.edu>
Patrick Fernandes <pattuga@gmail.com>
Pengfei Liu <stefanpengfei@gmail.com>
Philip Keung <keung@amazon.com>
Pratyush Maini <pratyushmaini@cmu.edu>
Qian Huang <qhwang@cs.stanford.edu>
Rahul Nadkarni <rahuln@cs.uw.edu>
Rebecca Roelofs <rofls@google.com>
Rishi Bommasani <nlprishi@stanford.edu>
Rohan Taori <rtaori@stanford.edu>
Rohith Kuditipudi <rohithk@stanford.edu>
Ruibo Liu <ruibo.liu.gr@dartmouth.edu>
Ruiqi Zhong <ruiqi-zhong@berkeley.edu>
Ruoqi Shen <shenr3@cs.washington.edu>
Saachi Jain <saachij@mit.edu>
Sachin Goyal <sachingo@andrew.cmu.edu>
Sadhika Malladi <smalladi@princeton.edu>
Sanjay Subramanian <sanjayss@berkeley.edu>
Saurabh Garg <garg.saurabh.2014@gmail.com>
Sewon Min <sewon@cs.washington.edu>
Shaojie Bai <shaojieb@cs.cmu.edu>
Sheng Shen <sheng.s@berkeley.edu>
Shengjia Zhao <sjzhao@stanford.edu>
Shibani Santurkar <shibani@stanford.edu>
Shiori Sagawa <ssagawa@cs.stanford.edu>
Shivam Garg <shivamgarg@stanford.edu>
Shuyan Zhou <shuyanzh@cs.cmu.edu>
Sidd Karamcheti <skaramcheti@cs.stanford.edu>
Siddharth Karamchetti <skaramcheti@cs.stanford.edu>
Simon Kornblith <skornblith@google.com>
Simran Arora <simarora@stanford.edu>
Stanislav Fort <stan@anthropic.com>
Steve Mussman <mussmann@cs.washington.edu>
Suchin Gururangan <sg01@cs.washington.edu>
Surbhi Goel <goel.surbhi@microsoft.com>
Suriya Gunasekar <suriyag@microsoft.com>
Tal August <taugust@cs.washington.edu>
Terra Blevins <blvns@cs.washington.edu>
Tete Xiao <jasonhsiao97@gmail.com>

Thao Nguyen <thaonguyen@cs.stanford.edu>
Tianle Cai <tianle.cai@princeton.edu>
Tianyu Gao <tianyug@cs.princeton.edu>
Tim Brooks <tim@timothybrooks.com>
Tim Dettmers <dettmers@cs.washington.edu>
Timo Schick <schickt@cis.lmu.de>
Ting Chen <iamtingchen@google.com>
Tri Dao <trid@stanford.edu>
Vaishnavh Nagarajan <vaishnavh@google.com>
Victor Zhong <victor@victorzhong.com>
Vishakh Padmakumar <vishakh@nyu.edu>
Wei Hu <vvh@umich.edu>
Weijia Shi <swj0419@uw.edu>
Weizhe Yuan <bellyapplerian@gmail.com>
Wenya Wang <wangwy@ntu.edu.sg>
William Peebles <peebles@berkeley.edu>
Xiang Lisa Li <xlisali@stanford.edu>
Xiang Wang <xwang@cs.duke.edu>
Xianyu Yue <xyyue@ie.cuhk.edu.hk>
Xuechen Li <lxuechen@cs.stanford.edu>
Yanda Chen <yc3384@columbia.edu>
Yang Song <yangsong@caltech.edu>
Yann Dubois <yanndubs@stanford.edu>
Yash Savani <ysavani@cs.cmu.edu>
Yi Zhang <zhayi@microsoft.com>
Yining Chen <cynnjjs@stanford.edu>
Yizhong Wang <yizhongw@cs.washington.edu>
Yoonho Lee <yoonholee95@gmail.com>
Yossi Gandelsman <yossi@gandelsman.com>
Yu Sun <yusun@berkeley.edu>
Yuanzhe Pang <yzpang@nyu.edu>
Yuhuai Wu <yuhuai@stanford.edu>
Yushi Hu <yushihu@uw.edu>
Zexuan Zhong <zzhong@princeton.edu>
Zhengbao Jiang <zhengbaj@cs.cmu.edu>
Zhiyuan Li <zhiyuanli@stanford.edu>
Zhuohan Li <zhuohan@cs.berkeley.edu>
Zi-Yi Dou <zdou@ucla.edu>

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei. Language Models are Few-Shot Learners. NeurIPS 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ACL, 2019.

[3] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, Noah Fiedel. PaLM: Scaling Language Modeling with Pathways. arXiv, 2022.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. ICML, 2020.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell,
Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. arXiv 2021.

[6] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. arXiv, 2021.

[7] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. ICML 2022.

[8] Anonymous Preprint on OpenReview. How to fine-tune vision models with SGD.

[9] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, Laurent Sifre. Training Compute-Optimal Large Language Models. arXiv, 2022.

[10] Xiang Lisa Li, Percy Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. ACL 2021.

[11] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, Percy Liang. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. ICLR 2022.

[12] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An Explanation of In-context Learning as Implicit Bayesian Inference. ICLR, 2022.

[13] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, Luke Zettlemoyer. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? EMNLP, 2022.

[14] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, J. Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. arXiv, 2021.

[15] Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, Percy Liang. Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? NeurIPS 2022.

[16] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou. Chain of Thought Prompting Elicits Reasoning in Large Language Models. arXiv, 2022.

[17] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv, 2022.

[18] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, Dario Amodei. Deep reinforcement learning from human preferences. arXiv, 2017.

[19] Zhang, Susan, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan et al. "Opt: Open pre-trained transformer language models." *arXiv preprint arXiv:2205.01068* (2022).

[20] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. ACL, 2023.

[21] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, Hannaneh Hajishirzi. Self-Instruct: Aligning Language Models with Self-Generated Instructions. ACL, 2023.

[22] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, Tom Goldstein. A Watermark for Large Language Models. ICML 2023.

[23] Srivastava, Aarohi, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown et al. "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models." arXiv preprint arXiv:2206.04615 (2022).

[24] Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo. "Are emergent abilities of Large Language Models a mirage?." arXiv preprint arXiv:2304.15004 (2023).

[25] Abbas, Amro, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. "SemDeDup: Data-efficient learning at web-scale through semantic deduplication." arXiv preprint arXiv:2303.09540 (2023).

[26] Raventós, Allan, Mansheej Paul, Feng Chen, and Surya Ganguli. "Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression." arXiv preprint arXiv:2306.15063 (2023).

[27] Panigrahi, Abhishek, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. "Task-Specific Skill Localization in Fine-tuned Language Models." arXiv preprint arXiv:2302.06600 (2023).

[28] Hsieh, Cheng-Yu, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. "Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes." arXiv preprint arXiv:2305.02301 (2023).

[29] Bricken, et al., "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning", Transformer Circuits Thread, 2023.

[30] Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, Adams Wei Yu. DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining. NeurIPS, 2023.

[31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, Ross Girshick. Masked Autoencoders Are Scalable Vision Learners, CVPR, 2022.

[32] Chen et al. Generative Pretraining From Pixels, ICML 2020.

[33] Albert Gu, Karan Goel, Christopher Ré. Efficiently Modeling Long Sequences with Structured State Spaces, ICLR 2022.

[34] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, Laurent Sifre. Improving language models by retrieving from trillions of tokens, arXiv, 2022.

[35] William Fedus, Barret Zoph, Noam Shazeer. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity, JMLR, 2022.

[36] Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, Simon Kornblith. Small-scale proxies for large-scale Transformer training instabilities, arXiv, 2023.

[37] Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, Jianfeng Gao. Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer, NeurIPS 2021.

[38] Hong Liu, Sang Michael Xie, Zhiyuan Li, Tengyu Ma. Same Pre-training Loss, Better Downstream: Implicit Bias Matters for Language Models, ICML 2023.

[39] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, Dario Amodei. Scaling Laws for Neural Language Models, arXiv, 2020.