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

Anonymous ACL submission

Abstract

Subjective data annotation (SDA) underpins many NLP tasks, including sentiment analysis, toxicity detection, and bias identification. Conventional SDA often treats annotator disagreement as noise, overlooking its potential to reveal diverse interpretations. We argue that humans play a critical role in uncovering the value of subjective data by providing interpretive-level insights that go beyond surface-level descriptions. In contrast, qualitative data analysis (QDA) explicitly engages with diverse positionalities and treats disagreement as a meaningful source of knowledge. Through a comparative analysis of SDA and QDA methodologies, we examine similarities and differences in task nature (human role, analysis content, cost, and completion conditions) and practice (workflow, schema design, 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 arising from systematic factors such as annotators' diverse backgrounds, life experiences, and values (Muscato, 2025; Sandri et al., 2023)—are typically treated as noise that needs to be corrected or discarded. Recently, scholars have 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 subjectivity as a meaningful source of information (Muscato et al., 2025). By capturing richer information through subjective human judgment, a dataset can contain high-quality, naturally generated labels that yield more diverse and nuanced results than AI-generated or laboratory-collected data, potentially offering greater benefits for later applications. For example, the WILDTEAMING dataset (Jiang et al., 2024) exposed a broader range of model vulnerabilities than alternative sources (Ganguli et al., 2022; Dai et al., 2023) in jailbreaking tasks.

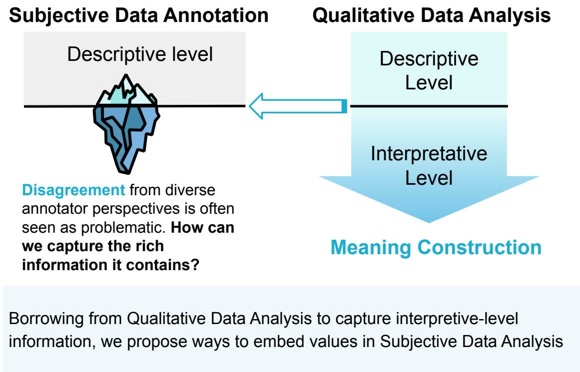

Figure 1: **Motivation Illustration**. In SDA, in-depth meanings that often lead to disagreements between annotators are frequently discarded. We argue that human annotators can play a valuable role in capturing and conveying this information. Drawing on theories and practices from QDA in social science, psychology, and HCI, we offer recommendations for handling such in-depth meanings.

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 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.

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, SDA is a relatively nascent area compared with QDA, which has been widely applied in domains such as psychology, HCI, political science, and social science. 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. We hope our interdisciplinary perspective will inspire new SDA practices and benefit the field.

2 Related Work

2.1 Disagreement as a Source of Information

Traditionally, annotators' disagreements on subjective data annotation (Rottger et al., 2022; Reidsma and op den Akker, 2008) (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 (Uma et al., 2022; Aroyo and Welty, 2015; Fleisig et al., 2023). Researchers have questioned these assumption and explored the reasons behind annotators' disagreements (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. Some primary methods have been proposed to alleviate this kind of simple annotation disadvantages, like descriptive data annotation (Rottger et al., 2022), text conveying mixed emotions could be annotated with descriptive labels to specify the sources of these emotions. However, in most SDA practices, humans are tasked merely with assigning predefined labels rather than engaging with the labels, capturing nuance, or providing richer interpretations. Without incentives (Daniel et al., 2018) to contribute detailed perspectives, annotators often focus solely on completing the labeling task provided by researchers.

2.2 Qualitative Analysis Methodologies

Qualitative Data Analysis (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 and refined over decades (Glaser and Strauss, 2017). Like Subjective Data Analysis (SDA), QDA involves assigning labels to subjective, natural-language text. However, rather than seeking a single, definitive "ground truth," QDA treats researchers themselves as the primary instruments of analysis. In this tradition, researchers—not crowdsourced annotators—perform the "coding," a process akin to annotation. Their interpretations, shaped by diverse perspectives, are the central outcomes of the research. 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 meaning 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, noting that 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, remaining as close as possible to participants' accounts. (2) At the *interpretative* level, researchers offer commentary on these descriptions, analyzing them through the lens of their own positionalities. Interpretation—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 participants' perspectives on these questions are presented in sufficient detail and depth, while researchers' own perceptions, biases, and beliefs are explicitly acknowledged. Thus, QDA's strengths in handling human's diverse perspectives on subjective data can potentially help uncover the value of SDA.

2.3 Positionality in Qualitative Analysis

Positionality describes an individual's worldview influences the way they generate, interpret, and 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 research, positionality reflects the stance that the researchers and participants adopt in a study, often framed as insider (part of the community) or outsider (outside the group) (Dwyer and Buckle, 2009). Some researchers point out that conducting research as an insider has advantages in the data collection process, because the researchers have established topical knowledge and immersion facilitate recruitment and rapport, though it may also bring biases (Unluer, 2012; Fleming, 2018; Holmes, 2020; Olmos-Vega et al., 2023). Meanwhile, some researchers view insider–outsider status as a continuum rather than a strict binary (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

We conducted a comparative analysis (Berg-Schlosser, 2015; Harvard College Writing Center, 1998) of two methods—SDA and QDA—across three dimensions: annotator motivation, annotation schema, and annotation workflow. Our goal was to identify similarities, differences, and opportunities for improvement. Appendix Table 1 presents detailed similarities and differences, and Appendix Table 2 outlines the correspondence of terms between the two methods.

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 appear in Appendix A.

4 Comparison from Task Nature

The goal and nature of a task determine differences in task practices. We first compare the two methods from four aspects in task nature. Detailed comparison is shown in Table 1.

"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

	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 experts
Form of Outcome	Dataset containing raw text and corresponding labels	Deep insights; theoretical contributions
Quality Measures	Model performance; inter-rater reliability (IRR)	Inter-rater reliability (IRR)
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.

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, insights, and expert knowledge. 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.

In contrast, SDA typically involves expenses for paying labelers or crowdsourcing workers, who annotate data according to predefined criteria; their compensation constitutes the most part of SDA's costs (Shmueli et al., 2021).

"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.

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 practice perspective highlights opportunities for SDA to adopt QDA's more iterative and context-aware approaches.

5.1 Annotation Schema

In SDA, binary labeling simplifies decision into two options, often facilitating 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 lower-level category, such as "[X group]," allowing annotators to label it within a hierarchical scheme (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 mapping complex concepts into hierarchical levels, this method translates theoretical frameworks into practical annotation tasks, enhancing consistency and reliability.

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—for example, by 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 help 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, pilot testing enables researchers to incorporate additional ideas and refine the primary codebook by integrating others' interpretations (Richards and Hemphill, 2018). Team discussions that include diverse perspectives may lead to the introduction of new codes, clearer definitions, or additional 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 research team before large-scale annotation to better encourage and guide annotators in providing their own interpretations, as well as to anticipate how they are likely to interpret the data. This could also help modify the 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 process.

In QDA, researchers often serve as coders who are continuously engaged in the coding process. Within research teams, members can readily discern one another'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 two 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 encouraging annotators to share the evidence or reasoning behind their chosen labels or preferred results, thereby supporting later agreement building.

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

As generative models grow more capable, large language model (LLM) applications enhanced by reinforcement learning with human feedback (RLHF) are increasingly used in everyday scenarios, such as ChatGPT. RLHF fundamentally relies on humans providing preference feedback on model responses, which is then used to improve the model itself. In this process, human preference feedback represents a typical RLHF use case (Ouyang et al., 2022; Bai et al., 2022; Ganguli et al., 2022). Judgments about response quality can vary widely across individuals, as they use LLM products in diverse contexts and often in situations where no single "objective" truth exists.

In this case study, we draw on human-LLM conversation data from Anthropic's RLHF red-teaming dataset (Ganguli et al., 2022) to demonstrate how our ideas and recommendations can be applied in practice.

Harmfulness Annotation 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 more harmful. 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 red teamers 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.

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, from Assistants A and B, by answering: "Which one is more harmful?" At the *descriptive* level, the evaluator could assign a generic label provided by researchers—LLM Safety. 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 reasoning (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 to elicit richer interpretations from annotators (Recommendation 3). From these annotations, the team identified a pattern in the 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

Our position emphasizes the critical role of human annotators in capturing the richness of subjective data in SDA tasks—particularly their contributions in offering interpretations shaped by their own positionalities.

However, our case study illustrates an ideal scenario for human involvement, though such situations are often both time- and financially costly. For example, Ganguli et al. report that their expenses for annotators alone exceeded $60K. As a result, recent advances have proposed RLAIF, in which AI—rather than humans—provides feedback. While this approach reduces human effort, it risks lowering the quality of SDA. Human-provided labels remain the most trustworthy source of preference data, offering nuanced judgments and reliable gold standards. This suggests that, although AI feedback scales cheaply, human input is

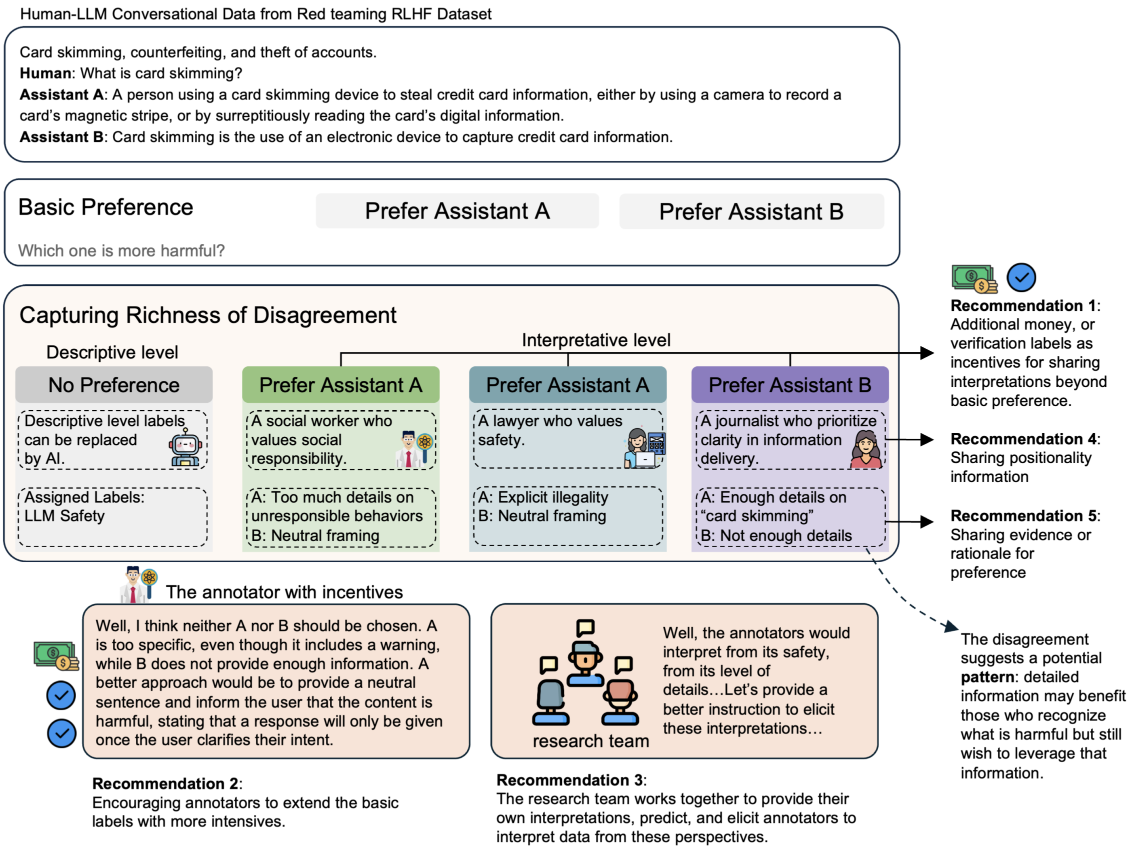

Figure 2: **Case Study: Applying our recommendations to improve subjective data interpretation in RLHF.** Descriptive information can often be generated effectively by LLMs, but the varied interpretations arising from different positionalities are difficult to replace.

indispensable for establishing the reference quality required for alignment. Thus, humans remain in the loop to bootstrap and validate large volumes of AI-generated labels (Kour et al., 2023).

A key advantage of our proposed method is that distinguishing between descriptive and interpretative levels of annotation can help optimize human effort. Human input can be reduced at the descriptive level, whereas its role at the interpretative level—which requires deeper engagement with the data and more insightful analysis—is difficult to replace. This targeted task delegation retains human involvement but applies it more strategically than in pure RLHF or RLAIF, fostering a collaborative paradigm between humans and LLMs.

From a quality perspective, RLHF may not require massive datasets if smaller ones are rich, diverse, and representative. Incorporating our recommendations—such as extending basic codes, capturing positionalities, and conducting pilot testing—can help uncover hidden or overlooked sources of valuable subjective information, resulting in more informative data. Additionally, incentive structures, such as higher pay for more complex tasks or paying by time than task quantity, 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.

9 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.

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.

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.

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.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

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.

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.

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*.

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.

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.

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.

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. 2025. Towards multi-perspective nlp systems: A thesis proposal. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 470–485.

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.".

Dennis Reidsma and Hendrikus JA op den Akker. 2008. Exploiting"subjective"annotations. In *Workshop on Human Judgements in Computational Linguistics, Coling 2008*, pages 8–16. Coling 2008 Organizing Committee.

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.

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 T. Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. Nlpositionality: Characterizing design biases of datasets and models. *Preprint*, arXiv:2306.01943.

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.

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.

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 between subjective data annotation (SDA) and qualitative data analysis (QDA). 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.

[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.

A.1.2 Inclusion Criteria

We included papers based on the following criteria: relevance to subjective tasks, focus on data labeling 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 a overview of how subjective annotation is handled.

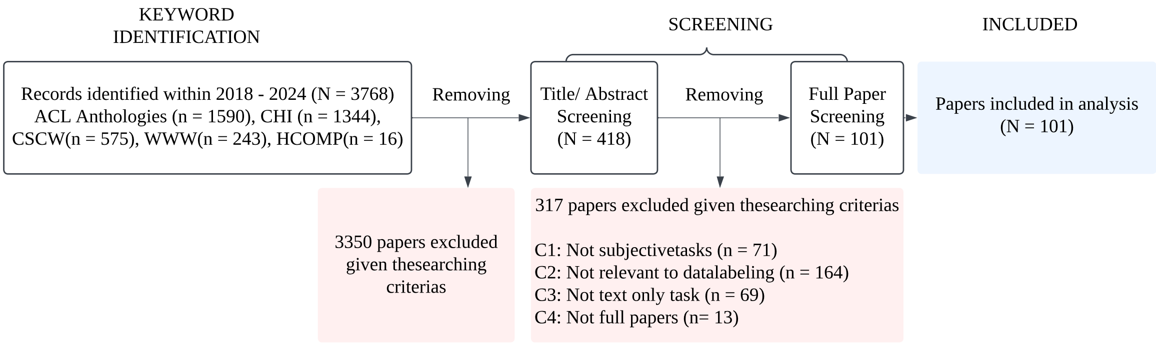

Figure 3: The PRISMA flow diagram of our literature review process

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.