

Revisiting Active Learning under (Human) Label Variation

Cornelia Gruber^{*1} Helen Alber^{*1,2} Bernd Bischl^{1,2}

Göran Kauermann^{1,2} Barbara Plank^{2,3} Matthias Aßenmacher^{1,2}

¹ LMU Munich, Department of Statistics, Germany

² Munich Center for Machine Learning (MCML), Germany

³ LMU Munich, Center for Information and Language Processing (CIS), Germany

^{*}Equal contribution **Correspondence:** cornelia.gruber@lmu.de, helen.alber@stat.uni-muenchen.de

Abstract

Access to high-quality labeled data remains a limiting factor in applied supervised learning. Active learning (AL), a popular approach to optimizing the use of limited annotation budgets in training ML models, often relies on at least one of several simplifying assumptions, which rarely hold in practice when acknowledging human label variation (HLV). Label variation (LV), i.e., differing labels for the same instance, is common, especially in natural language processing. Yet annotation frameworks often still rest on the assumption of a single ground truth, overlooking HLV, i.e., the occurrence of plausible differences in annotations, as an informative signal. In this paper, we examine foundational assumptions about truth and label nature, highlighting the need to decompose observed LV into signal (e.g., HLV) and noise (e.g., annotation error). We survey how the AL and (H)LV communities have addressed—or neglected—these distinctions and propose a conceptual framework for incorporating HLV throughout the AL loop, including instance selection, annotator choice, and label representation. We further discuss the integration of large language models (LLM) as annotators. Our work aims to lay a conceptual foundation for (H)LV-aware active learning, better reflecting the complexities of real-world annotation.

1 Introduction

Prediction algorithms play a central role in many natural language processing (NLP) tasks, like hate speech detection (Basile, 2020), sentiment analysis (Kenyon-Dean et al., 2018), or natural language inference (NLI; Pavlick and Kwiatkowski, 2019). For training such supervised machine learning (ML) models, a notable amount of labeled training data is necessary. However, acquiring high-quality labels is expensive as human crowd workers or, even more expensive, domain experts need to annotate the data. A popular scheme to efficiently guide the

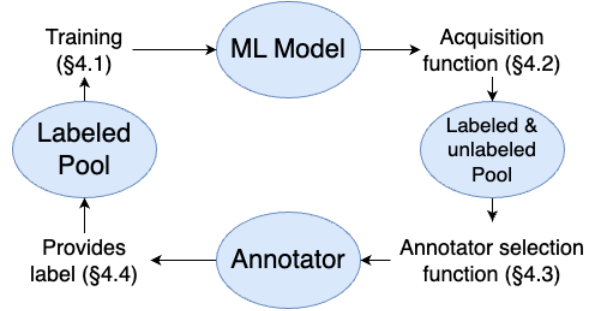


Figure 1: The traditional AL loop with possible adaptations in different steps, leading to generalized label variation aware AL

annotation process and allocate annotation budgets is active learning (AL; Abney, 2007; Settles, 2009). AL aims to maximize the expected predictive performance of the resulting model while minimizing the required number of annotations; often done by iterating the following three steps: (1) Training the ML model on available labeled data. (2) Selecting new instances for labeling from a pool of unlabeled data, usually based on an acquisition function. (3) Labeling these with an oracle. Those steps, which are repeated until the available annotation budget is depleted or the model has reached its target accuracy, rest on the following assumptions:

- A1** There exists a single ground truth label per instance.
- A2** The oracle provides the ground truth labels without any noise.
- A3** The annotation difficulty or cost is equal for all instances.

Equal annotation cost is not, strictly speaking, a critical assumption for AL, but is becoming increasingly important to consider. However, in NLP, those assumptions often are not or cannot be fulfilled. Especially in the presence of human label variation (HLV), i.e., differences in human anno-

tations that are plausible variability due to subjectivity or ambiguity and explicitly no sign of error (Plank, 2022; cf. §3), even the existence of such an omniscient oracle is questionable.

When we move away from these assumptions and acknowledge HLV, the AL loop is extended: an annotator selection function is introduced to choose among multiple annotators with varying perspectives or expertise, rather than assuming a single infallible oracle (cf. Figure 1).

Contributions In this work, we examine the consequences for the AL cycle when its conventional assumptions, i.e., A1 – A3, are violated due to plausible variation in labels, often coined HLV. We begin by discussing foundational assumptions about truth in annotation (§2), laying out different perspectives on label nature and emphasizing the need for a signal-noise decomposition of label variation (LV) into plausible variation (e.g., HLV) and noise. In what follows, we provide an overarching survey of the literature of both fields, i.e., (H)LV and AL, that reveals an emerging line of research integrating aspects of (H)LV into AL (§3), but simultaneously also uncovers shortcomings and misunderstandings between the fields. We then identify and categorize the adaptations required in the AL loop (§4), including modifications to the annotator selection function and considerations for incorporating LLMs. Altogether, we offer a holistic perspective on AL in the presence of (H)LV, aiming to establish a more structured ground for discussion and future empirical investigation by bridging ongoing debates across NLP, empirical ML, statistics, and philosophy.

2 Assumptions about Truth in Annotation

When observing LV in human annotations, it is important to recognize that this variation may arise from both error and HLV (Weber-Genzel et al., 2024), which can be present simultaneously. Throughout this work, we use LV to refer to the observed differences in annotation, which can be decomposed into signal, such as HLV, and noise, such as actual annotation error. Reflecting on the underlying assumptions about the true labels is crucial, as it helps to distinguish between these sources of LV, or, in other words, aids the “interpretation of any observed annotator disagreement” (Röttger et al., 2022, p. 3).

Task Dependence and Subjectivity The extent to which observed LV is attributed to HLV is often judged based on the (assumed) subjectivity of the task (Basile et al., 2021). In domains such as specific image classification tasks in computer vision (e.g., distinguishing between images of cats or dogs), lower levels of HLV may be expected, as the real-world categories constituting the data-generating process (e.g., actual cats or dogs) are typically less subjective. In such cases, higher shares of the observed label variation may be attributable to various types of errors, such as issues arising from imprecise measurement, the compression of real-world information into data, or noise, e.g., introduced during data collection like blurriness in images (Gruber et al., 2025), rather than to HLV.

The notion of inferring task subjectivity from observed LV introduces a certain circularity: LV is intuitively taken as evidence of subjectivity, while assumptions about subjectivity, in turn, inform how much of the variation is attributed to HLV. A more thorough discussion and a systematic approach to operationalizing subjectivity appear essential for future work when aiming to disentangle signal and noise in observed LV.

Worldviews and Nature of Truth Many NLP tasks, as well as certain computer vision tasks (e.g., image segmentation in medicine; Zhang et al., 2020), are assumed to involve a higher degree of subjectivity. Particularly when addressing such tasks, different underlying philosophical assumptions on the nature of truth and the closely related nature of reality can lead to varying methodological implications. For example, adopting a monistic worldview—drawing on the discussion of monism by Russell (1907)—may involve the assumption of a single underlying reality, with different annotations merely being different perspectives on it. In this context, no observed annotation could be fully true or false, and taking individual annotations into account as a distribution on the instance level may be a reasonable approach.

Label Non-Determinism and Levels Whether label variation is viewed from the annotator’s perspective (annotator level) or the instance’s perspective (instance level) can help clarify certain complexities. For example, on the annotator level, label non-determinism, defined as a probabilistic mapping between a real-world instance and a set of labels, can vary in degree between both subject-

tive and less subjective cases, and may even include label-deterministic subjective settings. In contrast, on the instance level, greater subjectivity inherently results in more label non-determinism. Ambiguity, here clearly distinguished from subjectivity, is linked to higher label non-determinism at both levels. While factors like these—label non-determinism, subjectivity, ambiguity, and annotator level vs. instance level—can, in principle, be treated separately, we assume substantial dependencies between them. For instance, even at the annotator level, tasks that are assumed to be more subjective may be likely prone to exhibiting a greater degree of label non-determinism.

Types of Label Nature Approaching the discussion from a more applied perspective, we provide an overview of possible types of labels: (a) discrete class label (also known as “hard label”), (b) label as probability for discrete classes (sometimes referred to as “soft label”, Uma et al., 2021, or “human judgment distribution”), and (c) label as continuous distribution for underlying fixed number of classes (cf. Figure 2). Note, that while the illustration depicts only scenarios with $k \in \{2, 3\}$ classes for simplicity, this schema is generally also applicable to settings with $k > 3$ classes. When viewing the annotation process from a statistical perspective, i.e. making assumptions about the data generating process, each label y_i can be regarded as a realization of a random variable Y . For discrete labels (a), an example in the binary setting is $y_i = 1$, with $Y \sim \text{Bin}(1, p)$; in the ternary case, i.e., three classes, an example is $y_i = [1, 0, 0]$, with $Y \sim \text{Multinom}(1, \mathbf{p})$, $\mathbf{p} = (p_A, p_B, p_C)$. Moving to probability labels (b), the label itself represents a probabilistic belief over class membership. For instance, $y_i = 0.75$ may arise from $Y \sim \text{Beta}(\alpha, \beta)$, and $y_i = [0.6, 0.2, 0.2]$ may be a realization from $Y \sim \text{Dir}(\boldsymbol{\alpha})$, $\boldsymbol{\alpha} = (\alpha_A, \alpha_B, \alpha_C)$. Finally, in the case of distribution labels (c), y_i takes the form of a full probability distribution—for example, $y_i = \text{Beta}(8, 3.5)$ in the binary case or $y_i = \text{Dir}(8, 3, 4)$ in the ternary case. Here, the label y_i is itself a distribution over class probabilities. The distribution of Y is modeled hierarchically by placing priors on the parameters of this distribution, e.g., on α, β in the Beta case or on $\boldsymbol{\alpha}$ in the Dirichlet case (Hechinger et al., 2024a).

We here challenge the common assumption of the first type (discrete class labels, sometimes also referred to as “single ground truths”) by propos-

ing the consideration of the latter two types, both as assumed true labels and requested annotations. The third label type appears to be the least studied of the ones listed; however, some work in uncertainty quantification has begun to explore different label representations (Bengs et al., 2022; Hechinger et al., 2024a; Sale et al., 2024; Wimmer et al., 2023).

In practice, a discrepancy can occur between the type of label assigned by the annotator and the assumed nature of the true label. This mismatch is especially likely when true labels are assumed to be continuous distributions over classes (cf. case (c) in Figure 2), as human annotators are not inherently equipped to give non-discrete annotations (cf. §4.4 for further discussion of the “oracle” in the AL cycle). This discrepancy introduces an irreducible uncertainty and may result in the interpretation that the observed label variation does not necessarily equate to HLV. This again emphasizes the importance of distinguishing between assumptions about the true labels and assumptions that may be required for practical reasons during annotation and the AL loop.

3 Views on Label Variation and Active Learning

In what follows, we outline key stages in how different fields have approached label variation, a phenomenon discussed under various terminologies and theoretical perspectives, and illustrate them with literature examples. Starting from work documenting its occurrence across diverse tasks, we move from approaches that neglect or seek to mitigate LV, to studies that measure variation mainly to steer away from high LV instances. Subsequently, we summarize recent perspectives that embrace (H)LV as a valuable signal, integrating it into learning objectives through distributional labels and motivating its decomposition. We then examine how the field of AL has responded to, incorporated, or overlooked these diverse understandings of LV in its methodological developments.

3.1 Label Variation

Supervised ML depends fundamentally on annotated data, making the quality and nature of labels a central part of the learning process. The phenomenon of LV, i.e., the occurrence of differing annotations for the same instance, both between and within annotators, is not limited to subjective tasks

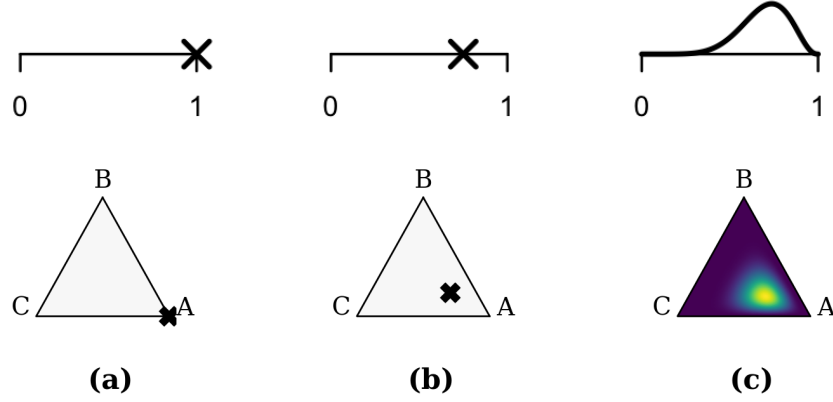


Figure 2: Types of labels visualized. Each label y_i is a realization of a random variable Y .

Top row: binary classes; bottom row: three classes.

(a) Discrete label: $y_i = 1$ with $Y \sim \text{Bin}(1, p)$ (top) and $y_i = [1, 0, 0]$ with $Y \sim \text{Multinom}(1, \mathbf{p})$ (bottom),
(b) probability label: $y_i = 0.75$ with $Y \sim \text{Beta}(\alpha, \beta)$ (top) and $y_i = [0.6, 0.2, 0.2]$ with $Y \sim \text{Dir}(\boldsymbol{\alpha})$ (bottom),
(c) distributional label: $y_i = \text{Beta}(8, 3.5)$ (top) and $y_i = \text{Dir}(8, 3, 4)$ (bottom) with a hierarchical model for the distribution of Y with priors on the parameters of the respective distributions. In the bottom row, ternary plots visualize the relative proportions of three classes as positions within a triangle. Each cross represents a single label, with its location indicating the class composition: the closer a point is to a corner, the higher the class proportion.

but has been found across a wide range of applications. In NLP, this includes tasks such as sentiment analysis (Kenyon-Dean et al., 2018), hate speech detection (Basile, 2020), veridicality judgments (De Marneffe et al., 2012), argumentation mining (Trautmann et al., 2020), natural language inference (Pavlick and Kwiatkowski, 2019), and even tasks traditionally considered “objective” like part-of-speech tagging (Plank et al., 2014b), word sense disambiguation (Passonneau et al., 2012), semantic role labeling (Dumitrache et al., 2019), and named entity recognition (Inel and Aroyo, 2017). Similar variation has also been observed in computer vision tasks like medical image classification and object identification (Uma et al., 2021), or remote sensing (Hechinger et al., 2024b), where annotator disagreement arises from ambiguity and subjectivity in visual interpretation. While most existing works listed treat this as *either* signal or noise, we refrain from exclusively assigning observed label variation to either category in the first place.

Mitigating Label Variation The assumption of a single ground truth label has long dominated ML practice, as reflected in foundational ML literature (Goodfellow et al., 2016; Hastie et al., 2009; Mitchell, 1997). Within this framework, LV is typically regarded as erroneous and to be minimized or corrected (Alm, 2011; Aroyo and Welty, 2015) with Cabitza et al. (2023), for example, documenting widespread practices of “disagreement removal”.

Treating “Hard” Cases Moving beyond the traditional view of LV, early work has begun to explore LV as a potential source of information. Exemplary, Reidsma and Op Den Akker (2008) advocate for analyzing patterns of disagreement, providing an overview of the various factors that may underlie annotator disagreement. However, this line of work uses information from LV to steer ML models away from “hard” cases (i.e., items with high LV), by, e.g., enabling classifiers to abstain from making predictions. Plank et al. (2014a) propose incorporating inter-annotator agreement measures into a cost-sensitive loss function, thereby explicitly integrating LV into the learning process as a signal of uncertainty. The next paragraph discusses approaches seeking to embrace LV more directly by explicitly modeling it, for instance, through adjustments to the nature and interpretation of the labels.

Human Label Variation There are two main bodies of literature relevant to this work addressing differences in human annotations: one that predominantly uses the term *variation* and another that refers to *disagreement*. We adopt the terminology of Plank (2022), who introduced the notion of HLV to conceptualize such differences as plausible and meaningful variations rather than as annotation errors. This perspective has been particularly motivated by developments in NLP, where subjectivity, leading to HLV, is recognized as an inherent property of many language-related tasks (Alm, 2011).

This framework further aligns with the concept of *perspectivism* introduced by Cabitza et al. (2023), which emphasizes that, rather than seeking a single ground truth, collecting multiple labels offers a way to sample the range of perceptions, opinions, and judgments present in a population.

The related body of literature that adopts the term disagreement rather than variation is more heterogeneous in its interpretation and evaluation of annotation differences. While some contributions view such disagreement as plausible or informative (Uma et al., 2021), others primarily treat it as a source of noise or error (Beigman Klebanov and Beigman, 2009). Throughout this section, we review work from both terminological traditions.

Distributional Labels Several contributions have moved beyond discrete labels by aggregating multiple annotations into distributional labels (cf. Figure 2 for the different types of label nature), aligning with a (strong) perspectivist stance. De Marneffe et al. (2012) frame veridicality assessment as a “distribution-prediction task”, using judgments from 10 annotators per instance. Similarly, Aroyo and Welty (2015) view disagreement as a signal and introduce the “Crowd Truth” framework, which incorporates distributional labels through annotation aggregation and addresses factors like the design of annotation guidelines and differing annotator expertise. In computer vision, Peterson et al. (2019) show that training convolutional neural networks on soft labels derived from multiple annotators improves generalization under distributional shift. For NLI, Pavlick and Kwiatkowski (2019) use slider-based annotations to capture uncertainty and argue for models that predict distributions over judgments. More recently, Chen et al. (2022) and Gruber et al. (2024) investigated whether to prioritize more annotators per instance or more annotated instances when working on the distributional level via label aggregation.

However, these contributions primarily address HLV by aggregating multiple annotations per instance, thereby treating distributional labels as post-hoc constructions rather than as *distributional by nature*—as in case (c) above—i.e., labels deliberately designed from the outset to capture uncertainty directly as a characteristic of the label.

Decomposing Label Variation Furthermore, the above contributions tend to conflate LV with HLV, overlooking the simultaneous presence of *both* noise and signal within LV. Incorporating this con-

ceptual distinction, Palomaki et al. (2018) highlight the need to distinguish between actual annotation errors and “disagreement that falls within the acceptable range”, introducing the concept of *acceptable variation*, which may differ across subsets of instances and has direct implications for task design. Weber-Genzel et al. (2024) extend this conceptual distinction to NLI. They address the challenge of identifying annotation error by incorporating validated annotator *labels with explanations* through a second round of validity judgments, rather than relying on post-hoc interpretation alone. This builds on earlier work by Jiang et al. (2023), who identify the phenomenon of *within-label variation*, where, even when the same label is assigned, annotators may vary in their explanations.

Data annotation remains a labor-intensive and complex process, particularly when aiming to analyze or leverage a signal–noise decomposition of LV. The following section, therefore, turns to the field of active learning, which focuses on strategies for optimizing annotation budgets and minimizing annotation effort.

3.2 Active Learning

Active Learning has been a vivid field of research for over 30 years (Aggarwal et al., 2014; Lewis and Catlett, 1994; Settles, 2009; Seung et al., 1992). Settles (2011) already discussed practical issues arising in active learning, including querying in batches, noisy oracles, and variable labeling costs. Zhang et al. (2022) provide a survey on AL for NLP, while Rauch et al. (2023) propose a tailored NLP benchmark for AL.

Annotation Costs and Quality The true costs of annotation are explored in Krishnamurthy et al. (2019); Margineantu (2005); Settles et al. (2008); Tomanek and Hahn (2010); Xie et al. (2018), challenging assumption **A3** (“*The annotation difficulty or cost is equal for all instances.*”) by modeling variation in annotation effort. Gao and Saar-Tsechansky (2020); Donmez et al. (2009) extend this by accounting for annotators with varying accuracies, while Donmez and Carbonell (2008) acknowledge that even oracles might be incorrect depending on task difficulty, both relaxing **A2** (“*The oracle provides the ground truth labels without any noise.*”). Yan et al. (2011, 2012) suggest jointly selecting an instance and an annotator. Furthermore, Zhang and Chaudhuri (2015) and Chakraborty (2020) incorporate both low-cost and expert annota-

tors by assuming a trade-off between cost and label quality. However, these approaches still assume a single ground truth label per instance (reliance on **A1**; “*There exists a single ground truth label per instance.*”) and treat label variation as noise.

In contrast, we highlight the underexplored setting where (H)LV is inherent and may carry an informative signal, arguing that its integration into the AL framework requires rethinking core components such as acquisition and annotation strategies.

Relabeling Relabeling, i.e., collecting additional annotations for previously labeled instances to reduce noise or correct errors, is explored in [Chen et al. \(2022\)](#); [Goh and Mueller \(2023\)](#); [Lin et al. \(2016\)](#); [Yuan et al. \(2024\)](#). These approaches implicitly challenge assumptions **A2** and **A3** by acknowledging annotation errors and varying difficulty. However, they treat disagreement as an error rather than a potentially meaningful signal.

HLV-aware AL A few recent studies have begun to explore how AL can be adapted to account for HLV. [Wang and Plank \(2023\)](#) and [van der Meer et al. \(2024\)](#) suggest strategies to choose *which* human annotator should label an instance. Furthermore, [Baumler et al. \(2023\)](#) suggest aligning model uncertainty with annotator uncertainty. While these works offer valuable insights, they address specific assumptions or propose targeted adaptations to the AL process. In §4, we build on these efforts by systematically analyzing their contributions and organizing them into a broader framework. There, we formalize and categorize key adaptations required for making AL effective in the presence of HLV, and point to open challenges and directions for future research.

4 The Active Learning Loop Revisited

In the following, we discuss the consequences of the assumptions about truth in annotation and the nature of the labels (§2) on each of the steps of the AL loop (as visualized in [Figure 1](#)).

4.1 Training Measure

Traditional AL assumes a single ground truth label provided by an oracle. This aligns naturally with classic supervised ML, where models are optimized based on hard-label measures like Bernoulli loss or cross-entropy. However, in cases where label variation is not due to error but comes from plausible causes, different soft-label

measures are necessary. In such cases, alternative loss measures based on label distributions, such as Kullback-Leibler (KL) divergence ([Koller et al., 2024](#)), Jensen-Shannon divergence, or label embeddings ([Schweden et al., 2025](#)) have been proposed. [Baumler et al. \(2023\)](#) offer solutions by comparing the predicted and observed label distribution, thus directly optimizing for a trustworthy representation of LV.

(C1) Consequence: In the presence of HLV, distributional measures must be used for optimizing and evaluating the classifier.

4.2 Acquisition Function

The acquisition function ranks all unlabeled instances by their usefulness if they were to be labeled. The oracle then provides labels to the most instructive cases. Traditional AL ([Zhang et al., 2022](#)) uses querying strategies based on either *informativeness* or *representativeness*, or *hybrid* approaches ([Ash et al., 2020](#)). Informative querying often uses uncertainty sampling ([Lewis and Gale, 1994](#)), where the samples with the highest predicted label entropy get labeled first, thus the ones with the highest uncertainty. However, with HLV, high entropy can also be integral to the task, and thus not necessarily a sign of uncertainty. This shows that classic entropy sampling is not suitable for AL in the presence of HLV. Representativeness sampling favors samples that represent the unlabeled pool well. However, classical representativeness sampling ignores the option of labeling some instances multiple times to represent HLV properly and is thus also unsuitable for HLV. Further, defining representativeness in distributions is not trivial. One option to take HLV into account is to precede the AL loop by training a prediction model for annotator disagreement (entropy) and then changing the acquisition function to query samples where the predicted annotator entropy and model entropy diverge the most ([Baumler et al., 2023](#)).

(C2) Consequence: In the presence of HLV, classical informativeness or representativeness sampling are unsuitable, as they ignore the option of labeling instances multiple times and fail to process distributional labels.

4.3 Annotator Selection Function

The assumption of having an oracle providing the single ground truth label is not suitable in subjective tasks, where the distribution of human opinions is of interest, or other tasks with high (assumed) HLV. Therefore, an additional step in the AL cycle needs to be considered: the selection of annotators. In many crowd worker settings, it is possible to inquire about labels from a specific annotator. Extending this thought, different “types” of annotators could be queried, e.g., not only human workers but also large language models (LLM). This is also known as “pre-annotation” (Zhang et al., 2022) in the pre-LLM era, and analogously as “LLM-as-annotator” (or “LLM-as-a-judge”; Wu et al., 2024; Zheng et al., 2023) today, where the idea is that a model’s predictions are given to human annotators to confirm or adjust. Consequently, an overarching annotator selection strategy needs to evaluate whether a language model or a human shall provide the label, and whether a specific annotator (e.g., representing a minority) or a specific LLM could provide the label. Recent work has extended the AL framework to include not only sample selection but also annotator selection. Wang and Plank (2023) introduce a multi-head model that jointly selects the most informative instance and the most suitable annotator. In contrast, van der Meer et al. (2024) focus on ensuring representativeness and diversity in annotator selection, proposing a strategy that balances labeler perspectives to reflect the underlying population of interpretations better. The idea of using LLMs as annotators is pursued in Bansal and Sharma (2023); Xia et al. (2025); Zhang et al. (2023).

(C3) Consequence: In the presence of HLV it matters *who* provides the label(s). An annotator acquisition function must decide not only whether to query a human or a language model, but also whether to obtain one or multiple annotations and from *which* specific annotator or model.

4.4 Quality of Label and Uncertainty

The quality of annotators is an important area of research in NLP, which becomes increasingly meaningful when diversity in annotations is present (or required; Sorensen et al., 2024) and label noise cannot be easily separated from the plausible share

of label variation. Currently, most work either assumes variation is noise (Goh and Mueller, 2023; Yan et al., 2016; Zhang et al., 2015; Zhao et al., 2011) or all variation in labels represents true HLV (van der Meer et al., 2024; Wang and Plank, 2023). Particularly, when the ground truth label is a distribution and multiple annotators provide labels, detecting annotation noise in HLV is a complex endeavor (Weber-Genzel et al., 2024). Now, when not only different humans annotate the data, but samples can also be processed by LLMs, assessing the label quality is non-trivial either (Ni et al., 2025). Also, in the process of labeling, human annotators usually provide a single label, while an LLM could directly provide distributions (Chen et al., 2024; Pavlovic and Poesio, 2024). This makes LLMs as annotators especially attractive in the presence of HLV and for providing labels for case (c) depicted in Figure 2.

(C4) Consequence: In the presence of HLV, it is non-trivial to distinguish true label variation from noise, especially when labels can be sourced from both humans and language models, each with differing capabilities and output formats.

5 Conclusion

In this work, we provide an overview of the crucial connection between the fields of (human) label variation and active learning. Our comprehensive overview of the existing literature in the individual fields helps building bridges between different, but connected, streamlines of research, paving the way for the identification of critical aspects to consider in the AL loop in the presence of HLV. Our critical assessment of these aspects aims to further point out potential avenues for future research to deal with them in a more nuanced and reflective manner. In doing so, we uncover several crucial assumptions about labels which are often *implicitly* made in traditional AL. However, we argue that they need to be made *explicit*. While providing a unified and implemented solution to the discussed problems is beyond the scope of the paper, we still hope to contribute to ongoing research debates on (H)LV by providing a fresh perspective from a different angle on existing problems and encourage new work addressing label-variation-aware active learning.

Limitations

While this work provides a structured discussion on active learning in the presence of human label variation, several limitations remain. The philosophical discussion on annotation truth is a conceptual suggestion rather than a prescriptive framework. For example, we do not address annotation tasks where it is assumed that *no* ground truth exists, or discuss other frameworks like imprecise probabilities for representing human label variation. Moreover, not all discussed adaptations are implemented in AL pipelines yet, requiring empirical validation. Additionally, we do not explore alternative methods for gathering human annotations that may better accommodate HLV in detail. Lastly, the reliability of “LLM-as-annotator” remains an open question. While LLMs can reduce costs and provide label distributions, their biases and lack of accountability pose challenges.

Acknowledgments

CG is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research. HA and MA are funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure – NFDI 27/1 - 460037581.

References

- Steven Abney. 2007. *Semisupervised Learning for Computational Linguistics*, 1st edition. Chapman & Hall/CRC.
- Charu C. Aggarwal, Xiangnan Kong, Quanquan Gu, Jiawei Han, and S. Yu Philip. 2014. Active learning: A survey. In *Data Classification: Algorithms and Applications*, pages 599–634. Chapman and Hall/CRC.
- Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112.
- 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.
- Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. [Deep batch active learning by diverse, uncertain gradient lower bounds](#). In *International Conference on Learning Representations*.
- Parikshit Bansal and Amit Sharma. 2023. [Large Language Models as Annotators: Enhancing Generalization of NLP Models at Minimal Cost](#). *arXiv preprint. ArXiv:2306.15766* [cs].
- Valerio Basile. 2020. [It’s the End of the Gold Standard as We Know It: Leveraging Non-aggregated Data for Better Evaluation and Explanation of Subjective Tasks](#). In *AIxIA 2020 – Advances in Artificial Intelligence: XIXth International Conference of the Italian Association for Artificial Intelligence, Virtual Event, November 25–27, 2020, Revised Selected Papers*, pages 441–453, Berlin, Heidelberg. Springer-Verlag.
- Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. [We need to consider disagreement in evaluation](#). In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.
- Connor Baumler, Anna Sotnikova, and Hal Daumé III. 2023. [Which Examples Should be Multiply Annotated? Active Learning When Annotators May Disagree](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10352–10371, Toronto, Canada. Association for Computational Linguistics.
- Beata Beigman Klebanov and Eyal Beigman. 2009. [From Annotator Agreement to Noise Models](#). *Computational Linguistics*, 35(4):495–503.
- Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. 2022. [Pitfalls of epistemic uncertainty quantification through loss minimisation](#). In *Advances in neural information processing systems*.
- Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. [Toward a Perspectivist Turn in Ground Truthing for Predictive Computing](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868. Number: 6.
- Shayok Chakraborty. 2020. [Asking the Right Questions to the Right Users: Active Learning with Imperfect Oracles](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3365–3372. Number: 04.
- Beiduo Chen, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, and Barbara Plank. 2024. [“seeing the big through the small”: Can LLMs approximate human judgment distributions on NLI from a few explanations?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14396–14419, Miami, Florida, USA. Association for Computational Linguistics.
- Derek Chen, Zhou Yu, and Samuel R. Bowman. 2022. [Clean or annotate: How to spend a limited data collection budget](#). In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 152–168, Hybrid. Association for Computational Linguistics.

- Marie-Catherine De Marneffe, Christopher D. Manning, and Christopher Potts. 2012. [Did It Happen? The Pragmatic Complexity of Veridicality Assessment](#). *Computational Linguistics*, 38(2):301–333.
- Pinar Donmez and Jaime G. Carbonell. 2008. [Proactive learning: cost-sensitive active learning with multiple imperfect oracles](#). In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 619–628, Napa Valley California USA. ACM.
- Pinar Donmez, Jaime G. Carbonell, and Jeff Schneider. 2009. [Efficiently learning the accuracy of labeling sources for selective sampling](#). In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 259–268, Paris France. ACM.
- Anca Dumitrache, Lora Aroyo, and Chris Welty. 2019. [A crowdsourced frame disambiguation corpus with ambiguity](#). 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 2164–2170, Minneapolis, Minnesota. Association for Computational Linguistics.
- Ruijiang Gao and Maytal Saar-Tsechansky. 2020. [Cost-Accuracy Aware Adaptive Labeling for Active Learning](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03):2569–2576. Number: 03.
- Hui Wen Goh and Jonas Mueller. 2023. [ActiveLab: Active Learning with Re-Labeling by Multiple Annotators](#). *arXiv preprint*. ArXiv:2301.11856 [cs].
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT Press.
- Cornelia Gruber, Katharina Hechinger, Matthias Assenmacher, Göran Kauermann, and Barbara Plank. 2024. [More labels or cases? assessing label variation in natural language inference](#). In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 22–32, Malta. Association for Computational Linguistics.
- Cornelia Gruber, Patrick Oliver Schenk, Malte Schierholz, Frauke Kreuter, and Göran Kauermann. 2025. [Sources of Uncertainty in Supervised Machine Learning – A Statisticians’ View](#). *arXiv preprint*. ArXiv:2305.16703 [stat].
- Trevor Hastie, Robert Tibshirani, Jerome H. Friedman, and Jerome H. Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.
- Katharina Hechinger, Christoph Koller, Xiao Xiang Zhu, and Göran Kauermann. 2024a. [Human-in-the-loop: Towards Label Embeddings for Measuring Classification Difficulty](#). *arXiv preprint*. ArXiv:2311.08874 [cs].
- Katharina Hechinger, Xiao Xiang Zhu, and Göran Kauermann. 2024b. [Categorising the world into local climate zones: towards quantifying labelling uncertainty for machine learning models](#). *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73(1):143–161.
- Oana Inel and Lora Aroyo. 2017. [Harnessing Diversity in Crowds and Machines for Better NER Performance](#). In *The Semantic Web*, pages 289–304, Cham. Springer International Publishing.
- Nan-Jiang Jiang, Chenhao Tan, and Marie-Catherine De Marneffe. 2023. [Ecologically Valid Explanations for Label Variation in NLI](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10622–10633, Singapore. Association for Computational Linguistics.
- Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhandari, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. [Sentiment analysis: It’s complicated!](#) In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895, New Orleans, Louisiana. Association for Computational Linguistics.
- Christoph Koller, Göran Kauermann, and Xiao Xiang Zhu. 2024. [Going Beyond One-Hot Encoding in Classification: Can Human Uncertainty Improve Model Performance in Earth Observation?](#) *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–11. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.
- Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé Iii, and John Langford. 2019. [Active Learning for Cost-Sensitive Classification](#). *Journal of Machine Learning Research*, 20(65):1–50.
- David D. Lewis and Jason Catlett. 1994. [Heterogeneous Uncertainty Sampling for Supervised Learning](#). In William W. Cohen and Haym Hirsh, editors, *Machine Learning Proceedings 1994*, pages 148–156. Morgan Kaufmann, San Francisco (CA).
- David D. Lewis and William A. Gale. 1994. [A Sequential Algorithm for Training Text Classifiers](#). *arXiv preprint*. ArXiv:cmp-lg/9407020.
- Christopher Lin, M Mausam, and Daniel Weld. 2016. [Re-Active Learning: Active Learning with Relabeling](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).
- Dragos D. Margineantu. 2005. [Active cost-sensitive learning](#). In *IJCAI*, volume 5, pages 1622–1623.
- Tom M. Mitchell. 1997. *Machine Learning*, nachdr. edition. McGraw-Hill series in Computer Science. McGraw-Hill, New York.

- Jingwei Ni, Yu Fan, Vilém Zouhar, Donya Rooein, Alexander Hoyle, Mrinmaya Sachan, Markus Leipold, Dirk Hovy, and Elliott Ash. 2025. [Can Large Language Models Capture Human Annotator Disagreements?](#) *arXiv preprint*. ArXiv:2506.19467 [cs].
- Jennimaria Palomaki, Olivia Rhinehart, and Michael Tseng. 2018. A case for a range of acceptable annotations. In *SAD/CrowdBias@ HCOMP*, pages 19–31.
- Rebecca J. Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012. [Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations](#). *Lang. Resour. Eval.*, 46(2):219–252.
- Ellie Pavlick and Tom Kwiatkowski. 2019. [Inherent Disagreements in Human Textual Inferences](#). *Transactions of the Association for Computational Linguistics*, 7:677–694. Place: Cambridge, MA Publisher: MIT Press.
- Maja Pavlovic and Massimo Poesio. 2024. [The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation](#). In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.
- Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625.
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014a. [Learning part-of-speech taggers with inter-annotator agreement loss](#). In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics.
- Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014b. [Linguistically debatable or just plain wrong?](#) In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.
- Lukas Rauch, Matthias Aßenmacher, Denis Huseljic, Moritz Wirth, Bernd Bischl, and Bernhard Sick. 2023. [ActiveGLAE: A Benchmark for Deep Active Learning with Transformers](#). In *Machine Learning and Knowledge Discovery in Databases: Research Track*, pages 55–74, Cham. Springer Nature Switzerland.
- Dennis Reidsma and Rieks Op Den Akker. 2008. [Exploiting ‘subjective’ annotations](#). In *Proceedings of the Workshop on Human Judgements in Computational Linguistics - HumanJudge ’08*, pages 8–16, Manchester, United Kingdom. Association for Computational Linguistics.
- Paul Röttger, 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.
- Bertrand Russell. 1907. [II.—On the Nature of Truth](#). *Proceedings of the Aristotelian Society*, 7(1):28–49.
- Yusuf Sale, Viktor Bengs, Michele Caprio, and Eyke Hüllermeier. 2024. Second-order uncertainty quantification: a distance-based approach. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML’24*, pages 43060–43076, Vienna, Austria. JMLR.org.
- Christoph Schweden, Katharina Hechinger, Göran Kauermann, and Xiao Xiang Zhu. 2025. [Can Uncertainty Quantification Benefit From Label Embeddings? A Case Study on Local Climate Zone Classification](#). *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14.
- Burr Settles. 2009. Active Learning Literature Survey. *University of Wisconsin, Madison*, 52.
- Burr Settles. 2011. From Theories to Queries: Active Learning in Practice. In *Active learning and experimental design workshop in conjunction with AISTATS 2010*, pages 1–18.
- Burr Settles, Mark Craven, and Lewis Friedland. 2008. Active Learning with Real Annotation Costs. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*.
- H. S. Seung, M. Opper, and H. Sompolinsky. 1992. [Query by committee](#). In *Proceedings of the fifth annual workshop on Computational learning theory, COLT ’92*, pages 287–294, New York, NY, USA. Association for Computing Machinery.
- Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A Roadmap to Pluralistic Alignment. In *Proceedings of the 41st International Conference on Machine Learning*, Vienna.
- Katrin Tomanek and Udo Hahn. 2010. [A comparison of models for cost-sensitive active learning](#). In *Coling 2010: Posters*, pages 1247–1255, Beijing, China. Coling 2010 Organizing Committee.

- Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. [Fine-Grained Argument Unit Recognition and Classification](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9048–9056. Number: 05.
- Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. [Learning from Disagreement: A Survey](#). *Journal of Artificial Intelligence Research*, 72:1385–1470.
- Michiel van der Meer, Neele Falk, Pradeep K. Murukanaiah, and Enrico Liscio. 2024. [Annotator-centric active learning for subjective NLP tasks](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18537–18555, Miami, Florida, USA. Association for Computational Linguistics.
- Xinpeng Wang and Barbara Plank. 2023. [ACTOR: Active learning with annotator-specific classification heads to embrace human label variation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2046–2052, Singapore. Association for Computational Linguistics.
- Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. [VariErr NLI: Separating annotation error from human label variation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.
- Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. 2023. [Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures?](#) In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 2282–2292. PMLR. ISSN: 2640-3498.
- Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. [Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge](#). *Preprint*, arXiv:2407.19594.
- Yu Xia, Subhojyoti Mukherjee, Zhouhang Xie, Junda Wu, Xintong Li, Ryan Aponte, Hanjia Lyu, Joe Barrow, Hongjie Chen, Franck Dernoncourt, Branislav Kveton, Tong Yu, Ruiyi Zhang, Jiuxiang Gu, Neseeren K. Ahmed, Yu Wang, Xiang Chen, Hanieh Deilamsalehy, Sungchul Kim, Zhengmian Hu, Yue Zhao, Nedim Lipka, Seunghyun Yoon, Ting-Hao Kenneth Huang, Zichao Wang, Puneet Mathur, Soumyabrata Pal, Koyel Mukherjee, Zhehao Zhang, Namyong Park, Thien Huu Nguyen, Jiebo Luo, Ryan A. Rossi, and Julian McAuley. 2025. [From selection to generation: A survey of LLM-based active learning](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14552–14569, Vienna, Austria. Association for Computational Linguistics.
- Kaige Xie, Cheng Chang, Liliang Ren, Lu Chen, and Kai Yu. 2018. [Cost-sensitive active learning for dialogue state tracking](#). In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 209–213, Melbourne, Australia. Association for Computational Linguistics.
- Songbai Yan, Kamalika Chaudhuri, and Tara Javidi. 2016. [Active Learning from Imperfect Labelers](#). In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.
- Yan Yan, Glenn M. Fung, Rómer Rosales, and Jennifer G. Dy. 2011. [Active learning from crowds](#). In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1161–1168.
- Yan Yan, Romer Rosales, Glenn Fung, Faisal Farooq, Bharat Rao, and Jennifer Dy. 2012. [Active Learning from Multiple Knowledge Sources](#). In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pages 1350–1357. PMLR. ISSN: 1938-7228.
- Bo Yuan, Yulin Chen, Yin Zhang, and Wei Jiang. 2024. [Hide and seek in noise labels: Noise-robust collaborative active learning with LLMs-powered assistance](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10977–11011, Bangkok, Thailand. Association for Computational Linguistics.
- Chicheng Zhang and Kamalika Chaudhuri. 2015. [Active learning from weak and strong labelers](#). In *Advances in neural information processing systems*, volume 28. Curran Associates, Inc.
- Jing Zhang, Xindong Wu, and Victor S. Sheng. 2015. [Active Learning With Imbalanced Multiple Noisy Labeling](#). *IEEE Transactions on Cybernetics*, 45(5):1095–1107. Conference Name: IEEE Transactions on Cybernetics.
- Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Ciccirelli, Frederik Barkhof, and Daniel C. Alexander. 2020. Disentangling human error from the ground truth in segmentation of medical images. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS ’20*, Red Hook, NY, USA. Curran Associates Inc.
- Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. [LLMaAA: Making large language models as active annotators](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103, Singapore. Association for Computational Linguistics.
- Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022. [A Survey of Active Learning for Natural Language Processing](#). In *Proceedings of the 2022 Conference*

on Empirical Methods in Natural Language Processing, pages 6166–6190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Liyue Zhao, Gita Sukthankar, and Rahul Sukthankar. 2011. [Incremental Relabeling for Active Learning with Noisy Crowdsourced Annotations](#). In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 728–733.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.