Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 24)

Antonio Valerio Miceli-Barone amiceli@ed.ac.uk
Fazl Barez fazl@robots.ox.ac.uk
Shay B. Cohen scohen@inf.ed.ac.uk
Elena Voita lenavoita@meta.com
Ulrich Germann ugermann@ed.ac.uk
Michal Lukasik mlukasik@google.com

1 Motivation

Neural Language Models generally seem to adhere to positive scaling laws: increasing the model size (in terms of the number of parameters, training data and training compute) predictably reduces token-prediction training loss, which in turn is a good predictor of performance in many downstream tasks (Kaplan et al., 2020; Hernandez et al., 2021; Hoffmann et al., 2022; Nijkamp et al., 2022). This scaling behavior has lead to Large Language Models (LLMs) becoming the dominant paradigm in Natural Language Processing, prompting speculations about all tasks becoming eventually solved by sufficiently large models. Sutton (2019) has called this the "Bitter Lesson" of AI research: simple models at scale tend to outperform attempts to implement more complex algorithms that reflect mental models of how we solve problems.

However, not all tasks follow these scaling laws. Some show *Inverse Scaling*, i.e., the performance drops with increasing model size, or a non-monotonic pattern such as *U-shaped Scaling* and *Inverse U-shaped Scaling*. Cases that defy the general trend of positive scaling include several of the BIG-Bench (Srivastava et al., 2022) benchmark tasks (several with non-monotonic scaling and some social bias tasks with inverse scaling, the latter also previously reported by Parrish et al. (2022)), unwanted memorization (Carlini et al., 2023), incorrect Python code generation under out-of-distribution (OOD) identifier redefinition (Miceli Barone et al., 2023), poor OOD compositional generalization of LLMs on a sematic parsing task (Kim et al., 2022), truthfulness of LLMs on an adversarial question answering dataset (Lin et al., 2022), LLM sensitivity to changes to the prompting format (Perez et al., 2021), etc. These research efforts, among others, culminated in the *Inverse Scaling Prize* (McKenzie et al., 2023) encouraging the community to contribute and systematically evaluate new tasks for different LLMs families. This led to the discovery and characterization of several types of tasks with inverse or non-monotonic scaling behaviors.

We believe that studying the scaling behavior of LLMs is of paramount importance in order to discover the trajectory of progress in NLP and AI. Determining whether and to what extent progress is due to investment in data and compute instead of traditional research (e.g., on different architectures and paradigms) can help the scientific community identify promising areas of research and inform the expectations of general public and decision makers. This is especially crucial now, when concerns about AI safety and its impact on society are becoming increasingly salient in the public discourse. We notice that without a central venue dedicated to promoting and sharing research on this topic, it has been mainly published on arXiv and promoted on social media, and much less commonly presented at peer-reviewed NLP and ML conferences and workshops. Our workshop aims to promote this research by providing such venue, establishing a place for the community to share their results and discuss the state of the art in this quickly developing field. While the original Inverse Scaling Prize was highly successful, it had limitations stemming from the need to be able to run automatic evaluations on all the submitted entries hence supporting only a handful of entry formats and evaluation metrics. As a result, the challenge ignored generation tasks, adaptive prompting strategies such as Chain-of-Thought (Wei et al., 2023), Least-to-Most (Zhou et al., 2023) or Tree-of-Thoughts (Yao et al., 2023a), use of external tools (Yao et al., 2023b; Schick et al., 2023; Qin et al., 2023), among others. Moreover, the prize only considered fixed model families and did not allow custom or fine-tuned models. Our aim is to elicit paper submissions not constrained to a fixed experimental format, allowing the authors to choose the task types, model

families, prompting strategies and evaluation metrics that they deem more appropriate. We also encourage submissions on studies on internal aspects of LLMs in relation with scale, such as attribution (Voita et al., 2021; Ferrando et al., 2023), mechanistic interpretation (Elhage et al. (2021); Wang et al. (2022), to name a few), modularity in computation (Csordás et al., 2021; Lepori et al., 2023), as well as research on deployment-critical metrics other than SoTA accuracy on benchmarks (e.g. uncertainty, effectiveness of pruning (Frantar and Alistarh, 2023; Sun et al., 2023), distillation, quantization (Dettmers et al., 2022; Dettmers and Zettlemoyer, 2023), etc.).

2 Main Workshop Topics

The workshop will provide focused discussions on multiple topics in the general field of Scaling behavior of Large Language Models, including, but not limited to the following:

1. Novel tasks that exhibit **Inverse**, **U-shaped**, **Inverse U-shaped** or other types of scaling;
2. Scaling behavior of fine-tuned or purpose-built models, in particular in-distribution (w.r.t. the fine-tuning dataset) vs. out-of-distribution;
3. Scaling with adaptive prompting strategies, e.g. allowing intermediate "reasoning" steps, model self-critique or use of external tools;
4. Scaling w.r.t. additional dimensions, such as the number of in-context/fine-tuning examples, the number of "reasoning" steps, or the intrinsic task complexity;
5. Scaling on non-English language tasks, in particular low-resource languages, where models might exhibit tradeoffs as high-resource language training data overwhelms low-resource language capabilities;
6. Scaling w.r.t. qualitative characteristics: internal aspects (e.g. modularity, mechanistic interpretability), calibration, uncertainty, effectiveness of various techniques (pruning, defences against adversarial attacks, etc.).

3 Workshop Format

We accept short and long paper submissions with no more than 4 and 8 pages, respectively, and an optional 4-page supplementary material. We plan to host at least one keynote talk, a few oral presentations and poster sessions during one day, as well as a panel discussion with 4-5 speakers.

4 Invited Talk

Najoung Kim agreed to give a keynote talk, either in-person or remotely. Dr. Kim is an Assistant Professor at Boston University and a researcher at Google. She is one of the authors of the Inverse Scaling Prize paper as well as other foundational works in this field.

5 Diversity

In order to promote diversity we strongly welcome contributions on non-English language tasks, in particular low-resource languages, which have been severely lacking in current research on LLM scaling. We, the organizers, are ourselves from different backgrounds, being from six different countries and being native speakers of languages other than English. We are also in different stages of our careers (Ph.D. students to research scientists and senior faculty) and we come from different institutions (both industry and academic groups).

6 Attendees

This would be the first edition of the proposed workshop, hence we do not have historical attendance data, however we ran a poll on Twitter[1] to gauge interest. The poll ran for a week from 24 July 2023, collecting 114 votes, with 79.8% asserting the need for the workshop (91 positive votes). Another data point is the original Inverse Scaling Prize which received 99 submissions. Therefore we estimate that our workshop will have 90-100 participants.

7 Special Requirements and Technical Needs

No special requirements. We would need suitable location for one or two poster sessions and an AV setup for an oral session and a panel discussion with 4-5 in-person and 4-5 remote participants.

7.1 Preferred Venues

From the most preferred to the least: ACL 2024, EMNLP 2024, EACL 2024, NAACL 2024.

[1] https://twitter.com/AVMiceliBarone/status/1683392089122390018

8 Organizers

Antonio Valerio Miceli-Barone is a senior researcher at the University of Edinburgh. His research interests center on deep learning models for machine translation and language modeling, with focus on machine translation for low-resource languages and programming code generation.

Fazl Barez is a research associate in the Department of Engineering sciences, University of Oxford and final year PhD student at Edinburgh Centre for Robotics. His research interest lies in the intersection of interpretability and safety.

Shay Cohen is a Reader at the University of Edinburgh (School of Informatics). Prior to this, he was a postdoctoral research scientist in the Department of Computer Science at Columbia University, and held an NSF/CRA Computing Innovation Fellowship. He received his B.Sc. and M.Sc. from Tel Aviv University in 2000 and 2004, and his Ph.D. from Carnegie Mellon University in 2011. His research interests span a range of topics in natural language processing and machine learning, with a focus on structured prediction (for example, parsing) and text generation.

Elena Voita is a research scientist at FAIR (Meta AI) working on analysis and interpretability for NLP. She is the current SIGREP president, served as an Area Chair at ACL, EMNLP, EACL, AACL, as a Senior Area Chair at AACL, and was a keynote speaker at several *ACL workshops (RepL4NLP, BlackBoxNLP, DeeLIO).

Ulrich Germann is a senior computing officer in the School of Informatics at the University of Edinburgh. His research interests are machine translation and neural models for NLP.

Michal Lukasik is a researcher in Google Research in New York, where his focus is on improving large scale machine learning models. Prior to this, he obtained his Ph.D. degree from the University of Sheffield. His research interests include deep retrieval and large language models.

9 Program Committee

The following 27 experts in the field have agreed to be members of the Program Committee (listed in alphabetical order):

- Adam Grycner, Google DeepMind
- Alicia Parrish, New York University / Google
- Ameya Prabhu, University of Oxford
- Andrea Madotto, Meta
- Barry Haddow, University of Edinburgh
- Biao Zhang, Google
- Colin Cherry, Google (Montreal)
- Ernest Davis, New York University
- Evelina Leivada, Autonomous University of Barcelona
- Gary Marcus, New York University
- Ian R. McKenzie, OpenAI
- Jacob Hilton, Alignment Research Center
- Julia Rayz, Purdue University
- Karl Stratos, Rutgers University
- Lucas Torroba Hennigen, CSAIL MIT
- Marcio Fonseca, University of Edinburgh
- Najoung Kim, Boston University / Google
- Parth Sarti, Stanford University
- Shihao Liang, Tsinghua University / ModelBest Inc
- Shunyu Yao, Princeton University
- Thomas Scialom, Meta
- Trevor Cohn, University of Melbourne / Google DeepMind
- Vaishnavh Nagarajan, Google Research (New York)
- Wenhu Chen, University of Waterloo / Vector Institute
- Yftah Ziser, University of Edinburgh
- Yifu Qiu, University of Edinburgh
- Yujia Qin, Tsinghua University

9.1 Submissions

Given that this is a new workshop that would hopefully turn into a series, it is difficult to assess the exact number of expected submissions. When considering the previous experience some of the workshop organizers have and the importance of the topic of the workshop, we would expect to have at least 20-30 high quality submissions, from which we will be able to comfortably cover a full day of activities. This should result in a workload of 2-3 submissions for each of our 27 reviewers. If the number of submissions exceeds our expectations, we will seek additional reviewers.

We prefer to use the OpenReview platform for the submission and review process. We do not wish to use ACL Rolling Reviews.

10 Funding

A prominent institution has already agreed to sponsor the workshop with 1,000 USD. At the time of writing other institutions have expressed interest and they are currently evaluating sponsorship options.

The funds will be used to reimburse travel expenses for the keynote speaker should she choose to participate in person, and depending on total availability, reimburse travel expenses and conference registration for a second keynote speaker, provide a best paper award or provide financial support to students from underprivileged backgrounds who would be otherwise unable to participate.

References

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models.

Róbert Csordás, Sjoerd van Steenkiste, and Jürgen Schmidhuber. 2021. Are neural nets modular? inspecting functional modularity through differentiable weight masks. In *International Conference on Learning Representations*.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*.

Tim Dettmers and Luke Zettlemoyer. 2023. The case for 4-bit precision: k-bit inference scaling laws. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 7750–7774. PMLR.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas, and Marta R. Costa-jussà. 2023. Explaining how transformers use context to build predictions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5513, Toronto, Canada. Association for Computational Linguistics.

Elias Frantar and Dan Alistarh. 2023. SparseGPT: Massive language models can be accurately pruned in one-shot. *arXiv preprint arXiv:2301.00774*.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. 2021. Scaling laws for transfer.

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, and Laurent Sifre. 2022. Training compute-optimal large language models.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Najoung Kim, Tal Linzen, and Paul Smolensky. 2022. Uncontrolled lexical exposure leads to overestimation of compositional generalization in pretrained models. *arXiv preprint arXiv:2212.10769*.

Michael A. Lepori, Thomas Serre, and Ellie Pavlick. 2023. Break it down: Evidence for structural compositionality in neural networks.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, Andrew Gritsevskiy, Daniel Wurgaft, Derik Kauffman, Gabriel Recchia, Jiacheng Liu, Joe Cavanagh, Max Weiss, Sicong Huang, The Floating Droid, Tom Tseng, Tomasz Korbak, Xudong Shen, Yuhui Zhang,

Zhengping Zhou, Najoung Kim, Samuel R. Bowman, and Ethan Perez. 2023. Inverse scaling: When bigger isn't better.

Antonio Valerio Miceli Barone, Fazl Barez, Shay B. Cohen, and Ioannis Konstas. 2023. The larger they are, the harder they fail: Language models do not recognize identifier swaps in python. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 272–292, Toronto, Canada. Association for Computational Linguistics.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. Codegen: An open large language model for code with multi-turn program synthesis.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 11054–11070. Curran Associates, Inc.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools.

Aarohi Srivastava, Abhinav Rastogi, Abhishek B Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Annasaheb Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmuller, Andrew M. Dai, Andrew D. La, Andrew Kyle Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakacs, Bridget R. Roberts, Bao Sheng Loe, Barret Zoph, Bartlomiej Bojanowski, Batuhan Ozyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Stephen Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, C'esar Ferri Ram'irez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Tatiana Ramirez, Clara Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Daniel H Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Gonz'alez, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, D. Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth P. Donoway, Ellie Pavlick, Emanuele Rodolà, Emma FC Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan J. Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fan Xia, Fatemeh Siar, Fernando Mart'inez-Plumed, Francesca Happ'e, François Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-L'opez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Han Sol Kim, Hannah Rashkin, Hanna Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hubert Wong, Ian Aik-Soon Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, John Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, J. Brooker Simon, James Koppel, James Zheng, James Zou, Jan Koco'n, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Narain Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jenni Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Oluwadara Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Jane W Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jorg Frohberg, Jos Rozen, José Hernández-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Ochieng' Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Luca Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Col'on, Luke Metz, Lutfi Kerem cSenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Madotto Andrea, Maheen Saleem Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, M Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew Leavitt, Matthias Hagen, M'aty'as Schubert, Medina Baitemirova, Melissa Arnaud,

Melvin Andrew McElrath, Michael A. Yee, Michael Cohen, Mi Gu, Michael I. Ivanitskiy, Michael Starritt, Michael Strube, Michal Swkedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Monica Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, T MukundVarma, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas S. Roberts, Nicholas Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W. Chang, Peter Eckersley, Phu Mon Htut, Pi-Bei Hwang, P. Milkowski, Piyush S. Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, QING LYU, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ram'on Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib J. Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Sam Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi S. Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo hwan Lee, Spencer Bradley Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Rose Biderman, Stephanie C. Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq A. Ali, Tatsuo Hashimoto, Te-Lin Wu, Theo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, T. N. Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler O'Brien Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, W Vossen, Xiang Ren, Xiaoyu F Tong, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yang Song, Yasaman Bahri, Ye Ji Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yu Hou, Yushi Bai, Zachary Seid, Zhao Xinran, Zhuoye Zhao, Zi Fu Wang, Zijie J. Wang, Zirui Wang, Ziyi Wu, Sahib Singh, and Uri Shaham. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv*, abs/2206.04615.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. 2023. A simple and effective pruning approach for large language models.

Richard Sutton. 2019. The bitter lesson. *Incomplete Ideas (blog)*, 13(1).

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1126–1140, Online. Association for Computational Linguistics.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models.