# Too Much Annotated Data is Bad. What Can We Do About It?

By Emma Harvey (Cornell Tech)

## Too Much Annotated Data is Bad. What Can We Do About It?

**By Emma Harvery (Cornell Tech)**

Last year, I failed to use a publicly available dataset. I was working on a project where my goal was to measure the extent to which chatbots produce less helpful responses to prompts written in “non-standard” English dialects. To do that, I needed text written in those dialects. Other researchers working on the same problem were starting to use a corpus called the *AAVE/SAE Paired Dataset* for this. It contains a set of tweets that were determined by a research team to be written in African American English as well as a corresponding set of “translations” of those tweets into so-called “Standard American English” by Amazon Mechanical Turk (MTurk) workers. But when I started looking into the dataset, I realized I couldn’t use it. When I compared the original and "translated" tweets, I saw that the MTurk workers had made changes like removing swear words from the original text that had nothing to do with dialect, but would affect how a chatbot might respond to the text. In other words, I found errors in the dataset that made it unsuitable for use in evaluating an AI system – and I had to create my own novel dataset for my project instead (Harvey et al., 2025a).

I am far from the only researcher with this experience. Researchers have found that label errors in datasets widely used as model benchmarks obfuscate measures of model performance (Northcutt et al., 2021), and that benchmark datasets produced by crowdworkers contain issues including logical fallacies and failures of basic quality control and consistency (Blodgett et al., 2021). In an interview study I conducted in 2025, I found that these issues are so pervasive that researchers and practitioners often struggle to use any publicly available data at all. For example, one practitioner who I interviewed told me, “Every single public benchmark we use…has a couple of rows that we look at by eye, and

we're like, 'that doesn't make sense.' And then that makes us question the entire validity of the benchmark" (Harvey et al., 2025b).

The errors I am describing arise during the process of *data annotation*, in which data is augmented with information in order to make it suitable for use in some downstream task. Data annotation is an extremely broad process: it can include everything from experts identifying the dialect of a speaker in an audio file to crowdworkers drawing bounding boxes around cars in video data to AI models determining whether text contains hate speech or bias. Perhaps because of this breadth, while various best practices for data annotation have been proposed, few have been widely adopted. In fact, a review found that a full 30% of recent papers involving annotation tasks published at leading computational linguistics venues did not conduct, or report conducting, any form of annotation quality management (Klie et al., 2024).

"Data work"' – the time and effort that goes into data development – has long been undervalued in AI research and practice, and as a result, the field has coalesced around data development practices that prioritize efficiency and scale over quality (Paullada et al., 2021; Sambasivan et al., 2021). Data annotation is no exception. When annotation quality management is conducted, it overwhelmingly involves measuring "agreement" as a proxy for annotation quality: two annotators label the same instance and "disagreement" is treated as "error." Agreement is simple to calculate, but it has several problems. For one thing, agreement does not necessarily imply correctness. For another thing, agreement cannot be straightforwardly applied to annotation tasks that are not simply labeling (like translating tweets across dialects).

Low-quality annotation leads to widespread errors in publicly available data that cause two distinct problems. First, researchers and practitioners who mistrust public data create custom and sometimes proprietary datasets for their projects, fracturing the data landscape in AI research and making it harder to evaluate AI systems in consistent ways.

At the same time, datasets with well-documented issues continue to be widely used for model development and evaluation.

I argue that in order to improve data annotation, researchers and practitioners should treat annotation as a measurement problem. Measurement, as outlined by Adcock and Collier (2001) is a multi-stage process. The first stage, systematization, involves giving a specific, explicit definition to a concept of interest. The next stage, operationalization, involves developing processes or instruments through which to measure the systematized concept. Measurements are then produced by applying that operationalization to an instance. Wallach et al. (2025) have proposed that a critical stage in measurement is interrogation, which entails assessing the extent to which a concept's systematization, operationalization, and resulting measurements meaningfully measure the concept. Interrogation involves assessing reliability, which is similar to statistical precision and asks whether a measurement can be repeated, as well as validity, which is similar to statistical unbiasedness and asks whether a measurement is correct (Jacobs and Wallach, 2021). I argue that the failure to treat data annotation as a measurement problem – and in particular, the omission of interrogation from the annotation process – is a primary cause of widespread issues with annotation.

Framing data annotation as a measurement problem immediately makes the limitations of current approaches clear. Measurement theory proposes a plethora of approaches for evaluating measurements: Jacobs and Wallach (2021) identify two tests of reliability and seven tests of validity that are (or should be) commonly used in measurement tasks. Of those, just one test of reliability (inter-rater reliability, or agreement between annotations produced by multiple annotators) and one test of validity (predictive validity, or whether models trained on annotated data perform well on predictive tasks) are regularly used to evaluate annotations. However, across multiple interview studies that I have conducted over the past few years, almost every researcher and practitioner that I spoke to reported that they attempted to validate annotations through manual review. I consider this to be an

excellent approach for assessing the face validity of annotations, or whether annotations pass the “sniff test.”

While many researchers and practitioners conduct face validity assessments (in the form of manual review of annotations) of publicly available annotated data, they usually do not document or publish the results of those assessments. However, datasets can be improved by public efforts to identify and correct errors, as has been the case with ImageNet (Denton et al., 2021; Northcutt et al., 2021). Thus, I call for the establishment of **community face validity assessments**. I propose that this could essentially look like a face validity version of a CAPTCHA. Before an individual can access a dataset for the first time, they must manually review some small number of randomly sampled data instances and determine whether they believe that annotations are correct. This data would then be stored and displayed in the same location where the data is hosted. Over time, it could be used to develop an estimate of the error rate of the annotated data and to potentially improve existing annotated datasets.

The AI models that are increasingly mediating our online lives are only as good as the data we use. Too much of that data is currently bad, and measurement theory should be used to make it better.

## References


Robert Adcock and David Collier. 2001. Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. American Political Science Review 95, 3 (Sept. 2001), 529–546. doi:10.1017/S0003055401003100

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) (ACL). Association for Computational Linguistics, Online, 1004–1015. doi:10.18653/v1/2021.acl-long.81

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 (July 2021), 20539517211035955. doi:10.1177/20539517211035955

Emma Harvey, Rene F. Kizilcec, and Allison Koenecke. 2025a. A Framework for Auditing Chatbots for Dialect-Based Quality-of-Service Harms. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT). ACM, Athens Greece, 2025–2039. doi:10.1145/3715275.3732137

Emma Harvey, Emily Sheng, Su Lin Blodgett, Alexandra Chouldechova, Jean Garcia-Gathright, Alexandra Olteanu, and Hanna Wallach. 2025b. Understanding and Meeting Practitioner Needs When Measuring Representational Harms Caused by LLM-Based Systems. In Findings of the Association for Computational Linguistics: ACL 2025, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 18423–18440. doi:10.18653/v1/2025.findings-acl.947

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT). Association for Computing Machinery, New York, NY, USA, 375–385. doi:10.1145/3442188.3445901

Jan-Christoph Klie, Richard Eckart De Castilho, and Iryna Gurevych. 2024. Analyzing Dataset Annotation Quality Management in the Wild. Computational Linguistics 50, 3 (Sept. 2024), 817–866. doi:10.1162/coli_a_00516

Curtis G Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS). https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/f2217062e9a397a1dca429e7d70bc6ca-Paper-round1.pdf

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 (Nov. 2021), 100336. doi:10.1016/j.patter.2021.100336

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI). ACM, Yokohama Japan, 1–15. doi:10.1145/3411764.3445518

Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas J Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2025. Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge. In Forty-second International Conference on Machine Learning Position Paper Track. https://openreview.net/forum?id=1ZC4RNjqzU

**Emma Harvey**
**DLI Doctoral Fellow**
**Cornell Tech**
**evh29@cornell.edu**
**More Info >**