

Data Annotation as Measurement

Emma Harvey^{1, 2}, Allison Koenecke^{1, 2}, René F. Kizilcec²

¹Cornell Tech

²Cornell University

evh29@cornell.edu, koenecke@cornell.edu, kizilcec@cornell.edu

Abstract

Modern AI systems depend on annotated data, but annotation is rarely treated as the act of measurement that it is. Instead, annotation quality is commonly reduced to agreement: if multiple annotators assign the same annotation to a data instance, the annotations are taken to be high-quality. Yet agreement does not establish whether annotations validly capture the underlying concept they are meant to represent. In this paper, we argue that data annotation should be understood as a measurement problem. Like other forms of measurement, annotation requires defining a concept, operationalizing it through an instrument, applying that instrument, and evaluating the reliability and validity of the resulting measurements. Drawing on a literature review of annotation quality research (N=132) and semi-structured interviews with annotation team members (N=10), we develop a framework for diagnosing and correcting annotation issues. First, we map key decision points across annotation processes—including task design, annotator management, quality assessment, quality improvement, and adjudication—that shape annotation outcomes. Second, we identify five distinct sources of annotation issues: error, ambiguity, impossibility, subjectivity, and annotator identity. Annotation problems that appear similar at the level of outcomes often require different process-level interventions based on their sources. Finally, we translate measurement theory into practical guidance for annotation teams, showing how assessments of reliability and validity can move beyond agreement alone. By reframing annotation as measurement, we offer a conceptual foundation for improving the quality of annotated data used in AI research and practice.

Appendix — <https://tinyurl.com/annotation-aies2026>

1 Introduction

Modern AI systems are built and evaluated using data. This data is often produced through *data annotation*, the process of augmenting raw data so that it can be used for some downstream task. In this work, we argue that annotation should be thought of as a form of measurement in which teams systematize and operationalize concepts and apply those operationalizations to data instances. To date, various best practices for data annotation have been proposed—but, despite the centrality of annotation to AI development, few

have been widely adopted (Klie, Castilho, and Gurevych 2024). Perhaps as a result, much annotated data is of shockingly poor quality. Researchers have found that label errors in widely-used evaluation benchmarks obfuscate measures of model performance, and that datasets produced by crowdworkers contain issues including logical fallacies and failures of basic quality control and consistency (Blodgett et al. 2021; Harvey, Kizilcec, and Koenecke 2025; Northcutt, Athalye, and Mueller 2021). Harvey et al. (2025) recently found that these issues are so pervasive that researchers and practitioners often struggle to use any annotated data at all.

To date, researchers have primarily identified these issues by calculating agreement, either between annotations and “ground truth,” or between multiple annotators. While this approach can reveal issues, we argue that it is insufficient for correcting them for two primary reasons. First, correcting issues at scale requires changing annotation processes, not only annotation outcomes.¹ But annotation processes, like data work in general, are under-specified and under-documented, leaving key decision points that lead to annotation issues invisible: treated as defaults, not documented, and not interrogated (Paullada et al. 2021; Sambasivan et al. 2021; Scheuerman, Hanna, and Denton 2021). Second, annotation teams often rely on agreement among annotators as their primary quality signal (Klie, Castilho, and Gurevych 2024). Agreement is simple to calculate and useful as one indicator of reliability, but treating it as evidence of validity is a mistake: annotators may agree with one another while systematically misinterpreting the concept, applying the same flawed assumptions, or following ambiguous instructions in the same way (Catanzariti 2025; Jacobs and Wallach 2021; Zhao, Liu, and Deng 2013).

In this work, we present the results of a qualitative study intended to support the correction of data annotation issues. Informed by a semi-systematic literature review of research on annotation quality (N=132) and semi-structured interviews of researchers on annotation teams (N=10),² we make

¹Using “ground truth” to identify issues is only possible if ground truth is available, and if ground truth were available for each data instance, annotation would not be needed. Correcting issues even when ground truth is not available thus requires changing an aspect of the annotation process and re-annotating.

²Annotation teams include *annotators* and *annotation managers*, who design and manage annotation tasks.

two primary contributions. First, **we make annotation processes visible**. We map key decision points throughout the annotation process that can lead to issues, showing where annotation processes can be changed to improve annotation outcomes. As part of this, we identify five distinct sources of annotation issues: error, ambiguity, impossibility, subjectivity, and annotator identity. These issues can all manifest as a misalignment between annotations and “ground truth,” but they arise from different parts of the annotation process and therefore require different interventions.

Second, **we argue that researchers and practitioners should treat annotation as a measurement problem**. In so doing, we build on prior work that has called for measurement theory to be applied to sociotechnical research (Jacobs and Wallach 2021; Wallach et al. 2025) and provide a framework for grounding calls to move beyond agreement when evaluating data annotation (Thomas et al. 2026). Drawing on measurement theory, we show how annotation teams can evaluate not only whether annotators agree, but whether annotations are reliable and valid. These contributions shift annotation quality from an outcome-level problem that asks whether labels agree with one another or with a nominal ground truth to a measurement problem that asks whether annotation processes produce reliable and valid representations of the concepts they purport to capture.

2 Background

Annotation, a component of data development (Paullada et al. 2021), is the process of augmenting data (text, image, audio, video, etc.) so that it can be used for some downstream task. There are three types of tasks that are commonly viewed as annotation (Shmueli et al. 2021):

- *Labeling*: identification (e.g., drawing a bounding box around people in an image) and categorization (e.g., determining the species of a plant in an image)
- *Evaluation*: judging a data instance, according to a rubric (e.g., scoring an essay) or to annotator preference (e.g., selecting which AI-generated response to a prompt is better)
- *Production*: producing new data (e.g., captioning a video)

Annotation tasks range from more objective (e.g., identifying the speaker in a video) to extremely subjective (e.g., determining the intent of the speaker).

Issues in Data Annotation. “Data work”—the time and effort that goes into data development—has long been undervalued in ML research and practice (Sambasivan et al. 2021; Gray and Suri 2019), and as a result, the field has coalesced around data development practices that prioritize efficiency and scale over quality (Paullada et al. 2021). Data annotation is no exception (Scheuerman, Hanna, and Denton 2021). At the same time, it is hard to fully account for the issues present in data annotation because, as Klie, Castilho, and Gurevych (2024) point out, many papers do not describe their annotation process in any capacity—despite the wide variety of documentation practices proposed by researchers (Hutchinson et al. 2021; Gebru et al. 2021; Holland et al. 2020; Fabris et al. 2022; Díaz et al. 2022; Bender and Friedman 2018). Therefore, issues in data annotation are

typically identified by looking at the outcomes of previously conducted annotation processes. For example, researchers have identified widespread errors in labeled data (e.g., where images are misidentified) (Northcutt, Athalye, and Mueller 2021; Klie, Webber, and Gurevych 2023) as well as in produced data (e.g., where crowdsourced text deviates from annotation instructions) (Blodgett et al. 2021; Harvey, Kizilcec, and Koenecke 2025). Harvey et al. (2025) found that practitioners regularly identify label errors in publicly available annotated data, although they generally do not document or publish those findings. In other cases, issues are identified by looking at the performance of models trained or evaluated on datasets. For example, researchers have found that image classification models trained on widely-used datasets like ImageNet can fail to generalize to novel images (Recht et al. 2019), at least in part because label categories are too loosely defined (Torralba and Efros 2011).

Ground Truth. Prior work has often identified issues as occurring when annotations are misaligned with a “ground truth.” It is important to note that *ground truth* itself is nebulous: it refers to the true value of the underlying concept that an annotation seeks to measure, and for any given data instance it may be unknown or contested—and the assumption that a single ground truth even exists may be flawed (Sloane et al. 2026).³ Thus, researchers often use proxies to identify misalignment between annotations and ground truth. For example, researchers may produce “gold labels” (which are often expert annotations) for a subset of data instances and treat gold labels as ground truth (Snow et al. 2008).

Agreement. Although work reviewing annotation outcomes often compares annotations to “ground truths,” annotation processes usually focus on maximizing *agreement* or *inter-rater reliability* (IRR). Agreement refers to the extent to which multiple annotations of the same data instance produced by different annotators agree with one another (Hayes and Krippendorff 2007; Cohen 1960; McDonald, Schoenebeck, and Forte 2019). It is appealing because it is easy to calculate: as long as multiple annotations are collected for each data instance, agreement can be calculated for any labeling task whether or not ground truth is available.⁴ In many cases, annotation teams treat agreement as a proxy for ground truth. Annotation managers aggregate annotations (e.g., by keeping the majority annotation for a data instance) and then treat the aggregated label as a “silver label” (Uma et al. 2021). However, prior work has repeatedly identified limitations of and assumptions inherent to this process, including that highly reliable annotations do not necessarily align with ground truth (Zhao, Liu, and Deng 2013; Reidsma and Carletta 2008; Bayerl and Paul 2011; Thomas, Borchers, and Koedinger 2025; Arhin et al. 2021; Ali et al. 2026).

³It is possible, however, to identify misalignments between an annotation and a ground truth even when ground truth is not known. For example, it is possible to determine that an annotation identifying a photo of a tree as a strawberry is incorrect even without knowing the ground truth (i.e., what species of plant the tree is).

⁴One limitation of agreement, however, is that it is not necessarily meaningful for preference-based evaluation or production tasks.

2.1 Measurement Theory

Measurement theory from the social sciences describes the process by which underlying concepts—i.e., ground truths, which may be unobservable, theoretical, or contested—are captured in instance-level measurements (Adcock and Collier 2001; Wallach et al. 2025). Measurement, as outlined by Adcock and Collier (2001) is a multi-stage process. The first stage, *systematization*, involves giving an 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 produced for specific data instances in the *application* stage. Finally, Wallach et al. (2025) argue that measurement processes should include an *interrogation* stage, in which the reliability and validity of measurements are assessed. *Reliability* asks whether a measurement can be repeated, while *validity* asks whether a measurement is correct (Jacobs and Wallach 2021). Researchers have identified multiple aspects of reliability and validity that can be interrogated (Cronbach and Meehl 1955; Messick 1987, 1995; Cook and Beckman 2006; Jackman 2008; Jacobs and Wallach 2021). We use these aspects of reliability and validity to suggest opportunities to improve data annotation in §5.

Measurement and Data Annotation. In recent years, researchers have increasingly argued that various aspects of sociotechnical research should be viewed through the lens of measurement theory (Jacobs and Wallach 2021; Wallach et al. 2025). Building on these arguments, a large body of work has used reliability and validity to interrogate data (Blodgett et al. 2021; Eriksson et al. 2025; Gehrman, Clark, and Sellam 2023; Gosciak et al. 2026; Harvey et al. 2025; Thomas, Borchers, and Koedinger 2025; Thomas et al. 2026; Van Der Wal et al. 2024; Xiao et al. 2023; Zhao et al. 2024; Zhou et al. 2022). For example, both Blodgett et al. (2021) and Eriksson et al. (2025) use interrogation to identify major validity issues in widely-used AI and ML benchmarks (including those based on labeled or produced data), and call for increased consideration of validity during dataset development. By building on this broad base of prior work to explicitly treat data annotation as a measurement problem, we offer a framework to ground existing and future work on evaluating data annotation.

3 Methods

Our qualitative methods followed an iterative and highly intertwined approach. First, we conducted a semi-systematic literature review of papers related to annotation quality (§3.1). We found two categories of papers: those that focused on annotation *processes* and those that focused on annotation *outcomes*. Within the papers that focused on processes, we found that papers tended to either propose “best practices” or survey “current practices” (at least to the extent that current practices are documented in published work). However, we struggled to find papers that described *how* annotation teams apply current or best practices—i.e., what decisions annotation teams make throughout the annotation process that affect the quality of annotations. Thus, we supplemented our literature review with a small-scale

semi-structured interview study of researchers on annotation teams in order to surface those accounts (§3.2).

3.1 Literature Review

We conducted a semi-systematic literature review of academic work related to annotation quality.⁵ We conducted a keyword search of the ACM, ACL, IEEE, and NeurIPS paper databases for terms related to annotation (annotat* | label* | coding) and quality (subjectiv* | objectiv* | ambigu* | eval* | valid* | reliab* | *agree* | bias* | ground truth | error*), requiring that keywords representing both annotation and quality were present in the title or abstract of the work. We selected relevant works by manually reviewing and filtering papers, first by title and then by abstract. Inclusion criteria for relevance were:

- Being about annotation specifically (as opposed to, e.g., crowdwork in general); and
- Evaluating, critiquing, or improving annotation outcomes (as opposed to, e.g., producing annotations); or
- Proposing or documenting annotation practices.

We supplemented this using a natural language search for “barriers or best practices in data annotation” using A12 Asta.⁶ Finally, to minimize selection bias, we reviewed the citations of papers surfaced by our search and added work that met the inclusion criteria for title and abstract relevance to our pool. Overall, we reviewed 132 publications spanning 1972 to 2026.

We categorized papers as focused on annotation processes or annotation outcomes. Papers focused on annotation processes either proposed “best practices” or surveyed “current practices,” while papers focused on annotation outcomes evaluated the quality of existing annotated data. We synthesized process-focused papers to develop a high-level understanding of the stages of the annotation process, which we used to scaffold our semi-structured interviews. We categorized outcome-focused papers according to their evaluation approach, attempting to map evaluation approaches to one or more of the reliability and validity aspects described in §2.1.

3.2 Interview Study

Between November and December of 2025, we conducted 10 semi-structured interviews with individuals who have been involved in data annotation projects. We recruited participants through purposive and snowball sampling in order to recruit individuals with varied experience with and opinions on data annotation. Interviews were 60 minutes long and were conducted and recorded via Zoom. Participants provided informed consent prior to the start of the

⁵A semi-systematic literature review is a research method that is designed for tasks like providing an overview of research that spans diverse disciplines (like data annotation). While semi-systematic literature reviews do not necessarily require strictly systematic search and research strategies (e.g., following PRISMA guidelines), they nevertheless require that strategies are transparently presented and well justified (Snyder 2019).

⁶<https://asta.allen.ai/>

	Role		Medium			Domain			# Projects			Race			Gender	
	Annotation Manager	Annotator	Text	Image	Video	General-Purpose ML	Education	Law	1 to 5	6 to 10	more than 10	Asian	Black	White	Male	Female
<i>P1</i>	✓	✓	✓		✓		✓			✓				✓	✓	
<i>P2</i>	✓	✓	✓					✓	✓					✓	✓	
<i>P3</i>	✓	✓	✓					✓	✓					✓		✓
<i>P4</i>	✓	✓	✓	✓		✓			✓			✓				✓
<i>P5</i>	✓	✓	✓					✓	✓					✓	✓	
<i>P6</i>	✓	✓	✓					✓	✓					✓		✓
<i>P7</i>	✓	✓	✓				✓		✓				✓	✓	✓	
<i>P8</i>		✓	✓					✓	✓			✓				✓
<i>P9</i>	✓	✓	✓	✓		✓					✓	✓			✓	
<i>P10</i>	✓	✓	✓	✓			✓				✓			✓		✓
Totals	9	10	10	3	1	2	3	5	7	1	2	3	1	7	5	5

Table 1: Characteristics of interview participants, including demographics and prior experience with annotation projects. Note that participants could report multiple roles, data mediums, domains, and racial identities.

interview, and all participants were compensated for their time with a \$50 gift card. The study was reviewed by an Institutional Review Board (IRB).

Participants. All participants have taken on annotator roles in prior projects, and all but one have also taken on annotation manager roles. Participants work across a variety of domains, including education, law, and general-purpose ML. All participants hold research roles, although several described working on annotation projects within industry labs or for non-academic reasons (e.g., for the purposes of mounting a legal appeal). All participants are based in North America. Additional background information about interview participants is provided in Table 1.

Interviews. To start, participants were asked to briefly list the data annotation projects that they had been involved with. They were then asked to describe each project, including the annotation goal, data medium (e.g., text, image, video), scale (e.g., data size, project timeline), and annotator population (e.g., experts, trained annotators, crowdworkers, AI). Participants were asked to describe their annotation processes in detail, focusing on key decision points and challenges that they faced throughout. We initially asked open-ended questions (e.g., “can you describe the data annotation process on your project?”) and asked more specific follow-ups as needed (e.g., “what tools or platforms did you use to collect annotations?”). For all decision points and challenges mentioned, participants were asked what action they took and what they thought the effects of that action on their resulting annotations were. The full semi-structured interview guide is available in OSF Appendix A.

Thematic Analysis. We then conducted a reflexive thematic analysis on the resulting interview transcripts, following an inductive-deductive coding approach (Braun and Clarke 2006, 2019). The first author, in discussion with

the other authors, developed an initial set of deductive codes that focused on key decision points related to misalignment between annotations and a “ground truth,” including how participants identified misalignments and what they did in their annotation processes to address them. The first author then manually reviewed the transcripts to identify the presence of those codes. While doing so, the first author identified additional key decision points raised by participants and inductively added them to the set of codes. Finally, the first author recoded all transcripts with the full set of codes.

Positionality. Our approach to thematic analysis is shaped by our own positionality. We are a team of academic researchers with backgrounds in computer science, data science, statistics, and philosophy. Before their career in academia, the first author worked in industry as a data scientist, and, as a result, personally values data work and is liable to notice and question when it is implicitly undervalued. At the same time, although all members of the research team have served as annotation managers or personally annotated data for various research projects, none have ever worked as crowdworkers or as full-time data annotators. Furthermore, because we are based in the US and English is the only language shared across all members of the research team, our interviews were conducted in English and our analysis is US-centric.

Limitations. We acknowledge several limitations of our approach. In particular, our interview study relied on a small (though in line with typical sample sizes for similar research; see Caine (2016)) and non-representative participant pool. We believe our interview study is nevertheless valuable as a supplement to our literature review, which lacked narrative accounts of decision points and associated challenges arising during annotation projects. Although our participant pool is not representative of all annotators

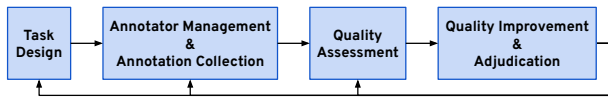


Figure 1: The high-level data annotation process consists of iterative stages of Task Design, Annotator Management & Annotation Collection, Quality Assessment, and Quality Improvement & Adjudication.

or annotation managers, it is not intended to be: we used targeted recruiting to reach out to individuals who we knew had varied experiences with and opinions on annotation in order to capture a breadth of perspectives.

4 Making Annotation Processes Visible

In this section, we focus on annotation processes. Drawing on our literature review, we provide a brief overview of the high-level annotation process (§4.1). Then, drawing primarily on our interview study, we provide a more detailed map of key decision points that occur throughout that process and that can lead to annotation issues (§4.2). Finally, drawing on both studies, we identify distinct sources of annotation issues stemming from different decision points (§4.3).

4.1 High-Level Annotation Process

There are multiple disciplines across which data annotation is conducted: while some projects draw on principles of qualitative coding from the social sciences, others view annotation as a more straightforward crowdsourcing challenge—and as AI-based annotation becomes widespread, annotation is increasingly seen as a prediction task (Callison-Burch and Dredze 2010; He et al. 2024; MacQueen et al. 1998). Different disciplines follow slightly different annotation processes, but all generally include the four stages of Task Design, Annotator Management & Annotation Collection, Quality Assessment, and Quality Improvement & Adjudication, as shown in Figure 1 (Beck 2023; Daniel et al. 2019; Fereday and Muir-Cochrane 2006; Klie, Castilho, and Gurevych 2024; MacQueen et al. 1998; Sabou et al. 2014; Törnberg 2024).

Task Design includes selecting data, writing a codebook or annotation guidelines, and planning what the annotation process will look like. *Annotator Management & Annotation Collection* includes selecting an annotator population, training annotators, and collecting annotations. *Quality Assessment* includes determining whether already-collected annotations are of sufficiently high quality. If quality is not sufficiently high, *Quality Improvement* involves changing either annotation processes or annotation outcomes to improve annotation quality; otherwise, *Adjudication* involves selecting the final set of annotations from all those collected.

4.2 Key Decision Points in Annotation Processes

Each stage of the annotation process has associated key decision points. These decision points, many of which are undocumented in published annotation studies and even treated

as “defaults” by annotation teams, have the potential to affect the quality of annotation outcomes. We summarize key decision points in Table 2 and discuss them in detail below.

Task Design. Before annotation can begin, data must be collected and pre-processed so that it is well-formatted for annotation. Key decision points in this stage include determining whether data is representative of the underlying concept that annotation seeks to capture and determining how data should be preprocessed. Interview participants who used “gold labels” reported that it was particularly important to determine whether those labels were representative of their underlying concept, as non-representative gold labels will systematically fail at catching annotation issues (P2, P5). Interview participants also reported struggling to determine the best way to preprocess data without introducing “*compounding*” (P1) or “*cascading*” (P10) issues (P1–P3, P10). For example, P2 described an annotation task in which crowdworkers evaluated the clarity of website terms of service. To set up the task, he scraped terms of service from sites and had to decide whether to strip HTML elements from the scraped pages (which risked dropping relevant information like bolding, font size, and font color) or to leave them in (which risked reducing readability of the text). Although prior work has similarly found that annotation quality is affected by data collection and preprocessing procedures, such decisions are rarely discussed in published annotation studies (Grondin-Verdon, Caillat, and Ouni 2024; Hays et al. 2023).

A decision that was particularly challenging for many participants was determining the unit of annotation (P1–P3, P6, P8, P10). For example, if text data is annotated, should it be annotated sentence-by-sentence or paragraph-by-paragraph? P6, who was annotating trial transcripts for aspects of gender stereotyping (e.g., hypersexualizing women), described the tradeoffs inherent to this decision:

“When you have a trial transcript...you have a lot of interruptions...So this is a little bit more difficult to code by sentence...it was very difficult for researchers to [know whether they should] code [an exchange] three different times, or [just] once. And we in the end decided that it would increase inter-rater reliability if we expanded the unit of text...That creates a problem, because if, quantitatively, you now want to say, ‘how many times was there a reference to her having sex, or wanting to have sex?’ we’ve only coded it once, instead of coding it three times. And what linguists will tell you is that repetition matters. You know, if somebody hears ‘she wanted to have sex’ five times, that’s gonna have more of an impact than ‘she wanted to have sex’ once...In the end, I think this is a process of what are you prioritizing?”

Other decision points identified during this phase were determining the modality of annotation (e.g., multiple choice vs. free response) (P4); determining whether and how to collect measures of annotator certainty (P1, P5, P6); navigating the tradeoff between writing annotation guidelines that were comprehensive without being overly long (P9) (see Dimara, Bezerianos, and Dragicevic 2017; Gadiraju, Yang, and

Stage	Components	Key Decision Points
Task Design	Data	Is data representative of the underlying concept? How should data be preprocessed?
	Schema	What is the modality of annotation? What is the unit of annotation? Should measures of certainty be collected? How?
	Guidelines	How to handle the comprehensiveness/readability tradeoff?
	Process	Through what interface should annotations be collected? What is the optimal level of coordination among annotators?
Annotator Management & Annotation Collection	Selection	What should the annotator population be? Should screening or qualification testing be applied?
	Training & Feedback	How should annotators be trained? Should annotator feedback be provided/solicited?
	Incentives	What is the optimal payment structure?
Quality Assessment	Metric Selection	Should agreement be used? If so, which agreement metric should be used? Should gold labels be created? If so, how?
Quality Improvement & Adjudication	Improvement	Should new annotations be collected or should existing annotations be filtered? If new annotations are collected, what changes should be made to the process? If existing annotations should be filtered, how?
	Adjudication	Should all annotations be retained, or should annotations be aggregated? If annotations are to be aggregated, how? How should certainty be expressed?

Table 2: Key decision points affecting the quality of annotations at each stage of the annotation process, identified through our literature review and interview study.

Bozzon 2017; Pradhan, Schaekermann, and Lease 2022); designing an optimal interface for collecting annotations, especially when the data to be annotated was multimodal (P7–P9); and determining the optimal level of coordination among annotators (e.g., should annotators come to a consensus or independently annotate instances?) (P1, P4, P7). Multiple participants noted that all of these decision points are complicated by the rise of AI annotators, as the extent to which best practices for human annotation can be applied to AI is not yet clear (P1, P2, P5, P7, P10) (see He et al. 2024).

Annotator Management & Annotation Collection. By far the biggest decision point in the annotator management phase, reported by all participants, was deciding how to select the annotator population: should annotators be human experts, human crowdworkers, AI, or some combination thereof? Participants reported having to determine whether annotation required expertise or could be done by crowdworkers (P3), and whether annotation required human perspectives or could be done by AI (P6, P8, P10). Participants also grappled with the fact that, although AI annotators tended to make fewer errors than humans, the errors they did make were often less explainable (P2, P5). They also struggled with resource constraints, such as cost and time, that prevented them from regularly using expert human annotators (P1–P3, P6, P7, P9, P10), who were seen as a gold standard.

There is a large body of research on managing and incentivizing crowdworkers which deals with questions of how to select an annotator population, how to decide whether

annotators should be screened (e.g., based on their prior tasks completed on a crowdwork platform) or tested (e.g., by completing an assessment prior to the start of the task), how to communicate with annotators, and how to determine optimal payment structures to incentivize high-quality annotation while still following fair labor practices (Alabduljabbar and Al-Dossari 2016; Alonso 2015b; Calderon, Reichart, and Dror 2025; Callison-Burch and Dredze 2010; Daniel et al. 2019; Eickhoff 2018; Hansen et al. 2013; Ho et al. 2015; Kovashka et al. 2016; Lease 2011; Li et al. 2017, 2019; Sabou et al. 2014; Snow et al. 2008; Vaughan 2018). Despite this, participants found themselves struggling with many of these questions, as well as with newer problems like preventing unwanted LLM use by crowdworkers (P4, P9).

Quality Assessment. Interview participants generally assessed the quality of annotations in three ways: manually reviewing a subset of annotations to see whether they passed the “sniff test” (P1–P5, P8–P10), calculating inter-rater reliability (P1–P3, P6, P7, P9, P10), and calculating agreement between annotations and gold labels (P1, P2, P7, P10). Deciding which approach to use (agreement, gold labels, or other) is a universal challenge. Although most annotation managers calculated and tried to reach high IRR, they also acknowledged that high agreement does not necessarily imply high correctness (P1, P7, P8, P10). Nevertheless, they felt that reaching high IRR was often a necessary task to have their annotations seen as valid by third parties. However, as has been extensively discussed in prior work, it is usually not clear which of many possible IRR metrics

to rely on (all have been criticized for resting on unrealistic assumptions), nor is it clear what a threshold for “good” IRR even is (Banerjee et al. 1999; Bayerl and Paul 2011; Braun 2024; Checco et al. 2017; McDonald, Schoenebeck, and Forte 2019; Reidsma and Carletta 2008; Zhao, Liu, and Deng 2013).⁷ Consistent with prior work, participants often prioritized agreement even when they felt that it was not the best indicator of annotation quality, a phenomenon that was summarized by P10 as “*a gap between what is publishable and what a peer reviewer will say [is] good enough, and what we actually all should be doing or considering as a research community.*”

Quality Improvement & Adjudication. In the last stage of the annotation process, teams must decide whether their annotations are sufficiently high-quality. If not, they must decide how to improve annotations: either by collecting new annotations or by removing lower-quality ones. If teams decide to remove lower-quality annotations, several key questions arise about how to do that filtering: what metric and threshold should be used to indicate low quality? Should filtering be done at the data instance level (so that every annotation for a given data instance is removed) or at the annotator level (so that every annotation produced by a particular annotator is removed)? If teams decide that their data is sufficiently high-quality, they must conduct adjudication, which is the process by which a final annotation (or set of annotations) for a data instance is determined. To perform adjudication, annotation teams must first decide whether to retain all labels for an instance or aggregate them into a single label. During this stage, annotation teams may also consider whether and how to express certainty in their final published dataset. Participants reported that their biggest challenge in this stage was deciding whether or not to aggregate labels: multiple participants reported believing that data instances in their setting were genuinely ambiguous, but nevertheless needing a single gold label for their downstream tasks (P2, P5, P9).

Prior work suggests many approaches for improving the quality of data annotations, including identifying “hard” data instances (Beigman Klebanov and Beigman 2009); identifying “low-quality” annotators (Dawid and Skene 1979; Gordon et al. 2021; Hovy et al. 2013); and returning to earlier stages of the annotation process to improve aspects of task design or annotator management (Alonso 2015b; Hube, Fetahu, and Gadiraju 2019; Pradhan, Schaekermann, and Lease 2022; Röttger et al. 2022; Sabou et al. 2014; Vaughan 2018). Interestingly, some prior work has also proposed that annotation errors—and disagreement generally—can be leveraged in ML processes for learning models that are robust to noise, and that may outperform models trained on “cleaner” data (Ali et al. 2026; Basile et al. 2021; Peterson et al. 2019; Plank 2022; Reidsma and Op Den Akker 2008; Schmarje et al. 2022; Sheng, Provost, and Ipeirotis 2008; Thebault-Spieker et al. 2023; Uma et al. 2021).

⁷Zhao, Liu, and Deng (2013) identify 22 published IRR metrics (many of which are mathematically equivalent). Popular metrics include percent agreement, Cohen’s κ (Cohen 1960), and Krippendorff’s α (Hayes and Krippendorff 2007).

4.3 Distinct Sources of Annotation Issues

Annotation issues can often be identified by finding misalignments between an annotation and a “ground truth.” But the action needed to correct an annotation issue depends on the source of that issue. Based on our literature review and interview study, we propose a taxonomy of five distinct sources of annotation issues (Table 3). While many of these sources have been identified and discussed by prior work, here, we offer a set of definitions that more clearly distinguish unique sources of issues from one another.

Errors occur when annotators do not follow given instructions (P1, P2, P4–P9). A large body of prior work is devoted to identifying errors that occur when crowdworkers try to “game” annotation tasks (Alonso 2015a; Marshall et al. 2023). In our interviews, participants were more likely to think of errors as good-faith mistakes stemming from fatigue (P1, P2, P6, P7) or over-reliance on heuristics-based annotation strategies (e.g., annotating for sentiment based on keywords as opposed to careful reading of text) (P2, P4, P7). *Ambiguity* occurs when annotation instructions do not provide enough information to determine what an annotation should be (P2, P3, P5–P8, P10). Ambiguity has been identified by prior work as a source of low agreement in annotation (Braun 2024; Geva, Goldberg, and Berant 2019; Schwirten et al. 2024). Unlike prior work, we distinguish ambiguity from *impossibility*, which occurs when data instances themselves do not provide enough information to determine what an annotation should be (P5, P7). *Subjectivity* occurs when an annotation depends on the personal values, feelings, or beliefs of an annotator (P1–P3, P5–P10). Many researchers have identified subjectivity, which also contributes to low agreement, as a major challenge in data annotation (Arhin et al. 2021; Ferracane et al. 2021; Hettiachchi et al. 2023; Kapania, Taylor, and Wang 2023; Pang et al. 2023; Parmar et al. 2023; Sap et al. 2022; Schumann et al. 2023; Wang et al. 2025; Wen-Yi et al. 2024). Subjectivity affects even tasks that may sound objective, such as identifying the skin tone of an individual captured in an image (Barrett, Chen, and Zhang 2023). Finally, we introduce *identity* as a source of annotation issues when the quality of an annotation depends not on the content of the annotation, but on the identity of the annotator (P4, P9). For example, AI-generated annotations are misaligned with a desired “ground truth” when they are created for an evaluation task intended to elicit human preferences.

The cause of the issue affects the solutions that can or should be applied to address it. It is important to have a clear language for understanding issues and their causes—without it, annotation managers may try to correct issues using inappropriate methods. For example, Zhang et al. (2025) find that annotation managers often treat all low-agreement annotations as errors, when in fact, large-scale low agreement is likely to indicate that annotations may be ambiguous or impossible. Nevertheless, the annotation managers studied by Zhang et al. (2025) responded to these cases by rejecting (and thus declining to pay) large swaths of annotators. In addition to constituting an objectionable labor practice, this action will not solve the problem if annotation tasks are ambiguous or impossible.

Source	Definition	Potential Solutions
Error	The annotator did not follow instructions	Retrain annotators; filter annotators
Ambiguity	There is not enough information in the annotation instructions to fully determine the annotation	Improve annotation instructions; refine annotation schema
Impossibility	There is not enough information in the data to fully determine the annotation	Re-process data; filter data
Subjectivity	The annotation depends on implicit values, beliefs, opinions, or assumptions	Improve annotation instructions; refine annotation schema
Identity	The identity of the annotator is not aligned with the desired annotator population	Filter annotators

Table 3: Sources of annotation issues. The source of an issue determines potential solutions to the issue.

5 Improving Annotation Reliability and Validity with Measurement Theory

Despite the fact that annotation issues are often thought of as misalignment between annotations and “ground truths,” annotation teams consistently focus on agreement (between multiple annotators on the same data instance) as their primary quality metric. This is a fundamental mismatch that is often justified by the fact that agreement is generally easy to calculate and “ground truth” is inevitably not available for all data instances (if it were, annotation would not be necessary). Measurement theory offers a broader vocabulary for evaluating annotation quality in terms not only of whether annotators agree, but of whether annotations are reliable and valid measures of the concepts they are intended to capture. In this section, we draw on measurement theory to identify relevant aspects of reliability and validity, and use findings from our literature review and interview study to show how they can be incorporated into annotation processes.

Test-Retest Reliability. While inter-rater reliability is the dominant approach to measuring annotation quality, it is rare to assess test-retest reliability (also referred to as intra-rater reliability), which asks whether annotations are consistent when produced by the same annotator at different points in time. Abercrombie et al. (2025) recently found that, of more than 80,000 papers in the ACL Anthology, just 56 of them (0.07%) report measuring test-retest reliability. However, researchers hypothesize that test-retest reliability can be used to identify sources of disagreement in annotation (e.g., distinguishing disagreement caused by subjectivity from disagreement caused by random error) (Abercrombie et al. 2025; Gordon et al. 2021). In subjective annotation tasks, measuring test-retest reliability may thus be a better way of identifying high-quality annotations than measuring IRR (Arhin et al. 2021). Test-retest reliability also has its challenges, however: due to its relative unpopularity, there are currently no agreed-upon best practices for how to best calculate it (for example, how long should researchers wait between original and subsequent annotations?) (Abercrombie et al. 2025).

Consistent with prior work, very few interview participants reported measuring test-retest reliability. Some seemed unaware that it existed, like P1, who asked: “*Like,*

check over your own work? I don’t know that that’s ever been done.” However, two participants reported that test-retest reliability had unique benefits not captured in other forms of reliability (P7, P10). P7 described it as “*low-hanging fruit for quality [assessment]...of the annotator.*” P10 argued that assessing test-retest reliability is particularly important for subjective tasks, noting that “*as you code data...your affinity with the data and your understanding of the construct will change.*”

Face Validity. Face validity assessments, in which researchers check whether annotated data instances “look right” or pass the “sniff test” are common—nearly every interview participant reported engaging in face validity assessments by conducting manual review of a subset of annotated data instances (P1–P5, P8–P10). Prior work has found that this practice is widespread (Harvey et al. 2025). However, these face validity assessments are often informal and thus neither documented nor published. In a rare exception describing a small-scale face validity assessment, Harvey, Kizilcec, and Koencke (2025) described identifying issues in a crowdsourced corpus of tweets written in both African American English and ‘Standard’ American English (Groenwold et al. 2020), and choosing not to use the dataset as a result. Similarly, Grondin-Verdon, Caillat, and Ouni (2024) conducted a face validity assessment of a gesture recognition corpus by re-annotating parts of the corpus and comparing their annotations to crowdsourced ones. For larger-scale face validity assessments, researchers may turn to automated methods (e.g., confident learning (Northcutt, Athalye, and Mueller 2021)) that are intended to identify likely annotation errors that can then be manually examined (Klie, Webber, and Gurevych 2023).

Content Validity. Content validity asks whether annotations “wholly and fully capture the substantive nature” of the concept that they seek to measure (Jacobs and Wallach 2021). In the context of annotation, content validity is most notably considered in social science research. Social science research, which often conceives of annotation as a qualitative coding task, discusses the importance of establishing clear definitions and coding rules for concepts of interest, and of grounding these definitions in specific theories (MacQueen et al. 1998). In contrast, critical

scholars have repeatedly pointed out that crowdsourced annotation work and annotation tasks intended for AI/ML systems often do not include clear systematizations or operationalizations of concepts of interest (Blodgett et al. 2021; Scheuerman et al. 2020).

Interview participants, especially those with backgrounds in social sciences, described attempting to establish content validity by using established theories to develop their codebooks or annotation guidelines (P1, P2, P10). P10 described how they interrogated the content validity of their annotation guidelines:

“We actually had raters with an understanding of the... theories. We gave them the codes [i.e., labels] and had them rate them for how clear are they, how concise are they, how aligned with the theory are they, how practical are they? And so we have these raw metrics of code quality.”

Convergent Validity. Convergent validity asks whether annotations correlate with other, valid, annotations of related concepts. Because content validity is rarely considered, assessing convergent validity in annotated data is nearly impossible: it is extremely hard to combine and compare different datasets. For example, in a review of language corpora annotated with information on negation, Jiménez-Zafra et al. (2020) found that one-third of datasets lacked annotation guidelines entirely. The others had incomplete or inconsistent guidelines, making it impossible to combine and compare for the purposes of a convergent validity assessment. Bostan and Klinger (2018) had similar findings in a review of corpora annotated for emotion classification.

Similarly, interview participants reported that they struggled to use existing annotated data and annotation guidelines, whether or not they had been involved in the creation of the data or guidelines in the first place (P1–P4, P6, P10). A primary reason for this was that participants felt that the concepts being annotated were often not consistently systematized or operationalized. Like prior researchers, they struggled with how to compare or combine different systematizations or operationalizations of related concepts.

Assessing convergent validity thus requires establishing content validity of annotation processes as a first step. Well-documented annotation processes can also be helpful here: for example, Recht et al. (2019) were able to conduct a convergent validity assessment of ImageNet (Deng et al. 2009) by replicating the original process used to select and label images and then testing the ability of models trained on the original ImageNet data to label the new images.

Predictive Validity. Predictive validity asks whether annotations can predict relevant related concepts. It is inadvertently assessed almost by definition for many data annotation projects conducted in computational fields: the explicit goal of annotation projects is often to build a model that performs a downstream prediction task. Thus, a view expressed by some interview participants is that annotations can be thought of as valid if they can be used to build an accurate predictive model (P1, P3). However, it is often challenging to specify what exactly the prediction task should be. For example, Catanzariti (2025) conducted an

interview study of individuals annotating videos of faces for affect; although the annotations were intended to measure fine-grained emotional states at specific points in time, practitioners assessed quality based on whether models built using the annotations could predict broader outcomes (“is this person generally sad?”).

Discriminant Validity. Discriminant validity asks whether annotations are too correlated with concepts not related to those that they seek to capture. While interview participants did not report engaging in discriminant validity assessments, prior work suggests opportunities for incorporating discriminant validity assessments into the annotation process. For example, researchers have trained models on annotated data that they then use to predict information that should not be contained in annotations—if the models perform well, this is a sign that annotations have low discriminant validity. For example, Geva, Goldberg, and Berant (2019) find that, in widely-used natural language understanding (NLU) benchmarks, it is possible to predict the IDs of annotators who produced particular examples. This suggests that rather than simply measuring NLU, the annotations are also inadvertently capturing information about individual annotators.

Taken together, we recommend that researchers and practitioners working on annotation projects should:

- (a) Determine the source of an annotation error before taking action to resolve it (see Table 3), e.g., by measuring test-retest reliability in addition to IRR (to distinguish ambiguity from subjectivity) or by adding an option for annotators to indicate that an instance cannot be annotated (to distinguish ambiguity from impossibility);
- (b) Clearly systematize and operationalize concepts of interest to establish content validity and facilitate interrogations of convergent and discriminant validity;
- (c) Document and report the results of face and predictive validity assessments that are already informally conducted (clearly specifying or pre-registering the predictive task for predictive validity assessments); and
- (d) Build infrastructure to support interrogation (see §6).

6 Discussion

In this work, we presented the results of a qualitative study that is intended to facilitate the correction of issues in data annotation. Here, we outline several avenues for future work that our findings suggest will be particularly fruitful.

Learning from Qualitative Coding. Multiple participants reported either that they struggled to identify best practices for their annotation tasks (P1, P7), or that they frequently observed other annotation projects that did not follow what they considered to be well-documented best practices (P9, P10). This is largely due to the fact that data annotation is a task that exists across multiple disciplines which are often not in conversation with each other. For example, P1, whose background is in machine learning, felt that “as a team we’re learning best practices that already exist... more from the qualitative research community.”

Similarly, P7, whose background is in data science, (jokingly) described the feeling of realizing that their annotation challenge could be addressed by adopting best practices from another field: “*political science, apparently, has been doing [pairwise annotation] for a long time and hiding it from everybody else.*” When computer science and machine learning researchers think of data annotation, they often focus on bringing in best practices from crowdwork and even from prediction tasks. We suggest that best practices from the social sciences could be a valuable source of ideas for improving the reliability and validity of data annotation.

Furthermore, critical scholarship from the social sciences is the primary avenue through which questions of *consequential validity* of data annotations (i.e., what are the societal impacts of annotations?) are examined. Catanzariti (2025), Denton et al. (2021), Lin (2022), Scheuerman et al. (2020), and Scheuerman, Hanna, and Denton (2021) argue that annotations “construct” ground truth, making subjective or value-laden judgments appear objective and immutable. Thus, annotations can reify social hierarchies and reinforce power imbalances (Smart et al. 2024). Scholars have also noted that power imbalances can be reinforced within the annotation supply chain (DiSalvo et al. 2024; Miceli, Schuessler, and Yang 2020). As crowdsourcing became an increasingly popular annotation paradigm in the early 2010s, Irani and Silberman (2013) raised the alarm that “human computation currently relies on worker invisibility.” Despite the efforts of researchers like Irani and Silberman (2013), who introduced the Turkopticon system to allow crowdworkers to share experiences, organize, and engage in mutual aid, and Miceli et al. (2025), who proposed Workers’ Inquiry as a Research Methodology to center crowdworkers in AI research, annotation labor remains largely invisible today (Gray and Suri 2019). Annotation tasks are often outsourced to low-wage workers located in lower-income countries who are subject to myriad risks, including psychological harm from labeling toxic data as well as privacy risks that are inherent to production tasks (Hao 2025; Le Ludec, Cornet, and Casilli 2023; Shestakofsky 2024; Shmueli et al. 2021; Smart et al. 2024). Annotation managers often view annotators as adversarial parties who may try to “game” the system by being paid to provide low-effort, low-quality data (Rothschild et al. 2024). As a result, annotation managers employ various surveillance and quality control measures to remove human uncertainty and subjectivity from annotations, often rejecting annotations and withholding pay from annotators who do not meet certain standards (Meisner, Duffy, and Ziewitz 2024; Rothschild et al. 2024; Zhang et al. 2025). To improve consequential validity in data annotation it is crucial to rely on critical scholarship that centers workers by emphasizing how leveraging worker errors, worker disagreement, and communication with workers can help improve both labor conditions and data quality (Lin and Jackson 2023; Schaekermann et al. 2018; Simons, Gurari, and Fleischmann 2020).

Accounting for AI. Current practices are being adjusted to account for the growing role of AI in data annotation pipelines (Desai, Card, and Jacobs 2026; Tan et al. 2024).

Researchers have proposed a variety of approaches for combining human judgment with AI precision and consistency in order to improve annotation processes (Calderon, Reichart, and Dror 2025; He et al. 2024; Schroeder, Roy, and Kabbara 2025; Tavakoli and Zamani 2025). An overarching theme is that AI annotators have the potential to prevent common human errors in annotation that are caused by fatigue or inattention (Wen-Yi et al. 2024). At the same time, however, researchers are raising alarms about the potential for AI annotators to replace human perspectives in annotation and potentially render crowdsourced data unusable (Wang, Morgenstern, and Dickerson 2025; Westwood 2025). Research shows that LLMs can produce dramatically different annotations depending on minor prompt variation (Röttger et al. 2024), raising concerns of “LLM-hacking” (Baumann et al. 2025). Co-annotation between humans and LLMs can also lead to humans adopting LLM annotations, even if humans on their own would annotate differently (He, Naphade, and Huang 2025). Exploring how annotation teams may use AI to reduce annotation issues—without introducing new ones—is thus an important avenue for future work (e.g., Kumar et al. 2026).

Building Infrastructure to Support Interrogation. Finally, future research should explore practical avenues for interrogating the reliability and validity of annotations. In particular, almost all interview participants (P1–P5, P8–P10) reported conducting manual review of annotations to see if they passed the “sniff test”—i.e., face validity assessments. However, these assessments were informal: participants did not document or publish their results. Building infrastructure to formalize face validity assessments (e.g., by creating repositories for data users to share and aggregate the results of their manual reviews or mechanisms for the research community to flag low-quality annotations in public datasets) could thus enable the improvement of annotated data at scale, in line with prior work that has shown that datasets can be improved by public efforts to correct errors (Beyer et al. 2020; Denton et al. 2021; Northcutt, Athalye, and Mueller 2021; Peychev et al. 2023).

7 Conclusion

Through a literature review (N=132) and interview study (N=10), we contribute a framework to support the correction of annotation issues. First, we map key decision points throughout the annotation process that can lead to annotation issues. By making these decision points, which are often taken for granted, visible, we show where annotation processes can be changed in order to improve annotations. Second, we argue that researchers and practitioners should treat annotation as a measurement problem. In so doing, we build on prior work that has called for measurement theory to be applied to sociotechnical research (Jacobs and Wallach 2021; Wallach et al. 2025) and provide a framework to ground prior work that has called for researchers to look beyond agreement when evaluating data annotation (Thomas et al. 2026). We make concrete suggestions for how annotation processes can move beyond agreement to interrogate and improve the reliability and validity of annotations.

Acknowledgments

We thank our anonymous interview participants for their time and candor. We thank Nikhil Garg; Karen Levy; Angelina Wang; and the members of the Future of Learning Lab, the KLEAR Lab, the AI in Policy and Practice group, and the Digital Life Initiative for their thoughtful and constructive feedback. This material is based upon work supported by grants from Apple, Inc. and by the National Science Foundation under Grant Number 2416996. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and should not be interpreted as reflecting the views, policies or position, either expressed or implied, of Apple Inc. or the National Science Foundation.

References

- Abercrombie, G.; Dinkar, T.; Cercas Curry, A.; Rieser, V.; and Hovy, D. 2025. Consistency is Key: Disentangling Label Variation in Natural Language Processing with Intra-Annotator Agreement. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, 63–74. Suzhou, China: Association for Computational Linguistics.
- Adcock, R.; and Collier, D. 2001. Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review*, 95(3): 529–546.
- Alabduljabbar, R.; and Al-Dossari, H. 2016. A Task Ontology-based Model for Quality Control in Crowdsourcing Systems. In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*, 22–28. Odense Denmark: ACM. ISBN 978-1-4503-4455-5.
- Ali, D.; Zhao, D.; Koencke, A.; and Papakyriakopoulos, O. 2026. Operationalizing Pluralistic Values in Large Language Model Alignment Reveals Trade-offs in Safety, Inclusivity, and Model Behavior. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(44): 37222–37231.
- Alonso, O. 2015a. Challenges with Label Quality for Supervised Learning. *Journal of Data and Information Quality*, 6(1): 1–3.
- Alonso, O. 2015b. Practical Lessons for Gathering Quality Labels at Scale. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, 1089–1092. Santiago Chile: ACM. ISBN 978-1-4503-3621-5.
- Arhin, K.; Baldini, I.; Wei, D.; Ramamurthy, K. N.; and Singh, M. 2021. Ground-Truth, Whose Truth? – Examining the Challenges with Annotating Toxic Text Datasets. ArXiv:2112.03529 [cs].
- Banerjee, M.; Capozzoli, M.; McSweeney, L.; and Sinha, D. 1999. Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, 27(1): 3–23.
- Barrett, T.; Chen, Q.; and Zhang, A. 2023. Skin Deep: Investigating Subjectivity in Skin Tone Annotations for Computer Vision Benchmark Datasets. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT, 1757–1771. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0192-4.
- Basile, V.; Fell, M.; Fornaciari, T.; Hovy, D.; Paun, S.; Plank, B.; Poesio, M.; and Uma, A. 2021. We Need to Consider Disagreement in Evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, 15–21. Online: Association for Computational Linguistics.
- Baumann, J.; Röttger, P.; Urman, A.; Wendsjö, A.; Plaza-del Arco, F. M.; Gruber, J. B.; and Hovy, D. 2025. Large Language Model Hacking: Quantifying the Hidden Risks of Using LLMs for Text Annotation. ArXiv:2509.08825 [cs].
- Bayerl, P. S.; and Paul, K. I. 2011. What Determines Inter-Coder Agreement in Manual Annotations? A Meta-Analytic Investigation. *Computational Linguistics*, 37(4): 699–725.
- Beck, J. 2023. Quality aspects of annotated data: A research synthesis. *ASta Wirtschafts- und Sozialstatistisches Archiv*, 17(3-4): 331–353.
- Beigman Klebanov, B.; and Beigman, E. 2009. From Annotator Agreement to Noise Models. *Computational Linguistics*, 35(4): 495–503.
- Bender, E. M.; and Friedman, B. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6: 587–604.
- Beyer, L.; Hénaff, O. J.; Kolesnikov, A.; Zhai, X.; and van den Oord, A. 2020. Are we done with ImageNet? arXiv:2006.07159.
- Blodgett, S. L.; Lopez, G.; Olteanu, A.; Sim, R.; and Wallach, H. 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, 1004–1015. Online: Association for Computational Linguistics.
- Bostan, L. A. M.; and Klinger, R. 2018. An Analysis of Annotated Corpora for Emotion Classification in Text. In *Proceedings of the 27th International Conference on Computational Linguistics*, ICCL. Santa Fe, NM, USA: ACL.
- Braun, D. 2024. I beg to differ: how disagreement is handled in the annotation of legal machine learning data sets. *Artificial Intelligence and Law*, 32(3): 839–862.
- Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2): 77–101.
- Braun, V.; and Clarke, V. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4): 589–597.
- Caine, K. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, 981–992. New York, NY, USA: Association for Computing Machinery. ISBN 9781450333627.
- Calderon, N.; Reichart, R.; and Dror, R. 2025. The Alternative Annotator Test for LLM-as-a-Judge: How to Statistically Justify Replacing Human Annotators with LLMs. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T.,

- eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16051–16081. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.
- Callison-Burch, C.; and Dredze, M. 2010. Creating Speech and Language Data With Amazon’s Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*. Los Angeles: Association for Computational Linguistics.
- Catanzariti, B. 2025. Taming Affect: On the Construction of Objectivity in Data Annotation Practices. *Digital Society*, 4(2): 40.
- Checco, A.; Roitero, K.; Maddalena, E.; Mizzaro, S.; and Demartini, G. 2017. Let’s Agree to Disagree: Fixing Agreement Measures for Crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 5: 11–20.
- Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1): 37–46.
- Cook, D. A.; and Beckman, T. J. 2006. Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *The American Journal of Medicine*, 119(2): 166.e7–166.e16.
- Cronbach, L. J.; and Meehl, P. E. 1955. Construct validity in psychological tests. *Psychological Bulletin*, 52(4): 281–302.
- Daniel, F.; Kucherbaev, P.; Cappiello, C.; Benatallah, B.; and Allahbakhsh, M. 2019. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Computing Surveys*, 51(1): 1–40.
- Dawid, A. P.; and Skene, A. M. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Applied Statistics*, 28(1): 20.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.
- Denton, E.; Hanna, A.; Amironesei, R.; Smart, A.; and Nicole, H. 2021. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, 8(2): 205395172111035955.
- Desai, M.; Card, D.; and Jacobs, A. Z. 2026. Validating LLMs in social science: Epistemic threats and emerging norms. arXiv:2607.07915.
- Dimara, E.; Bezerianos, A.; and Dragicevic, P. 2017. Narratives in Crowdsourced Evaluation of Visualizations: A Double-Edged Sword? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI, 5475–5484. Denver Colorado USA: ACM. ISBN 978-1-4503-4655-9.
- DiSalvo, C.; Rothschild, A.; Schenck, L. L.; Shapiro, B. R.; and DiSalvo, B. 2024. When Workers Want to Say No: A View into Critical Consciousness and Workplace Democracy in Data Work. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–24.
- Díaz, M.; Kivlichan, I.; Rosen, R.; Baker, D.; Amironesei, R.; Prabhakaran, V.; and Denton, R. 2022. CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. In *2022 ACM Conference on Fairness Accountability and Transparency*, FAccT, 2342–2351. Seoul Republic of Korea: ACM. ISBN 978-1-4503-9352-2.
- Eickhoff, C. 2018. Cognitive Biases in Crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM, 162–170. Marina Del Rey CA USA: ACM. ISBN 978-1-4503-5581-0.
- Eriksson, M.; Purificato, E.; Noroozian, A.; Vinagre, J.; Chaslot, G.; Gomez, E.; and Fernandez-Llorca, D. 2025. Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1): 850–864.
- Fabris, A.; Messina, S.; Silvello, G.; and Susto, G. A. 2022. Tackling Documentation Debt: A Survey on Algorithmic Fairness Datasets. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO, 1–13. Arlington VA USA: ACM. ISBN 978-1-4503-9477-2.
- Fereday, J.; and Muir-Cochrane, E. 2006. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods*, 5(1): 80–92.
- Ferracane, E.; Durrett, G.; Li, J. J.; and Erk, K. 2021. Did they answer? Subjective acts and intents in conversational discourse. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL, 1626–1644. Online: Association for Computational Linguistics.
- Gadiraju, U.; Yang, J.; and Bozzon, A. 2017. Clarity is a Worthwhile Quality: On the Role of Task Clarity in Micro-task Crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, 5–14. Prague Czech Republic: ACM. ISBN 978-1-4503-4708-2.
- Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.
- Gehrmann, S.; Clark, E.; and Sellam, T. 2023. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *Journal of Artificial Intelligence Research*, 77: 103–166.
- Geva, M.; Goldberg, Y.; and Berant, J. 2019. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, EMNLP-IJCNLP, 1161–1166. Hong Kong, China: Association for Computational Linguistics.
- Gordon, M. L.; Zhou, K.; Patel, K.; Hashimoto, T.; and Bernstein, M. S. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference*

- on *Human Factors in Computing Systems*, CHI, 1–14. Yokohama Japan: ACM. ISBN 978-1-4503-8096-6.
- Gosciak, J.; Boyce, L.; Wang, A.; and Koenecke, A. 2026. *Scrutinizing Index-Based Risk Assessments: A Case Study in NYC Decision-making for Heat Emergency Management*, 6659–6701. New York, NY, USA: Association for Computing Machinery. ISBN 9798400725968.
- Gray, M.; and Suri, S. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt. ISBN 9781328566249.
- Groenwold, S.; Ou, L.; Parekh, A.; Honnavalli, S.; Levy, S.; Mirza, D.; and Wang, W. Y. 2020. Investigating African-American Vernacular English in Transformer-Based Text Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, EMNLP, 5877–5883. Online: Association for Computational Linguistics.
- Grondin-Verdon, M.; Caillat, D.; and Ouni, S. 2024. Qualitative study of gesture annotation corpus : Challenges and perspectives. In *Companion Proceedings of the 26th International Conference on Multimodal Interaction*, 147–155. San Jose Costa Rica: ACM. ISBN 979-8-4007-0463-5.
- Hansen, D. L.; Schone, P. J.; Corey, D.; Reid, M.; and Gehring, J. 2013. Quality control mechanisms for crowdsourcing: peer review, arbitration, & expertise at family-search indexing. In *Proceedings of the 2013 conference on Computer supported cooperative work, CSCW*, 649–660. San Antonio Texas USA: ACM. ISBN 978-1-4503-1331-5.
- Hao, K. 2025. *Empire of AI: Dreams and Nightmares in Sam Altman’s OpenAI*. Penguin Publishing Group. ISBN 9780593657508.
- Harvey, E.; Kizilcec, R. F.; and Koenecke, A. 2025. 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*, 2025–2039. Athens Greece: ACM. ISBN 979-8-4007-1482-5.
- Harvey, E.; Sheng, E.; Blodgett, S. L.; Chouldechova, A.; Garcia-Gathright, J.; Olteanu, A.; and Wallach, H. 2025. Understanding and Meeting Practitioner Needs When Measuring Representational Harms Caused by LLM-Based Systems. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 18423–18440. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.
- Hayes, A. F.; and Krippendorff, K. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1(1): 77–89.
- Hays, C.; Schutzman, Z.; Raghavan, M.; Walk, E.; and Zimmer, P. 2023. Simplistic Collection and Labeling Practices Limit the Utility of Benchmark Datasets for Twitter Bot Detection. In *Proceedings of the ACM Web Conference 2023*, 3660–3669. Austin TX USA: ACM. ISBN 978-1-4503-9416-1.
- He, Z.; Huang, C.-Y.; Ding, C.-K. C.; Rohatgi, S.; and Huang, T.-H. K. 2024. If in a Crowdsourced Data Annotation Pipeline, a GPT-4. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI, 1–25. Honolulu HI USA: ACM. ISBN 979-8-4007-0330-0.
- He, Z.; Naphade, S.; and Huang, T.-H. K. 2025. Prompting in the Dark: Assessing Human Performance in Prompt Engineering for Data Labeling When Gold Labels Are Absent. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI, 1–33. Yokohama Japan: ACM. ISBN 979-8-4007-1394-1.
- Hettiachchi, D.; Holcombe-James, I.; Livingstone, S.; De Silva, A.; Lease, M.; Salim, F. D.; and Sanderson, M. 2023. How Crowd Worker Factors Influence Subjective Annotations: A Study of Tagging Misogynistic Hate Speech in Tweets. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 11(1): 38–50.
- Ho, C.-J.; Slivkins, A.; Suri, S.; and Vaughan, J. W. 2015. Incentivizing High Quality Crowdsourcing. In *Proceedings of the 24th International Conference on World Wide Web, WWW*, 419–429. Florence Italy: International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-3469-3.
- Holland, S.; Hosny, A.; Newman, S.; Joseph, J.; and Chmielinski, K. 2020. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. In Hallinan, D.; Leenes, R.; Gutwirth, S.; and De Hert, P., eds., *Data Protection and Privacy: Data Protection and Democracy*. Hart Publishing. ISBN 978-1-5099-3274-0 978-1-5099-3276-4 978-1-5099-3275-7 978-1-5099-3277-1.
- Hovy, D.; Berg-Kirkpatrick, T.; Vaswani, A.; and Hovy, E. 2013. Learning Whom to Trust with MACE. In Vanderwende, L.; Daumé III, H.; and Kirchhoff, K., eds., *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL*, 1120–1130. Atlanta, Georgia: Association for Computational Linguistics.
- Hube, C.; Fetahu, B.; and Gadiraju, U. 2019. Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI, 1–12. Glasgow Scotland Uk: ACM. ISBN 978-1-4503-5970-2.
- Hutchinson, B.; Smart, A.; Hanna, A.; Denton, R.; Greer, C.; Kjartansson, O.; Barnes, P.; and Mitchell, M. 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*, 560–575. Virtual Event Canada: ACM. ISBN 978-1-4503-8309-7.
- Irani, L. C.; and Silberman, M. S. 2013. Turkopticon: interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI, 611–620. Paris France: ACM. ISBN 978-1-4503-1899-0.
- Jackman, S. 2008. Measurement. In *The Oxford Handbook*

- of *Political Methodology*. Oxford University Press. ISBN 9780199286546.
- Jacobs, A. Z.; and Wallach, H. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT, 375–385. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8309-7.
- Jiménez-Zafra, S. M.; Morante, R.; Teresa Martín-Valdivia, M.; and Ureña-López, L. A. 2020. Corpora Annotated with Negation: An Overview. *Computational Linguistics*, 46(1): 1–52.
- Kapania, S.; Taylor, A. S.; and Wang, D. 2023. A hunt for the Snark: Annotator Diversity in Data Practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI, 1–15. Hamburg Germany: ACM. ISBN 978-1-4503-9421-5.
- Klie, J.-C.; Castilho, R. E. D.; and Gurevych, I. 2024. Analyzing Dataset Annotation Quality Management in the Wild. *Computational Linguistics*, 50(3): 817–866.
- Klie, J.-C.; Webber, B.; and Gurevych, I. 2023. Annotation Error Detection: Analyzing the Past and Present for a More Coherent Future. *Computational Linguistics*, 49(1): 157–198.
- Kovashka, A.; Russakovsky, O.; Fei-Fei, L.; and Grauman, K. 2016. Crowdsourcing in Computer Vision. *Foundations and Trends® in Computer Graphics and Vision*, 10(3): 177–243.
- Kumar, A.; Pounpeth, N.; Yang, D.; Farrell, E.; Lambert, B. L.; and Groh, M. 2026. When large language models are reliable for judging empathic communication. *Nature Machine Intelligence*, 8(2): 173–185.
- Le Ludec, C.; Cornet, M.; and Casilli, A. A. 2023. The problem with annotation. Human labour and outsourcing between France and Madagascar. *Big Data & Society*, 10(2): 20539517231188723.
- Lease, M. 2011. On Quality Control and Machine Learning in Crowdsourcing. In *Human Computation: Papers from the 2011 AAAI Workshop (WS-11-11)*.
- Li, G.; Zheng, Y.; Fan, J.; Wang, J.; and Cheng, R. 2017. Crowdsourced Data Management: Overview and Challenges. In *Proceedings of the 2017 ACM International Conference on Management of Data*, 1711–1716. Chicago Illinois USA: ACM. ISBN 978-1-4503-4197-4.
- Li, T.; Manns, C. J.; North, C.; and Luther, K. 2019. Dropping the Baton?: Understanding Errors and Bottlenecks in a Crowdsourced Sensemaking Pipeline. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–26.
- Lin, C. 2022. How Forest Became Data: The Remaking of Ground-Truth in Indonesia. In Goldstein, J.; Nost, E.; and Project Muse, eds., *The nature of data: infrastructures, environments, politics*. Lincoln: University of Nebraska Press. ISBN 978-1-4962-1715-8 978-1-4962-3278-6.
- Lin, C. K.; and Jackson, S. J. 2023. From Bias to Repair: Error as a Site of Collaboration and Negotiation in Applied Data Science Work. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–32.
- MacQueen, K. M.; McLellan, E.; Kay, K.; and Milstein, B. 1998. Codebook Development for Team-Based Qualitative Analysis. *CAM Journal*, 10(2): 31–36.
- Marshall, C. C.; Goguladinne, P. S.; Maheshwari, M.; Sathe, A.; and Shipman, F. M. 2023. Who Broke Amazon Mechanical Turk?: An Analysis of Crowdsourcing Data Quality over Time. In *Proceedings of the 15th ACM Web Science Conference 2023*, WebSci, 335–345. Austin TX USA: ACM. ISBN 979-8-4007-0089-7.
- McDonald, N.; Schoenebeck, S.; and Forte, A. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–23.
- Meisner, C.; Duffy, B. E.; and Ziewitz, M. 2024. The labor of search engine evaluation: Making algorithms more human or humans more algorithmic? *New Media & Society*, 26(2): 1018–1033.
- Messick, S. 1987. VALIDITY. *ETS Research Report Series*, 1987(2): i–208.
- Messick, S. 1995. Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9): 741–749.
- Miceli, M.; Dinika, A.-A.; Kauffman, K.; Salim Wagner, C.; Sachenbacher, L.; Hanna, A.; and Gebru, T. 2025. Methodological Considerations for Centering Workers’ Epistemic Authority in AI Research. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2): 1698–1710.
- Miceli, M.; Schuessler, M.; and Yang, T. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2): 1–25.
- Northcutt, C. G.; Athalye, A.; and Mueller, J. 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.
- Pang, R. Y.; Cenatempo, J.; Graham, F.; Kuehn, B.; Whisenant, M.; Botchway, P.; Stone Perez, K.; and Koennecke, A. 2023. Auditing Cross-Cultural Consistency of Human-Annotated Labels for Recommendation Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT, 1531–1552. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0192-4.
- Parmar, M.; Mishra, S.; Geva, M.; and Baral, C. 2023. Don’t Blame the Annotator: Bias Already Starts in the Annotation Instructions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, 1779–1789. Dubrovnik, Croatia: Association for Computational Linguistics.
- Paullada, A.; Raji, I. D.; Bender, E. M.; Denton, E.; and Hanna, A. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11): 100336.

- Peterson, J. C.; Battleday, R. M.; Griffiths, T. L.; and Rusakovsky, O. 2019. Human Uncertainty Makes Classification More Robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Peychev, M.; Mueller, M. N.; Fischer, M.; and Vechev, M. 2023. Automated Classification of Model Errors on ImageNet. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Plank, B. 2022. The “Problem” of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 10671–10682. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.
- Pradhan, V. K.; Schaekermann, M.; and Lease, M. 2022. In Search of Ambiguity: A Three-Stage Workflow Design to Clarify Annotation Guidelines for Crowd Workers. *Frontiers in Artificial Intelligence*, 5: 828187.
- Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do ImageNet Classifiers Generalize to ImageNet? In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 5389–5400. PMLR.
- Reidsma, D.; and Carletta, J. 2008. Reliability Measurement without Limits. *Computational Linguistics*, 34(3): 319–326.
- Reidsma, D.; and Op Den Akker, R. 2008. Exploiting ‘subjective’ annotations. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics - Human-Judge ’08*, 8–16. Manchester, United Kingdom: Association for Computational Linguistics. ISBN 978-1-905593-49-1.
- Rothschild, A.; Wang, D.; Jayakumar Vilvanathan, N.; Wilcox, L.; DiSalvo, C.; and DiSalvo, B. 2024. The Problems with Proxies: Making Data Work Visible through Requester Practices. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7 of *AIES*, 1255–1268.
- Röttger, P.; Vidgen, B.; Hovy, D.; and Pierrehumbert, J. 2022. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 175–190. Seattle, United States: Association for Computational Linguistics.
- Röttger, P.; Hofmann, V.; Pyatkin, V.; Hinck, M.; Kirk, H.; Schuetze, H.; and Hovy, D. 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL, 15295–15311. Bangkok, Thailand: Association for Computational Linguistics.
- Sabou, M.; Bontcheva, K.; Derczynski, L.; and Scharl, A. 2014. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).
- Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; and Aroyo, L. M. 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, 1–15. Yokohama Japan: ACM. ISBN 978-1-4503-8096-6.
- Sap, M.; Swayamdipta, S.; Vianna, L.; Zhou, X.; Choi, Y.; and Smith, N. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL, 5884–5906. Seattle, United States: Association for Computational Linguistics.
- Schaekermann, M.; Goh, J.; Larson, K.; and Law, E. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 1–19.
- Scheuerman, M. K.; Hanna, A.; and Denton, E. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–37.
- Scheuerman, M. K.; Wade, K.; Lustig, C.; and Brubaker, J. R. 2020. How We’ve Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1): 1–35.
- Schmarje, L.; Grossmann, V.; Zelenka, C.; Dippel, S.; Kiko, R.; Oszust, M.; Pastell, M.; Stracke, J.; Valros, A.; Volkman, N.; and Koch, R. 2022. Is one annotation enough? - A data-centric image classification benchmark for noisy and ambiguous label estimation. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35 of *NeurIPS*, 33215–33232. Curran Associates, Inc.
- Schroeder, H.; Roy, D.; and Kabbara, J. 2025. Just Put a Human in the Loop? Investigating LLM-Assisted Annotation for Subjective Tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, ACL, 25771–25795. Vienna, Austria: Association for Computational Linguistics.
- Schumann, C.; Olanubi, F.; Wright, A.; Monk, E.; Heldreth, C.; and Ricco, S. 2023. Consensus and Subjectivity of Skin Tone Annotation for ML Fairness. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36 of *NeurIPS*, 30319–30348. Curran Associates, Inc.
- Schwirten, L.; Scholz, J.; Kondermann, D.; and Keuper, J. 2024. Ambiguous Annotations: When is a Pedestrian not a Pedestrian? ArXiv:2405.08794 [cs].
- Sheng, V. S.; Provost, F.; and Ipeirotis, P. G. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery*

- and data mining, KDD, 614–622. Las Vegas Nevada USA: ACM. ISBN 978-1-60558-193-4.
- Shestakofsky, B. 2024. Cleaning up data work: Negotiating meaning, morality, and inequality in a tech startup. *Big Data & Society*, 11(3): 20539517241285372.
- Shmueli, B.; Fell, J.; Ray, S.; and Ku, L.-W. 2021. Beyond Fair Pay: Ethical Implications of NLP Crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL, 3758–3769. Online: Association for Computational Linguistics.
- Simons, R. N.; Gurari, D.; and Fleischmann, K. R. 2020. "I Hope This Is Helpful": Understanding Crowdworkers' Challenges and Motivations for an Image Description Task. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2): 1–26.
- Sloane, M.; Schellmann, H.; Mei, K. X.; Choi, A. S. G.; and Koenecke, A. 2026. The case for stakeholder-driven AI auditing in automatic speech recognition. *Nature Machine Intelligence*, 8(4): 493–494.
- Smart, A.; Wang, D.; Monk, E.; Díaz, M.; Kasirzadeh, A.; Van Liemt, E.; and Schmer-Galunder, S. 2024. Discipline and Label: A WEIRD Genealogy and Social Theory of Data Annotation. Version Number: 1.
- Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP.
- Snyder, H. 2019. Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104: 333–339.
- Tan, Z.; Li, D.; Wang, S.; Beigi, A.; Jiang, B.; Bhattacharjee, A.; Karami, M.; Li, J.; Cheng, L.; and Liu, H. 2024. Large Language Models for Data Annotation and Synthesis: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, EMNLP, 930–957. Miami, Florida, USA: Association for Computational Linguistics.
- Tavakoli, L.; and Zamani, H. 2025. Reliable Annotations with Less Effort: Evaluating LLM-Human Collaboration in Search Clarifications. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, 92–102. Padua Italy: ACM. ISBN 979-8-4007-1861-8.
- Thebault-Spieker, J.; Venkatagiri, S.; Mine, N.; and Luther, K. 2023. Diverse Perspectives Can Mitigate Political Bias in Crowdsourced Content Moderation. In *2023 ACM Conference on Fairness Accountability and Transparency*, 1280–1291. Chicago IL USA: ACM. ISBN 979-8-4007-0192-4.
- Thomas, D. R.; Borchers, C.; and Koedinger, K. R. 2025. Beyond Agreement: Rethinking Ground Truth in Educational AI Annotation. ArXiv:2508.00143 [cs].
- Thomas, D. R.; Borchers, C.; Vanacore, K. P.; Koedinger, K. R.; and Kizilcec, R. F. 2026. Modernizing Ground Truth: Four Shifts Toward Improving Reliability and Validity in AI in Education.
- Torrallba, A.; and Efros, A. A. 2011. Unbiased look at dataset bias. In *CVPR 2011*, CVPR, 1521–1528. Colorado Springs, CO, USA: IEEE. ISBN 978-1-4577-0394-2.
- Törnberg, P. 2024. Best Practices for Text Annotation with Large Language Models. Version Number: 1.
- Uma, A. N.; Fornaciari, T.; Hovy, D.; Paun, S.; Plank, B.; and Poesio, M. 2021. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72: 1385–1470.
- Van Der Wal, O.; Bachmann, D.; Leidinger, A.; Van Maanen, L.; Zuidema, W.; and Schulz, K. 2024. Undesirable Biases in NLP: Addressing Challenges of Measurement. *Journal of Artificial Intelligence Research*, 79: 1–40.
- Vaughan, J. W. 2018. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *Journal of Machine Learning Research*, 18(193).
- Wallach, H.; Desai, M.; Cooper, A. F.; Wang, A.; Atalla, C.; Barocas, S.; Blodgett, S. L.; Chouldechova, A.; Corvi, E.; Dow, P. A.; Garcia-Gathright, J.; Olteanu, A.; Pangakis, N. J.; Reed, S.; Sheng, E.; Vann, D.; Vaughan, J. W.; Vogel, M.; Washington, H.; and Jacobs, A. Z. 2025. Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge. In *Forty-second International Conference on Machine Learning Position Paper Track*, ICML.
- Wang, A.; Morgenstern, J.; and Dickerson, J. P. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3): 400–411.
- Wang, D.; Díaz, M.; Rastogi, C.; Davani, A.; Prabhakaran, V.; Mishra, P.; Patel, R.; Parrish, A.; Ashwood, Z.; Paganini, M.; Teh, T. H.; Rieser, V.; and Aroyo, L. 2025. "Just a Strange Pic": Rethinking 'Safety' in GenAI Image Safety Annotation Tasks from Diverse Annotators' Perspectives. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(3): 2612–2624.
- Wen-Yi, A. W.; Adamson, K.; Greenfield, N.; Goldberg, R.; Babcock, S.; Mimno, D.; and Koenecke, A. 2024. Automate or Assist? The Role of Computational Models in Identifying Gendered Discourse in US Capital Trial Transcripts. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7 of *AIES*, 1556–1566.
- Westwood, S. J. 2025. The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences*, 122(47): e2518075122.
- Xiao, Z.; Zhang, S.; Lai, V.; and Liao, Q. V. 2023. Evaluating Evaluation Metrics: A Framework for Analyzing NLG Evaluation Metrics using Measurement Theory. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, EMNLP, 10967–10982. Singapore: Association for Computational Linguistics.
- Zhang, B. Z.; Yang, T.; Miceli, M.; Haimson, O. L.; and Thomas, M. 2025. The Making of Performative Accuracy in AI Training: Precision Labor and Its Consequences. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI, 1–19. Yokohama Japan: ACM. ISBN 979-8-4007-1394-1.

Zhao, D.; Andrews, J. T. A.; Papakyriakopoulos, O.; and Xiang, A. 2024. Position: Measure Dataset Diversity, Don't Just Claim It. In *The Forty-first International Conference on Machine Learning*, ICML. Vienna, Austria. Version Number: 1.

Zhao, X.; Liu, J. S.; and Deng, K. 2013. Assumptions behind Intercoder Reliability Indices. *Annals of the International Communication Association*, 36(1): 419–480.

Zhou, K.; Blodgett, S. L.; Trischler, A.; Daumé Iii, H.; Suleman, K.; and Olteanu, A. 2022. Deconstructing NLG Evaluation: Evaluation Practices, Assumptions, and Their Implications. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL, 314–324. Seattle, United States: Association for Computational Linguistics.