Evaluation for Change

Rishi Bommasani
Center for Research on Foundation Models
Stanford University
nlprishi@stanford.edu

Abstract

Evaluation is the central means for assessing, understanding, and communicating about NLP models. In this position paper, we argue evaluation should be more than that: it is a force for driving change, carrying a sociological and political character beyond its technical dimensions. As a force, evaluation's power arises from its *adoption*: under our view, evaluation succeeds when it achieves the desired change in the field. Further, by framing evaluation as a force, we consider how it competes with other forces. Under our analysis, we conjecture that the current trajectory of NLP suggests evaluation's power is *waning*, in spite of its potential for realizing more *pluralistic* ambitions in the field. We conclude by discussing the legitimacy of this power, who acquires this power and how it distributes. Ultimately, we hope the research community will more aggressively harness evaluation to drive change.

1 Introduction

Evaluation plays a defining role in NLP research; in fact, evaluation has a very rich history. While this genealogy can be traced in many ways, since this piece (roughly) coincides with the 5th anniversary of the passing of one NLP's beloved pioneers and the first recipient of the ACL Lifetime Achievement Award, we look to Aravind Joshi's legacy. Best known for grammar formalism and discourse (see Webber, 2018), his research journey reflects broader field-wide trends towards evaluation. In early works (e.g. Joshi, 1969; Joshi et al., 1972; Grosz et al., 1983), evaluation went entirely unmentioned. Yet, over time, Aravind's work involved more evaluation (e.g. Joshi and Schabes, 1989), implicitly building new norms for evaluation in grammar formalism and discourse (Miltsakaki et al., 2004; Prasad et al., 2008, 2014). Liberman (2005) cites Joshi's standards for evaluation in conveying Joshi's signature belief in multidisciplinary approaches to human language.

Joshi's life and 5 decades of scholarship teaches us evaluation is important, but how should we reason about evaluation? Here, we present two perspectives that frame evaluation in considerably different ways. Under the first account, evaluation is technical in nature, functioning as a lens to study models. The motivation for this lens may depend on the specific evaluation, stakeholder, or both: evaluation may allow us to derive scientific insight. Or it can transparently document technology for broader audiences (e.g. practitioners, colleagues in other fields, policymakers, the public). Regardless, to determine if an evaluation is successful, under this account, the lens must yield the desired understanding about models.

In this work, we argue for a second perspective, which we believe is partially acknowledged but considerably less salient than the first perspective. Under our second account, evaluation is political in nature, functioning as a force to drive change. In contrast to the first account, this means evaluation pushes the research community in some direction, possibly referring to a specific social or scientific objective, with the emphasis being on future model development more so than existing models. Critically, under this account, to determine if an evaluation is successful, the force must yield the desired change in the community. By separating these two accounts, our goal is neither to suggest they are at odds nor that they are meaningfully separable, but to shed conceptual clarity on the merits of power-centric analysis.

In pushing for this position of viewing evaluation as a force, we explore what this force influences, what other forces it competes with, how it accrues power, whether its power is legitimate, and who it empowers. Motivated by the growing impact of language technology and our field, the abundant discord on the status quo, and the uncertainty on what lies ahead, we believe evaluation's potential for change presents a vital path forward.

Findings of the Association for Computational Linguistics: ACL 2023, pages 8227–8239
July 9-14, 2023 ©2023 Association for Computational Linguistics

2 Evaluation as a Force

If evaluation is a force, what domain does it act upon? And where does its power come from?

Domain. We will restrict our scope to how evaluation influences NLP research. Specifically, evaluation concretizes desired behavior for systems, thereby communicating an objective for model design. This allows for the community to coordinate on goals for modeling research. For this goal-setting to succeed, future research should then go on to make progress on the proposed evaluation. That is, successful evaluation requires that the evaluation be prioritized, redistributing research attention such that it is allocated towards making progress on the evaluation.

Adoption constructs power. As this suggests, the adoption of an evaluation (by others) generates its power and determines its success. It is in this sense that our account for evaluation success deviates from a purely technical/intrinsic characterization. Most evaluations are concrete instantiations of a broader agenda: for these evaluations to be effective, they must shift power, namely towards addressing this agenda and materially making progress. In spite of this, we generally find that evaluations in NLP research do not even mention how adoption will arise, and if evaluation creators will take any overt actions to accelerate adoption.

Accelerating adoption. If the power of evaluations come from adoption, and evaluation creators are incentivized to accrue such power to advance their broader agenda, are there ways to accelerate adoption? We observe at least two such approaches, though they have not been considered in this way to our knowledge. As a softer means for acquiring adoption/power, evaluations may be used as shared tasks (e.g. SemEval; see Parra Escartín et al., 2017; Nissim et al., 2017) or be built as part of workshops/conferences (e.g. BIG-bench; see WELM, 2021; Srivastava et al., 2022), which leans into the relationship between coordinating research and convening researchers. More aggressively, explicit competitions with prizes or other stronger incentives can more directly drive adoption, perhaps most famously in the Netflix Prize, which remarkably accelerated and shifted research on recommender systems (see Hallinan and Striphas, 2016).

Authority as a standard. As evaluations accrue influence, they eventually become reified as high-status standards like ImageNet, WMT, and SQuAD (Dotan and Milli, 2020; Dehghani et al., 2021). While it is difficult to directly assess the power these evaluations have (e.g. how would research have changed counterfactually in their absence; see Liu et al., 2021), strong norms emerge for modeling work to evaluate on these standards. And, consequently, improvements on these evaluations function as stand-ins for more fundamental progress (Porter, 1995; Liao et al., 2021; Raji et al., 2021). In fact, their authority is made clear in how serious improvements were seen as watershed moments, ushering in new paradigms. Famous examples include the performance of AlexNet (Krizhevsky et al., 2012) on ImageNet, which initiated the deep learning revolution, and Transformers (Vaswani et al., 2017) on WMT, which, by outperforming specialized machine translation approaches by a considerable margin, marked the dawn of the current dominance of Transformers.

Related work. This work is not the first to bring questions of power, values, reflection, and change to the fore in relation to evaluation/benchmarking (Spärck Jones and Galliers, 1995; Welty et al., 2019; Dotan and Milli, 2020; Ethayarajh and Jurafsky, 2020; Linzen, 2020; Scheuerman et al., 2021; Kiela et al., 2021; Dehghani et al., 2021; Bowman and Dahl, 2021; Raji et al., 2021; Koch et al., 2021; Denton et al., 2021; Paullada et al., 2021; Liu et al., 2021; Hutchinson et al., 2021; Jacobs and Wallach, 2021; Birhane et al., 2022; Liang et al., 2022b, *inter alia*). Prior work establishes that evaluations embed values, carry influence, encode broader power structures, and the nature of evaluation as ranking aligns with broader themes of hierarchy. They make clear how other disciplines can provide guidance on what we see in NLP, but also how our evaluation practices are distinctive (e.g. competitive tendencies in benchmarking, differences in standards for measure validity).

While we draw significant inspiration from these works, our work also significantly diverges in its objective. Rather than trying to make visible the tacit assumptions, norms, and infrastructure that animate evaluation's power, we instead set our sights on how evaluation's power can animate change. In this regard, our work more closely mirrors the aesthetic of Abebe et al. (2020), as can be seen in the similar titles.

3 Competing Forces

Having argued for where evaluation draws power from, how powerful is it? While difficult to state in absolute terms, we instead consider what other forces are in play and how they interact/compete.

Coexisting forces. NLP research is a fabric stitched through myriad social interactions: conversations with colleagues, talks at conferences, academic Twitter, scholarship from adjacent disciplines, and much more. Most of these interactions are poorly conceptualized as forces: while they exert influence, they are generally diffuse rather than concentrated and lack strong directionality. For this short-form analysis, we juxtapose evaluation with the force of *resources*. By resources, we refer to assets like money, compute, and engineering support, choosing to treat them as monolithic (rather than disaggregating) for brevity.

Language models. Given the central position language models occupy in modern NLP, we consider language models as a case study to relate evaluation and resources. Our thesis is resources, to a far greater extent than evaluation, dictate research on language models, which more broadly influences NLP research given the pervasive dependence on language models. Influential language models have near-exclusively been developed by resource-rich institutions. Further, we argue a resource-allocation mindset drives decision-making in their development. Namely, the use of scaling laws (e.g. Kaplan et al., 2020; Hoffmann et al., 2022) indicates development is framed as an efficient resource allocation problem. Evaluation does play a small role here: scaling laws relate resources (x-axis) with evaluated model performance (y-axis). But the evaluation scope is narrow: scaling laws generally center accuracy for a single task (generally upstream language model perplexity/loss), with *predictability* of this relationship being the principal concern (Ganguli et al., 2022, cf. Wei et al. (2022)).

In contrast, evaluation currently does not exert similar influence over language model development. Namely, while influential language models are similar in that they were developed by resource-rich institutions, they strikingly differ in the benchmarks they are evaluated on. Across all datasets evaluated for across language modeling works at the time, Liang et al. (2022b) find that RTE is the unique dataset evaluated for in more than 50% of the 32 language modeling works they consider (e.g. GPT-3, GPT-NeoX, BLOOM, PaLM, Gopher, OPT, GLM)[1], with some works sharing no evaluation datasets in common. Given this status quo, evaluations currently fail to achieve the widespread adoption required to drive change.[2]

Contrasting properties. Which forces orient NLP research is consequential: different forces profile differently. Resources are distributed very unevenly, so resources orienting progress implies a small subset of the community expresses outsized influence in shaping the field's trajectory. Further, by the nature of how these resource disparities came to be, these resource-rich actors tend to have specific incentives (e.g. commercial interest) and demographics (e.g. poor diversity), potentially causing them to advance particular agendas (see Rogaway, 2015). In contrast, we believe evaluation structurally is better equipped to enable broader participation (e.g. BIG-bench) and, critically, pluralism. Different values can be simultaneously foregrounded in evaluations (e.g. HELM (Liang et al., 2022b) highlights values/desiderata such as accuracy, robustness, fairness, uncertainty, and efficiency). For example, insofar as scaling laws drive language model development, greater pluralism would be achieved if scaling laws were studied, fit, and applied for a broader array of evaluation targets than just upstream accuracy/perplexity.

4 Legitimacy

Since evaluation accrues power, is this power legitimate? And who does this power distribute to?

Legitimacy. Evaluations are generally built by a small number of researchers, but could orient work across the broader research community. Consequently, in arguing for the greater use of evaluation as a means for shifting power, we should question whether this implicitly recommends *value imposition*: imposing values of the few onto the many. However, recall that evaluation's power derives not from its creation but its adoption. Consequently, for this power to emerge requires the consensual action of the early adopters, who choose to use the evaluation. To an extent, this (voluntary) choice suggests that the power of evaluation is generally and, at least, initially legitimate.

[1] This is likely a direct side effect of RTE being the unique dataset in both GLUE and SuperGLUE (Wang et al., 2019a,b).

[2] Recent high-profile evaluation efforts (e.g. BIG-bench, the HuggingFace Evaluate library, HELM) may change this.

If the power of evaluation is legitimate, then what does this imply when evaluations are shown to have issues with respect to their validity, reliability, relevance, or appropriateness (Gururangan et al., 2018; Kaushik and Lipton, 2018; Ethayarajh, 2020; Blodgett et al., 2021; Aribandi et al., 2021; Birhane and Prabhu, 2021, *inter alia*)? Here, we recognize that while the initial adoption of an evaluation is in most cases clearly legitimate, the subsequent sustained adoption can be more complicated.

In particular, we emphasize that evaluations tend to exhibit *inertia*: once an evaluation is widely adopted, it is hard for the evaluation to lose this status or for other evaluations to eclipse it (e.g. due to reviewing norms; Dehghani et al., 2021), even when there are strong reasons to demote or deprecate the evaluation (Peng et al., 2021). Most directly, we point to the strong norms of comparison in NLP, whereby model developers are expected to compare their models to prior models in head-to-head comparisons. While generally a useful norm, this does promote a certain conservatism. Notably, when prior models (i.e. those that are to be compared to) are not public (Bommasani et al., 2023) or laborious to re-evaluate on new datasets, developers of new models can most easily be compare to old models on the evaluations used in prior work. In this regard, paradigms where evaluations are continuously updated and refreshed (e.g. the evaluation rounds in ANLI (Nie et al., 2020) and versions in HELM (Liang et al., 2022b); inherently dynamic evaluations like DynaBench (Kiela et al., 2021)) more directly ensure the sustained power of specific evaluations is legitimate.

Distribution of power. Even if an evaluation's power is acquired legitimately, we should further question how the power distributes over different members of the community, especially as other forces (especially resources) are inequitably distributed. Koch et al. (2021) show the distribution of evaluation developers is also uneven, aligning strongly with institutional privilege (e.g. elite academic institutions like Stanford and Princeton, massive commercial organizations like Microsoft and Google). In part, this is likely a byproduct of the fact that evaluations themselves can be quite resource-intensive, especially when this scale is a virtue: ImageNet (Deng et al., 2009), especially for its time, was exceedingly costly in both money and time; large-scale model evaluation on HELM costs almost 40k USD in addition to 20k A100 GPU hours (Liang et al., 2022b).

With that said, we have significant optimism that evaluation can realize more pluralistic visions. Specifically, (i) the rise of foundation models in NLP has shifted the field towards few-shot evaluations (Brown et al., 2020; Bragg et al., 2021), which means evaluations need not include large-scale training subsets which constituted much of the cost for evaluations historically (e.g. 80%, or 80k+, of the examples in SQuAD (Rajpurkar et al., 2016) were allocated for training). This suggests that their development should be more broadly accessible (Bommasani et al., 2021, §4.4), though the dynamics of their adoption are less clear. Further, (ii) the practice of community-driven evaluation design has been successfully implemented in several instances: the EleutherAI LM Harness (Gao et al., 2021), GEM (Gehrmann et al., 2021), GEMv2 (Gehrmann et al., 2022), BIG-Bench (Srivastava et al., 2022), the Hugging Face Evaluate library (von Werra et al., 2022), with examples like Universal Dependencies (UD; Nivre et al., 2016; de Marneffe et al., 2021) even pre-dating them for many years. In most cases, these efforts did not push a very clear directional change/agenda in research priorities (UD as a partial exception), but we believe future efforts could more explicitly exert power while learning from these prior efforts. Finally, (iii) the community has grown to more properly recognize and value evaluation-type contributions (e.g. the NeurIPS datasets and benchmarks track, cf. Rogers (2020)). That is, while we argue evaluation's power is currently waning relative to resources, suggesting a trend towards less pluralism, we simultaneously believe the conditions are ripe for renewed commitment to evaluation to reverse this trajectory.

5 Conclusion

Evaluation wields power: we believe the community is largely aware of this, yet we foreground this power to understand how evaluation drives change. This perspective leads us to three conclusions: (i) adoption imbues evaluation with its power, (ii) evaluation's power relative to other competing social forces appears to be diminishing, and yet (iii) evaluation has attractive qualities, especially under current conditions, as a force for change relative to other forces with growing power. Overall, we hope the community reflects on the mantra "evaluation for change".

Limitations

This work puts forth a position: by the nature of a position paper, the work is deliberately intended to be evocative and opinionated, in some places not having unequivocal evidence for certain claims. This presents a clear limitation: the analysis presented may diverge from the realities of NLP at present or in the future, namely if the assumptions/conditions presented themselves prove to be untrue in practice. Nonetheless, we believe centering power and change, and understanding evaluation as a political and sociological phenomenon, is likely to be useful under all conditions.

Further, in understanding the qualities of evaluation relative to other social forces, we directly suggest that evaluation is more readily operationalized in more pluralistic ways than other key forces (primarily resources). While initial efforts indicate the potential for such holistic approaches that reflect many different desiderata (Liang et al., 2022b) as well as participatory approaches that permit contribution from different entities (e.g. Srivastava et al., 2022), it is still unclear how much adoption such approaches will get, and therefore how much power they will acquire. That is, the extent to which evaluation can realize this pluralistic vision still largely remains an unresolved aspiration than a readily realizable certainty. And, conversely, we do note that while current practices potentially put pluralism and resources at odds, they may be mutually compatible in other regimes (e.g. decentralized training through the pooling of shared/volunteered compute (Yuan et al., 2022), open-source software development (Wolf et al., 2020; Gao et al., 2021; von Werra et al., 2022)).

Finally, we do not discuss other forces that we believe have not exhibited strong influence on NLP research thus far, in favor of allocating focus to evaluation and resources, which have had clear influence. To enumerate some of these other (potential) forces, we specifically note (i) research norms, (ii) policy and regulation, and (iii) auditing/advocacy. For (i), we note that while the NLP research community has many established norms (e.g. reproducibility checklists, peer review guidelines, conference organization structure, policies on respectful conduct), most of these do not directly/significantly influence what research topics different researchers work on. We do note that is possible in the future that certain norms (e.g. the access to training data or model checkpoints; Liang et al., 2022a) would influence what research is conducted (e.g. we may have not seen as much work on the learning dynamics of language models and/or memorization of training data due to the relative inaccessibility of intermediary checkpoints and training data until recently). For (ii), we note that policy and regulatory efforts have had little to no salient impact on the deployment of most language technologies, let alone NLP research, to our knowledge. With that said, much as efforts like GDPR and privacy legislation has impacted scientific research on privacy (e.g. work that operationalizes the right to be forgotten as in Ginart et al., 2019), similar trends could occur in NLP research (e.g. in response to the EU AI Act).[3] Akin to (ii), for (iii), we also have seen fairly little impact from auditing/advocacy work on NLP research to our knowledge. But, much as work on auditing/advocacy around face recognition (Buolamwini and Gebru, 2018; Raji and Buolamwini, 2019; Raji et al., 2020, *inter alia*) influenced research in the computer vision community, we could see similar trends in NLP (e.g. in response to auditing/advocacy intervention around language models).

Ethics Statement

We do not find serious risks or ethical concerns with this work. We do note this work advances a specific position, which we clearly identify. It should not be assumed there is consensus in the community (or beyond) on any account for evaluation, let alone the account on power that we espouse. In this regard, we actively solicit response and interrogation of the positions presented in this work, especially given myriad relevant analyses of evaluation/measurement/benchmarking exist in other parts of AI, computer science, linguistics, and other disciplines.

Acknowledgements

I would like to thank Alex Hanna, Ben Recht, Chris Manning, Chris Potts, Claire Cardie, Dan Jurafsky,

[3] As described in Bommasani et al. (2023), we note the NIST AI Risk Management Framework (https://www.nist.gov/itl/ai-risk-management-framework) and the mandate for NIST to develop AI testbeds under the CHIPS and Science Act (https://www.congress.gov/bill/117th-congress/house-bill/4346/text) could change this status quo in the United States. Similarly, the draft EU AI Act outlines requirements for benchmarking foundation models on public benchmarks: https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence.

Deb Raji, Emily Denton, Henrik Kugelberg, Jacob Andreas, Jacob Steinhardt, John Hewitt, Judy Shen, Kawin Ethayarajh, Nelson Liu, Rediet Abebe, Rohan Taori, Sam Bowman, Stella Biderman, Tal Linzen, Yann Dubois, and Yoav Goldberg for being specific inspirations, whose writings and thoughts helped develop my current position on evaluation, with special thanks to Percy Liang. Thanks to Jason Wei for feedback on the initial version of this work. I would like to thank the CRFM community; the experience of designing and building HELM (Liang et al., 2022b) in particular helped sharpen my belief in this philosophy towards evaluation. I am supported by the NSF Graduate Research Fellowship Program under grant number DGE-1655618.

References

Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. 2020. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 252–260, New York, NY, USA. Association for Computing Machinery.

Vamsi Aribandi, Yi Tay, and Donald Metzler. 2021. How reliable are model diagnostics? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1778–1785, Online. Association for Computational Linguistics.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 173–184, New York, NY, USA. Association for Computing Machinery.

Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. 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 1004–1015, Online. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher Rè, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *ArXiv*.

Rishi Bommasani, Daniel Zhang, Tony Lee, and Percy Liang. 2023. Improving transparency in ai language models: A holistic evaluation. *Foundation Model Issue Brief Series*.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. 2021. FLEX: Unifying evaluation for few-shot NLP. In *Advances in Neural Information Processing Systems*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. The benchmark lottery. *ArXiv*, abs/2107.07002.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. 2021. On the genealogy of machine learning datasets: A critical history of imagenet. *Big Data & Society*, 8(2):20539517211035955.

Ravit Dotan and Smitha Milli. 2020. Value-laden disciplinary shifts in machine learning. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

Kawin Ethayarajh. 2020. Is your classifier actually biased? measuring fairness under uncertainty with bernstein bounds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2914–2919, Online. Association for Computational Linguistics.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. 2022. Predictability and surprise in large generative models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1747–1764, New York, NY, USA. Association for Computing Machinery.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation. *Version v0. 0.1. Sept*.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papangelis, Aman Madaan, Angelina McMillan-Major, Anna V. Shvets, Ashish Upadhyay, Bingsheng Yao, Bryan Wilie, Chandra Bhagavatula, Chaobin You, Craig Thomson, Cristina Garbacea, Dakuo Wang, Daniel Deutsch, Deyi Xiong, Di Jin, Dimitra Gkatzia, Dragomir Radev, Elizabeth Clark, Esin Durmus, Faisal Ladhak, Filip Ginter, Genta Indra Winata, Hendrik Strobelt, Hiroaki Hayashi, Jekaterina Novikova, Jenna Kanerva, Jenny Chim, Jiawei Zhou, Jordan Clive, Joshua Maynez, João Sedoc, Juraj Juraska, Kaustubh D. Dhole, Khyathi Raghavi Chandu, Leonardo F. R. Ribeiro, Lewis Tunstall, Li Zhang, Mahima Pushkarna, Mathias Creutz, Michael White, Mihir Kale, Moussa Kamal Eddine, Nico Daheim, Nishant Subramani, Ondrej Dusek, Paul Pu Liang, Pawan Sasanka Ammanamanchi, Qinqin Zhu, Ratish Puduppully, Reno Kriz, Rifat Shahriyar, Ronald Cardenas, Saad Mahamood, Salomey Osei, Samuel Cahyawijaya, Sanja vStajner, Sébastien Montella, Shailza, Shailza Jolly, Simon Mille, Tahmid Hasan, Tianhao Shen, Tosin P. Adewumi, Vikas Raunak, Vipul Raheja, Vitaly Nikolaev, Vivian Tsai, Yacine Jernite, Yi Xu, Yisi Sang, Yixin Liu, and Yufang Hou. 2022. Gemv2: Multilingual nlg benchmarking in a single line of code. *ArXiv*, abs/2206.11249.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. 2019. Making ai forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1983. Providing a Unified Account of Definite Noun Phrases in Discourse. In *21st Annual Meeting of the Association for Computational Linguistics*, pages 44–50, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Blake Hallinan and Ted Striphas. 2016. Recommended for you: The netflix prize and the production of algorithmic culture. *New Media & Society*, 18(1):117–137.

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, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*.

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 560–575, New York, NY, USA. Association for Computing Machinery.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 375–385, New York, NY, USA. Association for Computing Machinery.

Aravind K. Joshi. 1969. Properties of Formal Grammars with Mixed Type of Rules and their Linguistic Relevance. In *International Conference on Computational Linguistics COLING 1969: Preprint No. 47*, Sånga Säby, Sweden.

Aravind K. Joshi, S. Rao Kosaraju, and H.M. Yamada. 1972. String adjunct grammars: I. local and distributed adjunction. *Information and Control*, 21(2):93–116.

Aravind K. Joshi and Yves Schabes. 1989. An evaluation of lexicalization in parsing. In *Speech and Natural Language: Proceedings of a Workshop Held at Cape Cod, Massachusetts, October 15-18, 1989*.

Jared Kaplan, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv*, abs/2001.08361.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob Gates Foster. 2021. Reduced, reused and recycled: The life of a dataset in machine learning research. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Percy Liang, Rishi Bommasani, Kathleen A. Creel, and Rob Reich. 2022a. The time is now to develop community norms for the release of foundation models.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher R'e, Diana Acosta-Navas, Drew A. Hudson, E. Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan S. Kim, Neel Guha, Niladri S. Chatterji, O. Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas F. Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda.

2022b. Holistic evaluation of language models. *ArXiv*, abs/2211.09110.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. 2021. Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Mark Liberman. 2005. Franklin Medal to Aravind Joshi. *Language Log*.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Nelson F. Liu, Tony Lee, Robin Jia, and Percy Liang. 2021. Can small and synthetic benchmarks drive modeling innovation? a retrospective study of question answering modeling approaches. ArXiv:2102.01065.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Malvina Nissim, Lasha Abzianidze, Kilian Evang, Rob van der Goot, Hessel Haagsma, Barbara Plank, and Martijn Wieling. 2017. Sharing is caring: The future of shared tasks. *Computational Linguistics*, 43(4):897–904.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Carla Parra Escartín, Wessel Reijers, Teresa Lynn, Joss Moorkens, Andy Way, and Chao-Hong Liu. 2017. Ethical considerations in NLP shared tasks. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 66–73, Valencia, Spain. Association for Computational Linguistics.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

Kenneth L Peng, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Theodore M. Porter. 1995. *Trust in Numbers*. Princeton University Press, Princeton.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the Penn Discourse TreeBank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950.

Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 429–435, New York, NY, USA. Association for Computing Machinery.

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 145–151, New York, NY, USA. Association for Computing Machinery.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Phillip Rogaway. 2015. The moral character of cryptographic work. Cryptology ePrint Archive, Paper 2015/1162. `https://eprint.iacr.org/2015/1162`.

Anna Rogers. 2020. Peer review in nlp: resource papers.

Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? disciplinary values in computer vision dataset development. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Karen Spärck Jones and Julia R. Galliers. 1995. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Number 1083 in Lecture Notes in Computer Science. Springer Verlag, Berlin.

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 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Leandro von Werra, Lewis Tunstall, Abhishek Thakur, Alexandra Sasha Luccioni, Tristan Thrush, Aleksandra Piktus, Felix Marty, Nazneen Rajani, Victor Mustar, Helen Ngo, Omar Sanseviero, Mario vSavsko, Albert Villanova, Quentin Lhoest, Julien Chaumond, Margaret Mitchell, Alexander M. Rush, Thomas Wolf, and Douwe Kiela. 2022. Evaluate&evaluation on the hub: Better best practices for data and model measurements.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Bonnie Webber. 2018. Obituary: Aravind K. Joshi. *Computational Linguistics*, 44(3):387–392.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

WELM. 2021. Workshop on Enormous Language Models (WELM).

Chris Welty, Praveen K. Paritosh, and Lora Aroyo. 2019. Metrology for ai: From benchmarks to instruments. *ArXiv*, abs/1911.01875.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Binhang Yuan, Yongjun He, Jared Quincy Davis, Tianyi Zhang, Tri Dao, Beidi Chen, Percy Liang, Christopher Re, and Ce Zhang. 2022. Decentralized training of foundation models in heterogeneous environments. In *Advances in Neural Information Processing Systems*.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations (pg 5)

☑ A2. Did you discuss any potential risks of your work?
Ethics (pg 5)

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Yes (Abstract, Intro)

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☒ Did you run computational experiments?

Left blank.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.