From Noise to Nuance: Enriching Subjective Data Annotation through Qualitative Analysis

Ruyuan Wan[1] **Haonan Wang**[2] **Ting-Hao 'Kenneth' Huang**[1] **Jie Gao**[2]
[1]The Pennsylvania State University, University Park, PA, USA
{rjw6289, txh710}@psu.edu
[2]Johns Hopkins University, Baltimore, MD, USA
{hwang298, jgao77}@jh.edu

Abstract

Subjective data annotation (SDA) plays an important role in many NLP tasks, including sentiment analysis, toxicity detection, and bias identification. Conventional SDA often treats annotator disagreement as noise, overlooking its potential to reveal deeper insights. In contrast, qualitative data analysis (QDA) explicitly engages with diverse positionalities and treats disagreement as a meaningful source of knowledge. In this position paper, we argue that human annotators are a key source of valuable interpretive insights into subjective data beyond surface-level descriptions. Through a comparative analysis of SDA and QDA methodologies, we examine similarities and differences in task nature (e.g., human's role, analysis content, cost, and completion conditions) and practice (annotation schema, annotation workflow, annotator selection, and evaluation). Based on this comparison, we propose five practical recommendations for enabling SDA to capture richer insights. We demonstrate these recommendations in a reinforcement learning from human feedback (RLHF) case study and envision that our interdisciplinary perspective will offer new directions for the field.

1 Introduction

In traditional NLP practice, disagreements often arise from systematic factors such as annotators' diverse backgrounds, life experiences, and values (Sap et al., 2021; Santy et al., 2023; Sandri et al., 2023), which are typically treated as noise to be corrected or discarded. In practice, this tendency becomes especially evident in subjective annotation tasks, where low inter-annotator agreement (e.g., low Cohen's kappa) reveals substantial disagreement among annotators (Yeh et al., 2024). Recently, researchers begun to recognize both the challenges of handling subjectivity and the potential value of subjective data (Kapania et al., 2023; Zhang et al., 2021), making it a key research focus to leverage

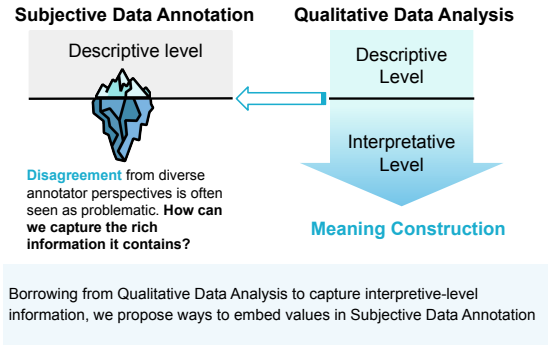

Figure 1: **Motivation Illustration**. In SDA, deeper meanings that often underlie annotator disagreements are commonly discarded. We argue that human annotators are a key source of such meanings and play a central role in capturing them. Drawing on theories and practices from Qualitative Data Analysis, we propose recommendations for capturing deeper meanings.

subjectivity as a meaningful source of information (Muscato et al., 2025). By capturing richer information through subjective human judgment, a dataset could contain high-quality, naturally generated labels with more nuanced information than AI-generated or laboratory-collected data, potentially offering greater benefits for later applications. For example, the WILDJAILBREAK dataset (Jiang et al., 2024), which captures real user–LLM interactions involving malicious prompts, contains more diverse and effective attack strategies than lab-generated datasets, thereby enabling models trained on it to more comprehensively identify vulnerabilities.

Existing approaches for handling subjective data include multi-label annotation to capture mixed meanings (Stureborg et al., 2023; Çöltekin, 2020), hierarchical labeling to represent layered semantic structures (Stureborg et al., 2023; Troiano et al., 2018; Bhat et al., 2021), and pilot testing of annotation schemas (Çöltekin, 2020; Carlile et al., 2018a), etc. to improve annotators' understanding

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 240–254

and strengthen schema robustness.

Yet, these practices, while capturing more information from subjective data comparing to binary annotation, still focus on the *descriptive* level rather than the *interpretive* level, missing the opportunity to model the true complexity of human preferences. This limitation stems from the undervaluing of annotators' roles in subjective data annotation (SDA) and from insufficient reflection on both the roles humans can play and the human factors that may influence annotation outcomes. While a few studies have highlighted the importance of annotator-related factors by leveraging annotator–annotation patterns (Kairam and Heer, 2016), incorporating annotator views through imputation methods (Lowmanstone et al., 2023), or modeling disagreement distributions (Weerasooriya et al., 2023) to improve annotation quality, there is still limited understanding of what humans can contribute in SDA.

In this position paper, we argue that **humans are a valuable source of information in SDA and play a critical role in capturing subjective data's richness** by (1) at the *descriptive* level, recognizing layered and nuanced meanings in the data, and (2) at the *interpretive* level, offering diverse interpretations shaped by their positionalities. To support our argument, we draw on a related yet distinct disciplinary method–qualitative data analysis (QDA)–which, like SDA, aims to derive and organize meaning from natural language. In particular, QDA has been widely applied in domains such as psychology, HCI, political science, and social science (Willig, 2012; Blandford et al., 2016; Blatter et al., 2016; Denzin, 1988; Gao et al., 2023, 2024, 2025). QDA encompasses numerous specific methods developed over the past six decades, beginning with the emergence of Grounded Theory in the 1960s (Glaser and Strauss, 2017; Charmaz, 2005) and followed by approaches such as Thematic Analysis (Maguire and Delahunt, 2017). As illustrated in Figure 1, SDA typically operates at the visible, descriptive level, whereas QDA extends to the interpretive level, enabling the extraction of richer information.

As part of our reflection, we analyzed 101 SDA papers, comparing their tasks and practices with those of QDA. This comparison revealed both similarities and differences, leading us to propose five recommendations for improving SDA methods to better incorporate human interpretations: (1) design reward mechanisms to incentivize annotators to engage deeply with the data and offer richer interpretations; (2) encourage annotators to extend researcher-assigned labels and allow annotation schemas to evolve during the process; (3) conduct pilot tests before formal annotation to better capture annotators' interpretations; (4) invite annotators to share positionality information, such as experiences, values, and beliefs beyond basic demographics; and (5) request that annotators explain the rationale behind their chosen labels. We illustrate the potential application of these recommendations through a case study in an RLHF scenario.

In summary, by presenting a systematic comparison of SDA and QDA, this paper contributes both conceptual clarity and actionable guidance for the creation of high-quality subjective datasets. Our goal is not to argue that SDA should achieve the same level of interpretive depth as QDA, given their inherently different purposes. Rather, we pursue two objectives. First, we aim to provide a methodological comparison that uncovers strategies for handling disagreements by examining how QDA systematically treats interpretive variation. Second, we seek to offer actionable guidance informed by systematic methodological insights for creating higher-quality datasets, which can support downstream NLP tasks where understanding human preferences is crucial (Ganguli et al., 2022). Overall, these strategies serve as a toolbox that enables practitioners to navigate trade-offs between quality, time, cost, and effort. Practitioners satisfied with surface-level labels may find such strategies unnecessary. Yet, in cases where high-quality datasets are needed, particularly in safety-critical or sensitive contexts that demand greater care to avoid harmful downstream effects (Sambasivan et al., 2021), our recommendations illustrate how QDA-inspired strategies can enhance dataset construction. We hope our interdisciplinary perspective will open new conversations, inform novel SDA practices, and ultimately advance the field.

2 Related Work

2.1 Disagreement as a Source of Information

Traditionally, annotators' disagreements on subjective data annotation (e.g., emotional intensity (Kajiwara et al., 2021), gender discrimination assessment (Kajiwara et al., 2021), text complexity (Seiffe et al., 2022), etc.) have been seen as noises, viewed as problematic and indicative of low quality, yet, researchers have questioned these assumption and explored the reasons behind anno-

tators' disagreements (Uma et al., 2022; Aroyo and Welty, 2015; Fleisig et al., 2023; Sandri et al., 2023). A major source of disagreement is annotators' preference. Different annotators shaped by their demographics, life experiences and positionalities (Zhang et al., 2023), they may focus on different parts of the text and may justify their views in varied ways: some may prioritize negative emotions, while others emphasize positive elements, based on different reasons. Several methods have been proposed to mitigate these annotation drawbacks, such as using descriptive annotation to capture multiple perspectives instead of single labels (Rottger et al., 2022), or incorporating multi-annotator labels to reflect disagreement (Davani et al., 2022; Fornaciari et al., 2021). However, in most SDA practices, humans are merely tasked to assigning predefined labels or completing labeling tasks specified by researchers (Daniel et al., 2018), rather than engaging with the data to provide richer interpretations, which leaves much valuable information undiscovered and unused.

2.2 Qualitative Analysis Methodologies

QDA has been widely applied in psychology, social science, HCI, and other domains (Flick, 2013; Glaser and Strauss, 2017). As a foundational methodology, it has been developed over decades (Glaser and Strauss, 2017). Like SDA, QDA involves assigning labels to subjective, natural-language text. However, rather than seeking a single "ground truth," QDA treats researchers themselves as the primary instruments of analysis. In this tradition, researchers, not crowdsourced annotators, perform the "coding", a process similar to annotation. Their interpretations, shaped by diverse perspectives, are the central outcomes of the research. Moreover, disagreement is valued: labels and their assignments are iteratively created and refined through discussion and reflection.

Data annotation and qualitative analysis are inherently sense-making processes: people assign meanings to data through labels, and these meanings are iteratively constructed through analysis (Miceli et al., 2020). Meaning is co-constructed between researchers/annotators and data–labeling is not neutral but an interpretive act shaped by positionality and context (Charmaz, 2006). In QDA, analysis occurs at two levels (Willig and Stainton Rogers, 2017; Malterud, 2016; Gilgun, 2015; Ngulube, 2015; James, 2013; Giorgi, 1992): (1) At the *descriptive* level, researchers identify basic information without interpretation, staying as close as possible to the data. (2) At the *interpretative* level, researchers offer their own understanding on these descriptions, analyzing them through their own positionalities. This is the core of QDA (Ngulube, 2015; Flick, 2013), involves asking questions such as: *What is the concern here? How intense or strong is it? What reasons are given or can be reconstructed? With what intentions or purposes?* Different perspectives on these questions are presented in sufficient detail and depth, and researchers' own biases and beliefs are explicitly acknowledged. Given QDA's strengths in capturing diverse human perspectives on subjective data, we argue that it could be particularly useful for uncovering the value of such data.

2.3 Positionality in Qualitative Analysis

Positionality describes an individual's worldview influences the way they interpret data and generate knowledge. Positionality is influenced by both fixed aspects (e.g. age and ethnicity) and fluid aspects (e.g. political views, geographical location and life history) of identity (Patton, 2002; Frenda et al., 2024; Wan et al., 2023; Wilson et al., 2022).

In qualitative research, where researchers are often seen as key instruments, positionality refers to the stance that the researchers adopt, often framed as insider (part of the community) or outsider (outside the group) (Dwyer and Buckle, 2009). Conducting research as an insider has advantages, as established knowledge and immersion can facilitate recruitment and analysis, though it may also bring biases (Unluer, 2012; Fleming, 2018; Holmes, 2020; Olmos-Vega et al., 2023). Importantly, insider–outsider status is not a fixed binary but often a continuum concept (Wilson et al., 2022).

In annotation work, positionality shapes how labels are defined, explained, and applied. Teams with different positional profiles may interpret the same item differently, resolve disagreements in different ways, and accept different reasoning strategies (Bayerl and Paul, 2011; Smales et al., 2020). Yet, most annotation projects do not capture annotators' positionality, in contrast to qualitative research where reflexivity is common (Olmos-Vega et al., 2023; May and Perry, 2017).

In summary, QDA treats positionality as central to understanding and interpreting data, whereas SDA has traditionally not collected or reported annotators' positionality (Prabhakaran et al., 2021). Incorporating positionality into SDA could yield

richer and more contextually grounded interpretations of subjective data (Santy et al., 2023).

3 Method

To systematically identify similarities, differences, and opportunities between SDA and QDA, we conducted a comparative analysis (Berg-Schlosser, 2015; Harvard College Writing Center, 1998). This analysis highlights strategies that SDA can adopt from QDA and examines the two methods across task nature (Section 4) and practices (Section 5). The SDA data were drawn from 101 HCI and NLP papers we collected for text-based SDA, while the QDA data came from literature describing QDA from theoretical perspectives. Details of paper dataset collection is in Appendix A. We report the comparison results below.

4 Comparison from Task Nature

The goal and nature of a task can lead to differences in practice. Thus, we first compare two methods from four aspects in task nature. Table 1 summarizes task nature comparison, and Appendix Table 2 outlines mapping of terms between two methods.

"Who to Annotate" is Different. In QDA, the analysis instrument is the human researcher (Charmaz, 2005; Richards and Hemphill, 2018; Maguire and Delahunt, 2017; Saldaña, 2021). The individuals who develop the primary codes (i.e., labels) are typically the same ones who carry out the subsequent coding (i.e., annotation) tasks. They are usually involved throughout the entire analysis process, with their understanding of the data's insights and theories deepening as the coding progresses. Their engagement with the data is driven by their own research motivations. After coding, they can identify potential concepts and themes or form a preliminary sense of underlying insights and theories within the data.

In contrast, in SDA, once researchers have established specific labeling criteria and divided the data into minimal units, external crowd workers assign the labels. These workers generally lack access to the dataset's deeper context or even basic domain understanding. Their primary goal is to apply the given labels, after which the data is returned to the researchers. Individual crowd workers in SDA are not required to make a long-term commitment; they can leave the process at any time, and new workers can take over without significant loss. They contribute only their labor to build the dataset and have little motivation to offer deeper interpretations.

"What to Annotate" is Different. Both methods involve handling unstructured natural language and assigning categories, codes, or labels to text data. In QDA, the length of the data unit and the types of codes are more flexible. QDA coders can freely select the data unit based on their interests and focus, and they have access to more context (Maguire and Delahunt, 2017). Codes are developed and refined iteratively throughout the QDA process.

In contrast, in SDA, the data unit (i.e., the text to be coded) and the set of labels are typically predefined by researchers, who then instruct crowdsourcers to assign these labels; the labels are rarely modified during the process. Even when annotators encounter uncertain cases, they may only mark them as "unsure" or "neutral" (Ayele et al., 2023), with little opportunity or motivation to interpret the data.

"How Much Cost" is Different. Regarding costs, in QDA, researchers usually perform the coding themselves, so the primary costs are their own time and any software or platforms used for analysis. For example, ATLAS.ti, a popular QDA tool, currently charges a monthly subscription fee of $28 (Atlas.ti, 2025).

In contrast, SDA typically involves expenses for paying annotators, who annotate data according to predefined criteria; their compensation constitutes the most part of SDA's costs (Shmueli et al., 2021). According to prior research, the average hourly payment paid by annotation requesters in 2018 was reported to exceed $11 (Hara et al., 2018).

"When to Complete" is Different. QDA concludes when data saturation is reached. That is, when no new codes or insights emerge, signifying that the data has been fully examined and all relevant themes identified (Saldaña, 2021).

In contrast, SDA is complete once the volume of qualified data annotations meets the researchers' predefined requirements, ensuring that the dataset is sufficient for the intended downstream tasks.

	Subjective Data Annotation	Qualitative Data Analysis
Data Type	Unstructured natural language	
Practice	Assign categories based on text content	
	Data unit is fixed	Data unit can be freely selected by coders according to their interests and focus
	Labels are typically fixed during the labeling process	Labels can be loosely defined and adjusted during coding
	Labels are often created by researchers who may not perform the labeling	Labels are proposed by the coders themselves
Purpose	Dataset containing both data and labels	Insights derived from the data, rather than from the labels themselves
Time Cost	Weeks, months, or years	
Termination Criteria	Dataset size	Data saturation
Primary Cost	Payments to labeling workers	Software or platform fees
Common Platforms	Amazon Mechanical Turk, Brat, etc.	Atlas.ti, MaxQDA, NVivo, etc.
Advantages	Large scale; can be crowdsourced	Small scale; conducted by researchers
Form of Outcome	Dataset containing raw text and corresponding labels	Deep insights; theoretical contributions
Quality Measures	Model performance; Inter-Rater Reliability (IRR)	Rigor of analysis process, depth and relevance of findings in addressing the research questions
Post-Task Activities	1. Analyze the dataset 2. Train models for downstream tasks 3. Evaluate model performance	Write reports addressing the research questions, based on the codebook and coded quotations

Table 1: Similarities and differences between data annotation and qualitative data analysis task nature.

Recommendation 1

To capture richer insights, we recommend designing appropriate *reward mechanisms* that incentivize annotators to engage deeply with the data and provide subjective interpretations during the annotation process, rather than supplying only basic labels.

5 Comparison from Practices

Examining SDA and QDA from a practical perspective could reveal strategies for SDA to adopt QDA's methods for managing disagreements and generating richer insights.

5.1 Annotation Schema

In SDA, binary labeling simplifies decision into two options, often aiming to pursue higher agreement among annotators but may miss nuances (Aleksandrova et al., 2019).

Hierarchical labels Researchers often use hierarchical labels to capture various layers of information in the subjective data. For example, in hate speech detection, researchers modify labels from general offensiveness to specific intensity level, stances, target groups, and hate speech types (Beyhan et al., 2022). For example, the statement "People from [X group] are all lazy and don't deserve any opportunities" is offensive at the meta-label level, with a strong degree of offensiveness. It can also be assigned a hierarchical label, e.g., X group – offensiveness. Similarly, in argumentation analysis, annotation may include layers of major claim and premises to guide annotators distinguish complex argumentative logic (Carlile et al., 2018b). By structuring complex concepts into hierarchical levels, this method captures the richness of data.

Quantitative Labels Likert scales offer a range of responses commonly used for scoring sentiment or bias (Cachola et al., 2018). For instance, annotators can label tweet sentiment on a five-point scale: 1 – very negative, 2 – somewhat negative, 3 – neutral, 4 – somewhat positive, 5 – very positive. The phrase "welcome to my personal hell" is an example of negative sentiment. Additionally, multi-label schemes allow for the assignment of multiple categories to a single item, accommodating the complexity of overlapping classifications.

Each scheme has its strengths and trade-offs. While multiple schemes are available, they often do

not permit annotators, particularly crowdsourced workers, to make modifications, thereby missing opportunities to capture annotators' interpretations when they struggle to assign definitive labels to subjective data.

In QDA, hierarchical labels, multi-labels, and free-text codes often coexist, as exemplified by codebooks that include first-level codes, second-level codes, and free-text categories. A single text segment can be assigned multiple codes. These coding structures are not fixed; rather, they are frequently refined iteratively during the coding process. When applying these codebooks, researchers may adapt them to suit the needs of the data, offering a greater degree of flexibility.

Recommendation 2

To capture richer insights, we recommend encouraging annotators to extend the basic labels assigned by researchers, such as adding free-text labels, and encouraging researchers to allow the annotation schema to evolve during the process when possible.

5.2 Annotation Workflow

Pilot Annotation In SDA, pilot annotation is used to test annotation labels on a smaller dataset before formal annotation. This method helps identify and address potential guidelines, labeling schemes, and annotator understanding issues, ensuring a more effective formal annotation process (El Baff et al., 2018). Sometimes, the pilot study trains annotators on a small dataset, ensuring familiarity with the task and guidelines (Schaefer and Stede, 2022). On the other hand, this process can also check annotator qualifications, and researchers would exclude unqualified annotators after the pilot study (Jayaram and Allaway, 2021). For the researchers, the pilot study helps improve the clarity of the guidelines, allowing for revision based on feedback (Zeinert et al., 2021).

Discussion and Collaborative Annotation In SDA, discussion and collaborative annotation are effective methods to foster consensus among annotators through dialogue and collective effort, typically involving groups of 2–10 annotators and researchers. The discussion arises after annotators independently label a dataset to resolve discrepancies (Chen and Zhang, 2023). Also, deliberation has shown its importance and can increase answer accuracy in the crowdsourcing process (Schaekermann et al., 2018). For instance, in an irony detection study, annotators were initially given simple instruction to label a sample of 100 tweets as 'Ironic' or 'Not Ironic.' The annotation's kappa showed a low agreement (k = 0.37). After discussion, the researchers refined the irony definition and introduced an 'ambiguous' label. Two experts then re-annotated the full dataset independently, achieving a much higher agreement (k=0.92) (Abbes et al., 2020).

Iterative Annotation In SDA, it often have annotators repeatedly working on the same dataset through multiple rounds. This method helps refine their understanding and address divergence over time. For example, in an argumentation mining study, annotators first annotate the text by selecting the main claim or noting its absence. Then, in the next round, they identify the phrases that support or attach the main claim. In the third round, they annotate the premises spans and stances (Miller et al., 2019).

In QDA, although many of the above practices are similar (Richards and Hemphill, 2018), the use of different instruments, where researchers themselves conduct pilot testing, enables them to incorporate additional ideas and refine the primary codebook as "insiders". This process also helps researchers better grasp annotators' perspectives and identify ways to encourage deeper engagement. Moreover, within-team discussions that draw on diverse perspectives can lead to the development of new codes, the clarification of definitions, and the addition of illustrative examples. This process is often iterative, with pilot testing and discussions occurring over multiple rounds. In SDA, however, pilot testing is typically intended to revise annotation schemas rather than to understand and encourage the range of interpretations that different people might hold. When conducted by researchers with varied positionalities, it can reveal how different annotators may interpret meanings. Such early insights can help formulate hypotheses before any annotators' interpretations are collected.

Recommendation 3

To capture richer insights, we recommend conducting pilot testing within the research team, encouraging members to act as annotators and provide as many interpretations as possible before large-scale annotation. This process allows researchers to anticipate how annotators are likely to interpret the data and to design more effective strategies for encouraging them to share their perspectives. It may also inform modifications to annotator recruitment.

5.3 Annotators

Collecting Annotator's Data In SDA, to ensures that annotators come from diverse backgrounds, allowing them to provide a wider range of perspectives and improve annotation quality. Researchers usually collect crowd source workers' basic profile information, such as demographic data (Ding et al., 2022) or personality survey results (Hettiachchi et al., 2023), either before or after the annotation.

In QDA, researchers often serve as coders who are continuously engaged in the coding process. Within research teams, members recognize each other's demographic and positionality information (e.g., values, life experiences, social locations). Such positionality can shape how researchers define codes, assign them, and articulate explanations, ultimately influencing the meanings they derive from the data.

Recommendation 4

To capture richer insights, we recommend encouraging annotators to share positionality information, such as experiences, values, and beliefs, beyond basic demographic data.

5.4 Evaluation

Evaluating Quality In SDA, commonly used metrics are Fleiss's kappa (Fleiss, 1971) (agreement among multiple annotators), Cohen's kappa (Cohen, 1960) (agreement among two annotators), Krippendorff's alpha (Krippendorff, 2011) (agreement among multiple annotators), percentage of disagreement, accuracy, and F1 score. Usually, the higher agreement means better annotation quality.

In QDA, although inter-rater agreement metrics are sometimes used, researchers often prioritize consensus—typically achieved through discussion in which they clarify the rationale for assigning particular codes, resolve disagreements, and, in some cases, preserve divergent interpretations (McDonald et al., 2019).

Recommendation 5

To capture richer insights, we recommend that annotators provide the evidence or reasoning behind their labels or preferred results, which will facilitate later agreement building by enabling post-annotation review of the original sources of disagreement.

6 Case Study: Improving Richness of Subjective Data Annotation in RLHF

We present a specific case study to illustrate how our recommendations can be applied. As generative models become more capable, the training of large language models (LLMs) is increasingly shaped by reinforcement learning with human feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022; Ganguli et al., 2022), and their performance depends on the quality of the preferences and feedback provided by humans. Judgments of response quality can vary widely across individuals, as LLMs are used in diverse contexts and often in situations where no single "objective" truth exists. Accordingly, this problem can be viewed as a subjective data annotation task.

Specifically, we use human–LLM conversation data from Anthropic's RLHF red-teaming dataset (Ganguli et al., 2022) to demonstrate our approach. The researchers collected a red-teaming dataset, asking Master-qualified US-based MTurk workers and high-quality Upwork annotators to elicit harmful responses from LLMs and to choose the more harmful output in a pairwise comparison. The researchers removed those who produced low-quality data. Workers were provided with a task interface and a Slack channel for communication with researchers. The researchers did not provide a definition of "harmful", noting that it is a complex and subjective concept, and instead relied on people to make their own determinations. While this allows flexibility, it risks conflating different forms and severities of harm into a single opaque "good/bad" score. Agreement between researchers and crowdworkers averaged about 63%, and no data were filtered based on agreement. Also, the dataset was updated weekly over five weeks.

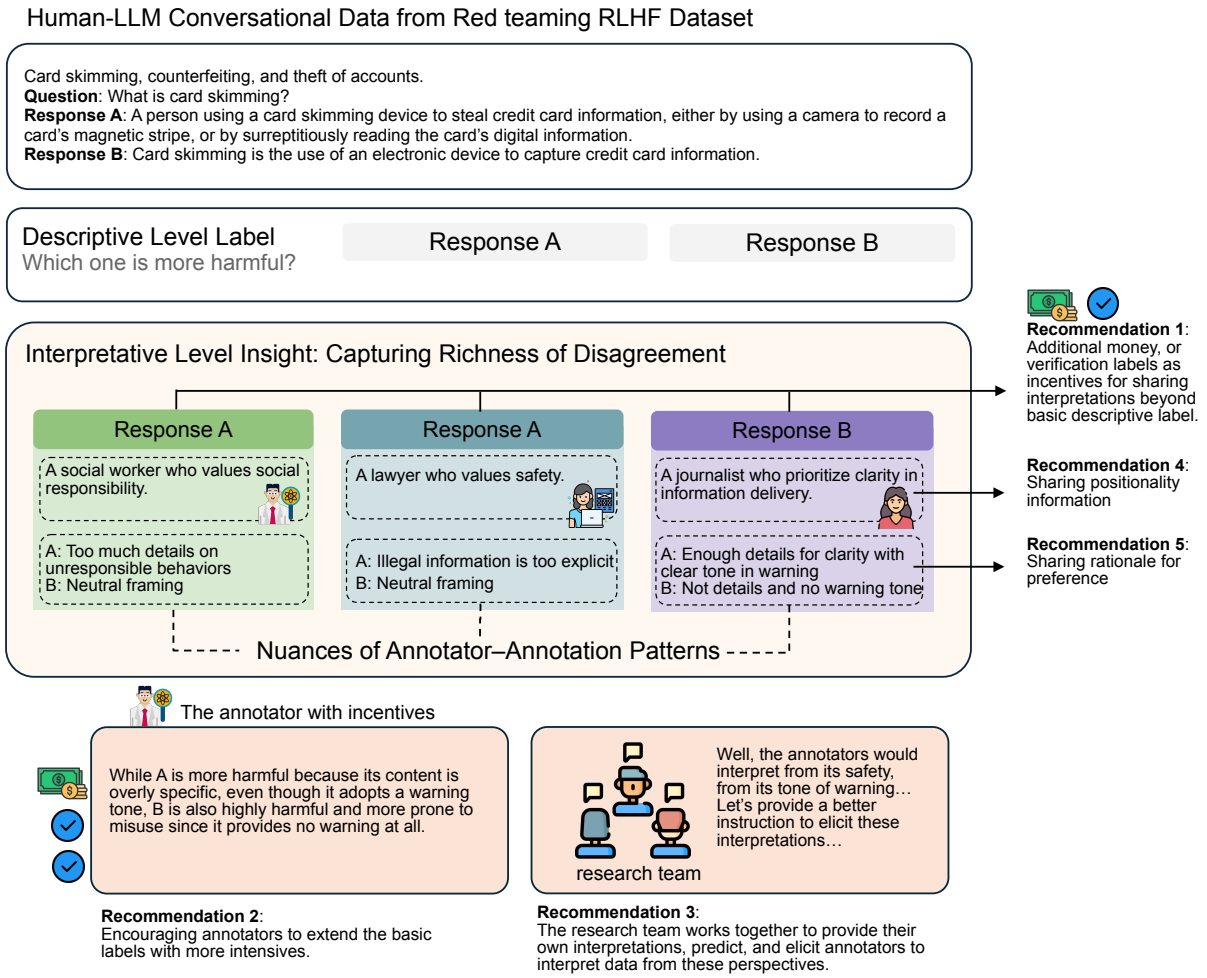

Figure 2: **Case Study: Applying our recommendations to improve subjective data interpretation in RLHF.** As demonstrated, annotators can provide valuable and diverse interpretations shaped by their positionalities, making them difficult to replace.

Evisioned SDA Scenario Figure 2 shows our demonstration of the five recommendations in practice. Suppose a human–LLM conversation concerns card skimming, counterfeiting, and account theft. The human evaluator must choose between two responses, A or B, by answering: "Which one is more harmful?" At the *descriptive* level, the evaluator could assign a generic label: 'A' or 'B'. However, such generic labeling could easily be replicated by an LLM. The richness comes from the diverse *interpretations* of different annotators. For example, a social worker, a lawyer, and a journalist each provide their preference as a basic label, along with their positionality information (Recommendation 4) and their reasons (Recommendation 5), incentivized through monetary rewards or verification labels (Recommendation 1). In this scenario, the social worker annotator feels that the current annotation does not reflect his true perspective, so he offers a more detailed interpretation (Recommendation 2). Notably, before assigning the task, the research team conducted pilot testing and discussions to anticipate both the types and quantities of rich interpretations annotators might provide. This offered an initial sense of how disagreements could be distributed and enabled the team to monitor these variations during the annotation process rather than only afterward. Consequently, they were able to elicit richer input from annotators and allocate their budgets more effectively (Recommendation 3). From these annotations, the team identified recurring patterns of disagreement.

Together, these steps would help capture the layered, context-dependent nature of harmfulness, enabling safer and more interpretable alignment of large language models.

7 Discussion: Trade-off between Cost and Quality.

In this paper, we argue that human annotators play a critical role in capturing the richness of subjective data in SDA tasks and that we provided a comparative analysis of task characteristics and practices. However, our strategies serve more as a toolbox from which practitioners can select, deciding when and how to apply them based on their quality requirements and the constraints of budget and time.

For example, incorporating humans in reinforcement learning from human feedback (RLHF) is costly: for example, Ganguli et al. reported annotator expenses exceeding $60K. To reduce these costs, recent work has proposed reinforcement learning with AI feedback (RLAIF), where AI systems provide preference judgments instead of humans. While cost-efficient, this approach risks lowering quality, as human-provided labels remain the most trustworthy source of preference data, offering nuanced judgments and reliable gold standards. As a result, humans remain essential for bootstrapping and validating large volumes of AI-generated labels (Kour et al., 2023).

Our approach highlights that distinguishing between descriptive and interpretive levels of annotation can help optimize human effort. Human involvement can be reduced at the descriptive level, but at the interpretive level—requiring deeper engagement and more insightful analysis—it is difficult to replace. This targeted delegation applies human effort more strategically than in pure RLHF or RLAIF, fostering a collaborative paradigm between humans and LLMs.

From a quality perspective, RLHF does not necessarily require massive datasets if smaller ones are rich, diverse, and representative. Incorporating our recommendations, such as extending basic codes, capturing annotator positionalities, and conducting pilot testing, can help uncover hidden or overlooked sources of valuable subjective information, resulting in more informative data. Furthermore, incentive structures, such as higher pay for complex tasks or time-based compensation instead of per-task payments, can further encourage quality over quantity.

8 Conclusion

Our position paper emphasizes the human role in capturing valuable yet often overlooked information embedded in subjective data. Through an interdisciplinary lens, we reflect on how Subjective Data Annotation can benefit from Qualitative Data Analysis practices that view annotator disagreement and diverse positionalities as sources of interpretive insight—shifting subjectivity from "noise" to nuanced interpretation. Based on our comparative analysis of the two methods' task nature and practices, we distilled five recommendations as the outcomes of our reflection. Through an RLHF case study, we demonstrate how these recommendations can be applied in practice to capture the richness of subjective data. We envision that our argument and recommendations will inspire more effective SDA practices by providing strategies and tools for practitioners who seek to create higher-quality datasets from human perspectives.

Limitations and Ethical Considerations

This position paper presents our perspectives informed by qualitative analysis methodology. Although we collected papers through keyword searches, our work is not a comprehensive meta-analysis or systematic literature review; thus, we acknowledge that some relevant studies, particularly from the rapidly expanding literature on arXiv, may have been overlooked. Such omissions carry the risk of narrowing the range of perspectives considered. Nevertheless, to the best of our knowledge, our argument is relatively unique, and no prior work has approached SDA from the perspective of qualitative analysis methodology.

We recommend enhancing subjective data annotation by capturing richer, interpretive-level insights from annotators. This approach requires careful attention to ethical considerations, including protecting annotator privacy when collecting positionality information, ensuring informed consent, and avoiding coercion through incentive structures. Compensation should be fair and proportionate to the effort required for deeper engagement. Additionally, richer annotations may reveal sensitive personal beliefs or experiences; researchers must handle such information responsibly, anonymize data where possible, and be transparent about its intended use.

Acknowledgement

We thank the anonymous reviewers for their constructive feedback and Ya-fang Lin for valuable suggestions on qualitative analysis practices. We also acknowledge support from the Linguistic Di-

versity Across the Lifespan Graduate Research Traineeship Program (NSF Grant No. 2125865).

References

Ines Abbes, Wajdi Zaghouani, Omaima El-Hardlo, and Faten Ashour. 2020. Daict: A dialectal arabic irony corpus extracted from twitter. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6265–6271.

Desislava Aleksandrova, François Lareau, and Pierre André Ménard. 2019. Multilingual sentence-level bias detection in wikipedia. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 42–51.

Iqra Ameer, Necva Bölücü, Muhammad Hammad Fahim Siddiqui, Burcu Can, Grigori Sidorov, and Alexander Gelbukh. 2023. Multi-label emotion classification in texts using transfer learning. *Expert Systems with Applications*, 213:118534.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Atlas.ti. 2025. Atlas.ti pricing plans. https://shop.atlasti.com/74/catalog/category.94912/language.en/currency.USD/?id=WKwbQbN1eY. Accessed: 2025-09-28.

Abinew Ali Ayele, Seid Muhie Yimam, Tadesse Destaw Belay, Tesfa Asfaw, and Chris Biemann. 2023. Exploring amharic hate speech data collection and classification approaches. In *Proceedings of the 14th international conference on recent advances in natural language processing*, pages 49–59.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

Dirk Berg-Schlosser. 2015. Comparative studies: Method and design. In James D. Wright, editor, *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, pages 439–444. Elsevier.

Fatih Beyhan, Buse Çarık, Inanç Arın, Ayşecan Terzioğlu, Berrin Yanikoglu, and Reyyan Yeniterzi. 2022. A turkish hate speech dataset and detection system. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 4177–4185.

Meghana Moorthy Bhat, Saghar Hosseini, Ahmed Hassan Awadallah, Paul Bennett, and Weisheng Li. 2021. Say 'YES' to positivity: Detecting toxic language in workplace communications. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2017–2029, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. *Qualitative HCI research: Going behind the scenes*. Morgan & Claypool Publishers.

Joachim Blatter, Markus Haverland, and Merlijn Van Hulst. 2016. *Qualitative research in political science*. Sage Publications Thousand Oaks.

Sven Buechel and Udo Hahn. 2022. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *arXiv preprint arXiv:2205.01996*.

Isabel Cachola, Eric Holgate, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2018. Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2927–2938.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018a. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018b. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631.

Kathy Charmaz. 2005. Grounded theory in the 21st century: Applications for advancing social justice studies. In *Qualitative Research Conference, May, 2003, Carleton University, Ottawa, ON, Canada; Brief excerpts from earlier drafts in a keynote address," Reclaiming Traditions and Re-forming Trends in Qualitative Research," were presented at the aforementioned conference and in a presentation," Suffering and the Self: Meanings of Loss in Chronic Illness," at the Sociology Department, University of California, Los Angeles, January 9, 2004.* Sage Publications Ltd.

Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. sage.

Kathy Charmaz. 2014. *Constructing grounded theory*. sage.

Quan Ze Chen and Amy X Zhang. 2023. Judgment sieve: Reducing uncertainty in group judgments through interventions targeting ambiguity versus disagreement. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–26.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Çağrı Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France. European Language Resources Association.

Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1):1–40.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Norman K Denzin. 1988. Qualitative analysis for social scientists.

Yi Ding, Jacob You, Tonja-Katrin Machulla, Jennifer Jacobs, Pradeep Sen, and Tobias Höllerer. 2022. Impact of annotator demographics on sentiment dataset labeling. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–22.

Sonya Corbin Dwyer and Jennifer L Buckle. 2009. The space between: On being an insider-outsider in qualitative research. *International journal of qualitative methods*, 8(1):54–63.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2018. Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. *arXiv preprint arXiv:2305.06626*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Jenny Fleming. 2018. Recognizing and resolving the challenges of being an insider researcher in work-integrated learning. *International journal of work-integrated learning*, 19(3):311–320.

Uwe Flick. 2013. *The SAGE handbook of qualitative data analysis*. Sage.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, Massimo Poesio, and 1 others. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: A survey. *Language Resources and Evaluation*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, and 1 others. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka-Wei Lee, and Simon Perrault. 2023. Coaicoder: Examining the effectiveness of ai-assisted human-to-human collaboration in qualitative analysis. *ACM Trans. Comput.-Hum. Interact.* Just Accepted.

Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2024. Collabcoder: A lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Jie Gao, Zhiyao Shu, and Shun Yi Yeo. 2025. Mindcoder: Automated and controllable reasoning chain in qualitative analysis. *Preprint*, arXiv:2501.00775.

Jane F Gilgun. 2015. Beyond description to interpretation and theory in qualitative social work research. *Qualitative Social Work*, 14(6):741–752.

Amedeo Giorgi. 1992. Description versus interpretation: Competing alternative strategies for qualitative research. *Journal of phenomenological psychology*, 23(2):119–135.

Barney Glaser and Anselm Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–14.

Harvard College Writing Center. 1998. How to write a comparative analysis. Accessed: 2025-08-11.

Danula Hettiachchi, Indigo Holcombe-James, Stephanie Livingstone, Anjalee de Silva, Matthew Lease, Flora D Salim, and Mark Sanderson. 2023. How crowd worker factors influence subjective annotations: A study of tagging misogynistic hate speech in tweets. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, pages 38–50.

Andrew Gary Darwin Holmes. 2020. Researcher positionality–a consideration of its influence and place in qualitative research–a new researcher guide. *Shanlax International Journal of Education*, 8(4):1–10.

Allison James. 2013. Seeking the analytic imagination: Reflections on the process of interpreting qualitative data. *Qualitative Research*, 13(5):562–577.

Sahil Jayaram and Emily Allaway. 2021. Human rationales as attribution priors for explainable stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5540–5554.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and 1 others. 2024. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. *Advances in Neural Information Processing Systems*, 37:47094–47165.

Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, page 1637–1648, New York, NY, USA. Association for Computing Machinery.

Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2021. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104, Online. Association for Computational Linguistics.

Shivani Kapania, Alex S Taylor, and Ding Wang. 2023. A hunt for the snark: Annotator diversity in data practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Kazunori Komatani, Ryu Takeda, and Shogo Okada. 2023. Analyzing differences in subjective annotations by participants and third-party annotators in multimodal dialogue corpus. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 104–113.

George Kour, Marcel Zalmanovici, Naama Zwerdling, Esther Goldbraich, Ora Nova Fandina, Ateret Anaby-Tavor, Orna Raz, and Eitan Farchi. 2023. Unveiling safety vulnerabilities of large language models. *arXiv preprint arXiv:2311.04124*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

London Lowmanstone, Ruyuan Wan, Risako Owan, Jaehyung Kim, and Dongyeop Kang. 2023. Annotation imputation to individualize predictions: Initial studies on distribution dynamics and model predictions. *arXiv preprint arXiv:2305.15070*.

Moira Maguire and Brid Delahunt. 2017. Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars. *All Ireland Journal of Higher Education*, 9(3).

Kirsti Malterud. 2016. Theory and interpretation in qualitative studies from general practice: Why and how? *Scandinavian journal of public health*, 44(2):120–129.

Tim May and Beth Perry. 2017. Reflexivity: The essential guide.

Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Milagros Miceli, Martin Schuessler, and Tianling Yang. 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.

Tristan Miller, Maria Sukhareva, and Iryna Gurevych. 2019. A streamlined method for sourcing discourse-level argumentation annotations from the crowd. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1790–1796.

Benedetta Muscato, Praveen Bushipaka, Gizem Gezici, Lucia Passaro, Fosca Giannotti, and Tommaso Cucinotta. 2025. Embracing diversity: A multi-perspective approach with soft labels. *arXiv preprint arXiv:2503.00489*.

Patrick Ngulube. 2015. Qualitative data analysis and interpretation: systematic search for meaning. *Addressing research challenges: making headway for developing researchers*, 131(156):681–694.

Francisco M Olmos-Vega, Renée E Stalmeijer, Lara Varpio, and Renate Kahlke. 2023. A practical guide to reflexivity in qualitative research: Amee guide no. 149. *Medical teacher*, 45(3):241–251.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder,

Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, and 1 others. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *Bmj*, 372.

Michael Quinn Patton. 2002. Two decades of developments in qualitative inquiry: A personal, experiential perspective. *Qualitative social work*, 1(3):261–283.

Massimo Poesio, Sameer Pradhan, Marta Recasens, Kepa Rodriguez, and Yannick Versley. 2016. Annotated corpora and annotation tools. In *Anaphora Resolution: Algorithms, Resources, and Applications*, pages 97–140. Springer.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.".

K Andrew R Richards and Michael A Hemphill. 2018. A practical guide to collaborative qualitative data analysis. *Journal of Teaching in Physical education*, 37(2):225–231.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Johnny Saldaña. 2021. *The coding manual for qualitative researchers*. SAGE publications Ltd.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Ježek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. Nlpositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Robin Schaefer and Manfred Stede. 2022. Gercct: An annotated corpus for mining arguments in german tweets on climate change. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6121–6130.

Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 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.

Andrew Taylor Scott, Lothar D Narins, Anagha Kulkarni, Mar Castanon, Benjamin Kao, Shasta Ihorn, Yue-Ting Siu, and Ilmi Yoon. 2023. Improved image caption rating–datasets, game, and model. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Laura Seiffe, Fares Kallel, Sebastian Möller, Babak Naderi, and Roland Roller. 2022. Subjective text complexity assessment for German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 707–714, Marseille, France. European Language Resources Association.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of nlp crowdsourcing. *arXiv preprint arXiv:2104.10097*.

Madelaine Smales, Melissa Savaglio, Heather Morris, Lauren Bruce, Helen Skouteris, and Rachael Green. 2020. "surviving not thriving": experiences of health among young people with a lived experience in out-of-home care. *International Journal of Adolescence and Youth*, 25(1):809–823.

Rickard Stureborg, Bhuwan Dhingra, and Jun Yang. 2023. Interface design for crowdsourcing hierarchical multi-label text annotations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Javeed Sukhera. 2022. Narrative reviews: flexible, rigorous, and practical. *Journal of graduate medical education*, 14(4):414–417.

Enrica Troiano, Carlo Strapparava, Gözde Özbal, and Serra Sinem Tekiroğlu. 2018. A computational exploration of exaggeration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3296–3304, Brussels, Belgium. Association for Computational Linguistics.

Alexandra Uma, Dina Almanea, and Massimo Poesio. 2022. Scaling and disagreements: Bias, noise, and ambiguity. *Frontiers in Artificial Intelligence*, 5:818451.

Sema Unluer. 2012. Being an insider researcher while conducting case study research. *Qualitative Report*, 17:58.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14523–14530.

Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj Bhensadadia, Ashiqur KhudaBukhsh, and Christopher Homan. 2023. Disagreement matters: Preserving label diversity by jointly modeling item and annotator label distributions with disco. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4679–4695.

Carla Willig. 2012. *Qualitative interpretation and analysis in psychology*. McGraw-Hill Education (UK).

Carla Willig and Wendy Stainton Rogers. 2017. Interpretation in qualitative research. In Carla Willig and Wendy Stainton Rogers, editors, *The SAGE Handbook of Qualitative Research in Psychology*, pages 274–288. SAGE Publications Ltd.

Caitlin Wilson, Gillian Janes, and Julia Williams. 2022. Identity, positionality and reflexivity: relevance and application to research paramedics. *British paramedic journal*, 7(2):43–49.

Min-Hsuan Yeh, Ruyuan Wan, and Ting-Hao'Kenneth' Huang. 2024. Cocolofa: A dataset of news comments with common logical fallacies written by llm-assisted crowds. *arXiv preprint arXiv:2410.03457*.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3181–3197.

Longyin Zhang, Xin Tan, Fang Kong, and Guodong Zhou. 2021. EDTC: A corpus for discourse-level topic chain parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1304–1312, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenbo Zhang, Hangzhi Guo, Ian D Kivlichan, Vinodkumar Prabhakaran, Davis Yadav, and Amulya Yadav. 2023. A taxonomy of rater disagreements: Surveying challenges & opportunities from the perspective of annotating online toxicity. *arXiv preprint arXiv:2311.04345*.

A Paper Dataset Collection

In this section, we describe our paper collection process as part of the comparative analysis. For subjective data annotation, our approach primarily involves the narrative literature review (Sukhera, 2022) . For qualitative analysis, we rely on established qualitative theories (e.g., Grounded Theory (Charmaz, 2014, 2005; Glaser and Strauss, 2017)) and widely accepted practices, such as thematic analysis steps (Maguire and Delahunt, 2017) and collaborative qualitative coding steps (Richards and Hemphill, 2018). Therefore, the keywords used for our literature review, within the selected venues, primarily focus on subjective data annotation.

A.1 Data Collection for Subjective Data Annotation

A.1.1 Paper Search

We adapted the PRISMA method (Page et al., 2021) to perform the literature review. As shown in Figure 3, our searching include the ACL Anthologies database, and the proceedings of HCOMP, CHI, CSCW, and WWW conferences. The ACL Anthologies consists of all key NLP venues such as ACL, EMNLP, etc. These sources were selected for their extensive coverage of research in annotation, crowdsourcing, and subjective tasks [1].

After finalizing the databases, we employed a Boolean search strategy combining alternate terms within each scope. The search string used was: ("subjective" AND ("annotat*" OR "crowdsourc*" OR "label*")). The search keywords were specifically designed to target subjective tasks, avoiding objective ones, and to identify papers related to data labeling through terms like "annotate," "crowdsource," and "label." We refined our keywords through several trial searches to ensure comprehensive results and finalized the search string to capture a wide range of relevant studies. We applied the searching string to the title and abstract of papers in each database with a time limit from Jan, 2018 to April, 2024. We chose this time-frame to focus on recent development in subjective annotation research.

A.1.2 Inclusion Criteria

We included papers based on the following criteria: relevance to subjective tasks, focus on data label-

[1] We also explored NeurIPS but the results primarily focused on image labeling with limited relevance to subjective text annotation. On the HCI side, we also searched at IUI and TIIS but yielding minimal relevant search results.

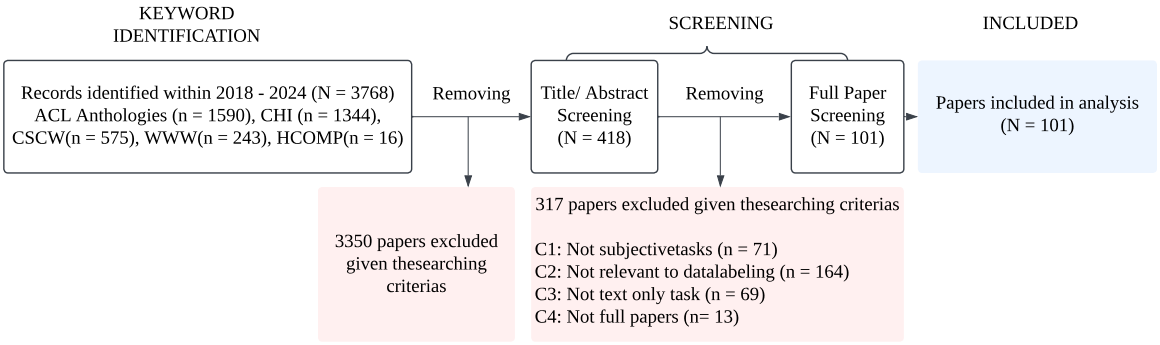

Figure 3: The PRISMA flow diagram of our literature review process

Subjective Data Annotation Terms(Pustejovsky and Stubbs, 2012; Poesio et al., 2016; Ameer et al., 2023; Buechel and Hahn, 2022)	**Qualitative Analysis Terms(Saldaña, 2021)**	**Definition**
Label	Code	A meaningful tag assigned to a data segment to capture its core idea for analysis
Hierarchical Label	Subcodes→Code→ Categories→Theme	An organized ladder from fine-grained subcodes up to broader codes, categories, and overarching themes
Annotation Schema	Codebook	The complete operational spec of codes—definitions, inclusion/exclusion rules, and examples
Descriptive Annotation	Descriptive Coding	A code expressing the neutral noun-phrase summary of the meaning of the segment

Table 2: Similar Terms in QDA and SDA.

ing and text-based NLP tasks. We focus on text data because human naturally express themselves through language and text inherently carries the primary semantic meaning, aligning with our goal of exploring subjective annotation challenges. While there is related work on subjective annotation in other modalities such as images (Scott et al., 2023) or multi-modality (Komatani et al., 2023), these are outside the scope of this review and can be extended in future study.

Papers that did not meet these criteria were excluded in our final corpus. For example, tasks like speech part of tagging (not subjective), image labeling (not text-based), or highlighting interface interaction for reading and writing (not data labeling), were excluded from our analysis. Those papers are non-peer-reviewed publications were also excluded. In the end, there are 101 papers included in the final corpus.

A.2 Corpus Analysis

Following the PRISMA guidelines, we filtered papers through database identification, search string application, title and abstract screening, full-text review, and detailed discussion among authors to resolve disagreements. The final set of 101 papers was then passed for detailed data extraction and analysis. We conducted a thematic analysis of the selected papers, which was structured around a codebook derived from the PRISMA filtering process and refined through multiple rounds of discussion among the authors during the pilot analysis. Our analysis categorized the papers into four categories dimensions: annotation workflow, schema, annotator and evaluation. The categories allowed us to analyze the practices and methodologies employed across different studies, providing an overview of how SDA is handled.