

The incentive gap in data work in the era of large models

Katy Ilonka Gero, Payel Das, Pierre Dognin, Inkit Padhi, Prasanna Sattigeri & Kush R. Varshney



There are repeated calls in the AI community to prioritize data work – collecting, curating, analysing and otherwise considering the quality of data. But this is not practised as much as advocates would like, often because of a lack of institutional and cultural incentives. One way to encourage data work would be to reframe it as more technically rigorous, and thereby integrate it into more-valued lines of research such as model innovation.

As artificial intelligence (AI) becomes more and more reliant on large datasets¹, there is growing recognition of the value of data quality. Performance problems – such as gender bias in text generation, or overreliance on background details in object detection – are often traced back not to modelling decisions, but to training data. Datasets may encode social biases, misrepresent or underrepresent certain classes, lose relevance over time, or otherwise not accurately reflect the phenomenon of interest. Although there is no lack of recommendations for how to collect, curate and maintain datasets², in reality AI researchers and professionals tend to use whatever datasets they can get their hands on, and once a dataset has begun to be used, it is rarely studied or modified³.

In this Comment, we consider the incentive gap in data work, and how we might better recognize and credit data work as an important and valued contribution. We first consider three reasons why AI practitioners may not do the data work: some practitioners believe ‘more data is good data’, regardless of source or quality, and prioritize collection over curation and maintenance; practitioners who do believe in the value of data work may be culturally disincentivized from doing it; and practitioners may be pressured to work at a pace that does not allow for rigorous (re)consideration of datasets. Then, given such problems, we propose reframing data work as a technical contribution, requiring sophisticated understanding of modelling details and the impact of data on downstream performance. We outline a case study demonstrating one example of ‘data work as technical contribution’, and then more explicitly articulate how to achieve this reframing.

The ‘more data is good data’ maxim

Recent trends in AI across many application domains (such as biology, vision and language) have focused on increasing the size of the model and with it the amount of training data. The maxim that ‘more data is good data’ goes back at least to early successes in automatic speech recognition in the 1970s – where the shift from using linguistic rules

to machine-learning models resulted in performance increases⁴ – and continues to enthrall AI practitioners today. This shift was largely due to more data becoming available; when data were scarce, careful curation and analysis paid dividends. But as more and more data became available, it seemed as though model training would sort out all the finicky details in the data. A focus on collecting ever larger datasets does not inherently imply a lack of interest in data work, but does typically discourage practitioners from taking on the kind of questions that data work entails, such as how well the data cover the phenomenon of interest.

One recent example of a mismatched phenomenon of interest can be found in an investigation of the consequences of injecting expressions of uncertainty into prompts for language models⁵. It was found that appending the expression ‘I’m 100% certain’ results in a dramatic drop in accuracy compared with ‘I’m 90% certain’, and hypothesized that this may be due to the usage of ‘I’m 100% certain’ in the training data for emphasis, rather than as an indicator of confidence. In this case, the phenomenon of interest (uncertainty as it relates to accuracy in question-answering systems) is not necessarily reflected in the training data.

In addition, work on scaling laws, which often argues in favour of collecting larger datasets (see, for example, ref. 6), ignores the impact and availability of data, encouraging practitioners to collect more data at essentially any cost. Such a singular focus misses out on understanding a multitude of performance issues. Progress on offensive prediction results and lower performance for underrepresented groups in image classification was made through the investigation of labelling issues in the popular image dataset ImageNet⁷. Recent work on large language models often implicitly acknowledges the importance of data quality, for instance by weighting some datasets more heavily than others during training, but fails to follow through and seriously consider the impact of data⁸.

Even if we take the ‘more data is good data’ maxim at face value, there are plenty of reasons to invest in data work. Investigating the distribution and impact of data can act as an early-warning sign for model problems that may otherwise be seen only when a model is deployed in a real-world setting. Data work does not require one to filter the data; instead, practitioners can investigate how data influence models, whether it be towards variable model performance or more tricky issues such as encoding societal biases. Available data will probably not increase forever, and as some models shift towards self-supervision, certain datasets will be more impactful than others. Acknowledging the limitations of scale will encourage a new wave of innovation that is likely to be necessary for continued AI performance.

Prestige of modelling

AI researchers often have to argue that their research is novel: applying an existing technique to a new dataset is typically not considered

enough to get published or promoted. For instance, at the NeurIPS conference, there is a separate ‘Benchmarks and Datasets’ track, indicating that data work is considered a peripheral, rather than central, aspect of AI research. Other disciplines, such as the social sciences, are often the opposite, preferring to apply existing techniques to new datasets, and putting interdisciplinary researchers in a bind⁹.

Such a phenomenon can be hard to measure, but is reflected in our experience as AI researchers. There have also been a few quantitative investigations: one study¹⁰ found that many more highly cited machine-learning papers focus on ‘performance’ and ‘novelty’ than ‘robustness’ or ‘reproducibility’, reflecting the fact that modelling and algorithmic work endows more cultural prestige than data collection and preparation. This is likely to be tied to other cultural hierarchies. The ‘hard’ sciences are often considered more difficult than the ‘soft’ ones; invention is considered more important than maintenance. (See ref. 11 for an ethnographic investigation of hierarchies, dualism, and gender in engineering.) In these terms, data work might be seen as the ‘softer’, maintenance side of AI: necessary, but not a path to individual success.

This cultural prejudice exists despite several arguments against modelling as the most important factor. The ‘Rashomon set’ argument states that lots of models perform equally well, and thus modelling work, while important, will not necessarily lead to performance improvements if the training data are held constant¹². Moreover, the performance of a model is limited by what can be learned from the data at hand; even the best model cannot learn what is not there in the first place.

We are not advocating that AI practitioners should boycott modelling work. Clearly, modelling innovation has resulted in progress. Instead, we propose that progress comes from a combination of modelling and data work, and that future innovation will be better served if both are done more equally and with more understanding of each other’s challenges.

Moving fast

AI has become a large, thriving field. This means that research and innovations come at an impressive clip, resulting in pressure to work at a pace matching that of the field as a whole. Combined with the prestige of modelling, practitioners are doubly encouraged to reuse existing datasets, such that they can spend more time on modelling innovations and making improvements to existing benchmarks¹³. When practitioners need more data, they collect them quickly and with little retrospection.

As an example, the BookCorpus has been used a multitude of times – including in highly influential language models such as BERT, GPT-3 and XLNet – but was introduced in a paper that (a) did not intend it to be used for language modelling and (b) described it in just two sentences. Seven years later, a retrospective analysis found that it contained many duplicates, copyrighted material, and probably much sexually explicit content¹⁴.

Typically, as a society, we like socially important things to move more carefully. As the technology developed by AI practitioners becomes more integrated into real-world systems, we are likely to want more work to focus on robustness and reliability, and this must include the data work. Just as modelling improvements are grounds for publication, so should be data improvements, in which existing datasets are analysed and their impact on modelling performance better understood.

Case study

In this section we briefly outline an example of a specific research agenda, to demonstrate how data work can be seen as technical work. We return to the BookCorpus, a collection of self-published books scraped from the website smashwords.com¹⁴, often described as ‘high quality’ when used to train large language models⁸. We have carried out preliminary analyses to investigate the details of this corpus from two angles: metadata attributes and hate-speech detection.

For the first, we align the BookCorpus with metadata available on smashwords.com, in particular the category tags and blurbs provided by the author. We learn that there are several differences between versions of this corpus; in particular, our version of BookCorpus included erotic content, which not all versions do¹⁴, probably because of inconsistent usage of a built-in erotic content filter provided by smashwords.com. We propose using these metadata to explicitly create versions of the data set, such as one containing all available data and one with books from certain categories (such as ‘erotic literature’) removed.

For the hate-speech detection, we ran a toxic-language-classification model¹⁵ over all sentences in the corpus, and find that the books with the highest percentage of toxic sentences tend to have notable category tags, such as ‘black comedy’, ‘humour and satire’, and ‘erotic literature’. Although hate-speech detection has several issues – such as bias against certain identity groups¹⁶ – we find that, in combination with metadata analysis, it can be a useful tool for discovering potentially problematic content.

This preliminary exploration suggests that combined use of different data-filtering techniques will provide distinct and more-comprehensive data views. In particular, future work should consider how to measure the impact of different data views. One avenue to explore would be using influence scores¹⁷ to understand how individual data points affect, say, toxic-speech detection. Another would be to investigate how the incorporation of different genres, such as romance novels, influences gender stereotypes in text generation. We see this small case study as highlighting interesting technical problems in data work – namely, identifying problematic datapoints and tracing problematic outputs back to datapoints.

How to reframe data work

We propose to encourage data work by reframing it as a technical and rigorous endeavour. This would reduce the incentive gap, and align data work with the same incentives we see for modelling innovation. We know that data influence model training and performance, but exactly how data affect downstream measures – and how we might measure these effects – is an underdeveloped yet fruitful area of study. There has been some movement on better documenting commonly used datasets¹⁴, as well as providing better access to these datasets (for example, the [ROOTS search tool](https://roots.cornell.edu))¹⁸. Still, we should encourage AI practitioners to track and study ‘data imprints’, where data documentation is combined with an understanding of modelling dynamics and performance results to understand how data ‘imprint’ on a model. Such a research agenda goes beyond collection and curation, entering the realm of analysis and discovery. We make no claim that such work is not being done, but rather suggest that it might be elevated in the AI community if its technical rigour is better appreciated.


One promising idea is that of ‘adversarial collaboration’, wherein different groups of scientists come together to design a joint experiment aimed at disproving the others’ hypotheses. We imagine that data and modelling researchers might jointly design experiments intended

to test the limits of what data or modelling can do, putting data and modelling on a more-equal footing.

Again, we argue that this work can be inherently technical and requires rigorous analysis, in addition to being extremely practical. Downstream applications can be high stakes, such as usage in health-care and financial applications, as well as novel knowledge discovery within the original datasets¹⁹. Naturally, consideration of data imprints²⁰ on the models should be easier to consider earlier, rather than after the large models have been trained and tweaked. It is time to emphasize the value of data work, not just as an endeavour that adds value, but as one that is central to the progress of artificial intelligence overall.

Katy Ilonka Gero ^{1,2} , **Payel Das** ¹ , **Pierre Dognin** ¹, **Inkit Padhi**¹, **Prasanna Sattigeri**¹ & **Kush R. Varshney** ¹ 

¹IBM Research–T. J. Watson Research Center, Yorktown Heights, NY, USA. ²Columbia University, New York, NY, USA.

 e-mail: katy@cs.columbia.edu; daspaa@us.ibm.com; krvarshn@us.ibm.com

Published online: 22 June 2023

References

1. Jiang, M., Rocktäschel, T. & Grefenstette, E. Preprint at <https://arxiv.org/abs/2211.07819> (2022).
2. Liang, W. et al. *Nat. Mach. Intell.* **4**, 669–677 (2022).
3. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. & Aroyo, L. M. in *Proc. 2021 CHI Conference on Human Factors in Computing Systems* 1–15 (Assoc. Computing Machinery, 2020).
4. Liberman, M. *Comp. Linguistics* **36**, 595–599 (2010).

5. Zhou, K., Jurafsky, D. & Hashimoto, T. Preprint at <https://arxiv.org/abs/2302.13439> (2023).
6. Kaplan, J. et al. Preprint at <https://doi.org/10.48550/arXiv.2001.08361> (2020).
7. Yang, K., Qinami, K., Fei-Fei, L., Deng, J. & Russakovsky, O. in *Proc. 2020 Conference on Fairness, Accountability, and Transparency* 547–558 (Assoc. Computing Machinery, 2020).
8. Brown, T. B. et al. in *Advances in Neural Information Processing Systems* 33 <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html> (NeurIPS, 2020).
9. Narayanan, A. *The Limits of the Quantitative Approach to Discrimination* (James Baldwin Lecture, 2022).
10. Birhane, A. et al. in *2022 ACM Conference on Fairness, Accountability, and Transparency* 173–184 (Assoc. Computing Machinery, 2022).
11. Faulkner, W. *Social Studies Sci.* **30**, 759–792 (2000).
12. Semenova, L., Rudin, C. & Parr, R. in *2022 ACM Conference on Fairness, Accountability, and Transparency* 1827–1858 (Assoc. Computing Machinery, 2022).
13. Koch, B., Denton, E., Hanna, A. & Foster, J. G. in *35th Conference on Neural Information Processing Systems* (2021).
14. Bandy, J. & Vincent, N. in *Proc. Neural Information Processing Systems Track on Datasets and Benchmarks 1* https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021 (NeurIPS Datasets and Benchmarks, 2021).
15. Caselli, T., Basile, V., Mitrović, J. & Granitzer, M. in *Proc. 5th Workshop on Online Abuse and Harms* <https://aclanthology.org/2021.woah-1.3/> (WOAH, 2021).
16. Borkan, D., Dixon, L., Sorensen, J., Thain, N. & Vasserman, L. in *Companion Proc. 2019 World Wide Web Conference* 491–500 (ACM, 2019).
17. Sattigeri, P., Ghosh, S., Padhi, I., Dognin, P., & Varshney K. in *Advances in Neural Information Processing Systems* 35 (2022).
18. Srivastava, A. et al. Preprint at <https://arxiv.org/abs/2206.04615> (2022).
19. Das, P. & Varshney, L. R. *IEEE Signal Proc. Mag.* **39**, 85–95 (2022).
20. Rothschild, A. et al. in *Proc. ACM on Human–Computer Interaction* 6 article 307 (Assoc. for Computing Machinery, 2022).

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Machine Intelligence* thanks Margaret Mitchell for their contribution to the peer review of this work.