Understanding and Addressing Uncertainty of the Crowd

Quan Ze Chen

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Amy X. Zhang, Chair

Katharina Reinecke

Daniel S. Weld

Program Authorized to Offer Degree:
Computer Science & Engineering

University of Washington

Abstract

Understanding and Addressing Uncertainty of the Crowd

Quan Ze Chen

Chair of the Supervisory Committee:
Chair Amy X. Zhang
Computer Science & Engineering

Uncertainty is crucial to understanding human judgments. Whether it's annotators producing data used to train and evaluate machine learning systems, teaching staff assigning grades to open-ended student responses, or online communities adjudicating the moderation action to apply to a piece of content, many groups and individuals need to account for and address the uncertainty that comes along with making judgments. As the application of computing technology expands to more areas of society, groups and individuals are faced with the need to make judgments on increasingly complex, subjective, and nuanced tasks at scale.

This dissertation presents a set of novel tools and processes that improve upon how we capture, distinguish, and address various sources of uncertainty present in individual and collective human judgments. I will start by introducing, Goldilocks, a tool for conducting scalar rating annotations that enables different sources of uncertainty to be distinguished while also improving consistency. Then I will introduce case law crowdsourcing as a process that enables capturing similar insights about uncertainty on complex categorical classification judgment tasks by utilizing prior decisions in the form of precedent cases. Following this, I will present Cicero, a tool that addresses one specific source of uncertainty – disagreement – through multi-turn, contextual deliberation. Finally, I will tie together individual tools for understanding and addressing uncertainty through a dynamic workflow that applies a targeted intervention on a per-instance scale to reduce uncertainty using measurements

that allow us to distinguish the source of uncertainty. I conclude by discussing the limitations of current tools and give some insights for future work in designing new tools and processes that natively support judgment under uncertainty.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

GLOSSARY

AMBIGUITY: Uncertainty observed in judgments that manifests through each individual adjudicator judging a case.

ADJUDICATOR: In this work, we will refer to a person conducting scale-based judgments on cases generally as an 'adjudicator'.

CASE: A general term to refer to a specific instance of a problem in the human judgment process, encompassing the item being judged and any associated contextual information. For example, a case in an image annotation problem may take the form of a single image, while a case in a content moderation problem may consist of the post being moderated as well as background context like metadata about the post author.

DISAGREEMENT: An emergent uncertainty resulting from interpreting the judgments by multiple different adjudicators on a case in aggregate.

SCALE-BASED JUDGMENT: A scale-based judgment is a judgment task where a single case is judged by an adjudicator against a scale, such as a rating scale or set of categorical labels.

ACKNOWLEDGMENTS

My work would not have been possible without the support and guidance provided by these amazing people. I would like to thank all of you who have helped me throughout my journey. To all the following, I dedicate this thesis:

- my parents Yun and Xueliang, and my sister Emily, for providing invaluable emotional support through my Ph.D.;

- my advisor Amy Zhang, whose advice and support helped turn many of the ideas in this work into reality;

- my dissertation committee members, Dan Weld, Katharina Reinecke, and Tanu Mitra, who have provided valuable feedback on my work;

- my undergraduate research advisors, Wei Xu and Chris Callison-Burch, whose encouragement and support brought me to pursue undergraduate research and eventually graduate school;

- my collaborators and mentors that I have been fortunate to meet through my internships, Besmira Nushi, Tobias Schnabel, Saleema Amershi, Mike Schaekermann, and Matt Lease, all of whom have made my internship experiences amazing;

- my friends and office mates, Tongshuang (Sherry) Wu, Gagan Bansal, Jonathan Bragg, Chris Lin, Octavian Vlad Murad, Ruotong Wang, Kevin Feng, Raymond Fok, and Marissa Radensky, who have contributed many bits of inspiration throughout our random discussions;

- my lab mates from the Lab for Human-AI Interaction (formerly CrowdLab) and Social Futures Lab, who have encouraged me to become more social;

- my mentees, Andre Ye and Teanna Barrett, who are continuing the work of making contributions to building better uncertainty tools;

- and finally, all of the users and crowd workers who participated in my user studies and contributed your time and effort in providing data—thank you!

Chapter 1

INTRODUCTION

The digital revolution and the advent of modern computing have brought significant changes to human civilization, of which one of the most paradigm shifting is that it has enabled humanity to automate intellectual work at a scale never seen before. Throughout the years, advances in computing have expanded our perception by allowing us to make sense of vast swaths of information, increased our productivity by allowing us to coordinate across space and time, and democratized the accessibility of knowledge by allowing us to retrieve information on demand. With the recent advances in the capabilities of artificial intelligence (AI) and machine learning (ML), computing has also presented opportunities to assist us in a new frontier—the realm of making judgments.

Historically, the ability to make judgments—by considering a collection of observations, forming an understanding, and coming to a conclusion—has been seen as a uniquely human trait. However, the pervasiveness of modern computing has driven the demand for formalizing these components of judgments through the medium of *data*, be it measurements from sensors to decisions coded up by humans. In the field of medicine, doctors rely on digitally recorded diagnostic tests and imaging, and note down their diagnosis and treatment in electronic health records [103]. In the field of education, teachers are increasingly utilizing digital learning management systems to collect student assignments, conduct exams, and assign grades [246, 163, 33]. Even communities are taking advantage of data to inform and record their judgments, from academic communities using preference data to coordinate the planning of conference sessions [53] to Wikipedia moderators using logs to investigate attempts to distort consensus via sockpuppetry [251]. Consequently, all of this data has then opened up the opportunity for computers to automate the task of making judgments at scale by simulating how we do it ourselves. As these models become more capable at capturing complexity, though, we are also creating higher demand for the data that these

models fit to, and inevitably, as with any data, we run into the problem of uncertainty. While mechanisms for measuring, understanding, and addressing uncertainty in measurements like sensor readings are often built into the data collection process, the same cannot be said for those decisions coded up by humans, which are often captured through tools that over-estimate the confidence of individuals and then aggregated with models that make under-informed assumptions about agreements of a group.

In this dissertation, I argue that an important and oft-overlooked uncertainty lies in the scale-based human judgments we collect and that if we are to create *data-driven* computational tools for assisting human judgments, we need new **tools** and **processes** to empower groups and individuals to create better data under uncertainty. More concretely:

> We need better tools and processes to collect and interpret human judgments that account for the presence of diverse sources of uncertainty. And in order to address this demand, we need to build: (1) tools for *mitigating and capturing* various sources of uncertainty during the **initial collection of human judgments**; (2) workflows to *address* uncertainty observed **after judgments are collected**; and (3) meta-processes to dynamically *coordinate* when and how to apply uncertainty interventions **throughout a human judgment collection task**.

This dissertation serves to explore some directions for how the broader space of tooling around collecting, interpreting, and addressing uncertainty in human judgments can be designed and improved. In this work, I will introduce and evaluate tools used to collect scale-based human judgments across several input modalities such as continuous scalar rating and categorical classification, and demonstrate how they can be used to capture and distinguish various sources of uncertainty. I will also explore the space of interventions that help to more effectively address one of the sources of uncertainty—disagreement. Finally, I'll present a design for a higher-level meta-process that ties together tools for measuring uncertainty with interventions to address uncertainty, making for a more efficient approach to systematically account for uncertainty. In combination, the systems and workflows I will describe in this thesis provide the initial foray toward building tooling that natively account

for uncertainty and empower both those collecting judgments and making judgments to evaluate and express uncertainty.

1.1 Uncertainty and Scale-based Human Judgments

While the broad space of "human judgments" covers a wide variety of scenarios—ranging from high-stakes judgments in the legal space that involve months or years of effort by expert lawyers all the way down to everyday snap decisions like a group of friends choosing which restaurant to go to—much of the human judgments we formalize into data for training computation systems is created through **scale-based** judgments where groups of non-experts are asked to place an individual **case** (the entity being judged, such as a digital photo, an online comment *etc.*) onto a **scale** (a set of pre-defined judgments, such as a set of labels for object recognition, or a continuous numeric scale for toxicity *etc.*). This kind of human judgment task is quite common largely because of its simplicity: Scale-based judgments often require little effort from the people producing the judgments (the *human adjudicators*) as they only need to make their decision by selecting from a pre-defined set of possible outcomes. Having judgments selected out of a pre-defined set also reduces the amount of expertise needed, since it can be easier to arrive at an answer using strategies like the process of elimination. Lastly, unlike other types of judgments, such as pairwise comparisons where *relationships between* cases are captured rather than judgments on the case, scale-based judgments produce data that is easy to consume later since each judgment is a single standalone answer. As a result of this simplicity, scale-based judgment tasks have seen massive adoption in the realm of crowdsourced dataset creation, where non-expert crowds are quickly assembled to do a large amount of simple "microtasks". With relatively minimal effort, *requesters* on crowdsourcing platforms can define the problem they want to collect judgments for in the form of a scale-based judgment task template, automatically assemble individual cases from their data, and then deploy these cases as a large batch of microtasks to crowd marketplaces where non-expert crowd *workers* can pick up and complete them. To support this, a variety of tools and workflows geared towards improving the quality of these judgments has been developed, mainly manifesting in the form of *quality control* mechanisms and designs for crowd tasks. However, as problems we are collecting

scale-based judgments on have increased in complexity and subjectiveness, limitations in how these tools engage with uncertainty has meant that it is time for us to develop better tools for scale-based judgments.

1.1.1 Engaging with Uncertainty in Human Judgment Tools

Many early forms of these scale-based judgment tasks centered around collecting data on objective properties of the cases being judged and largely using human adjudicators as a (somewhat noisy) sensor. For example, scale-based judgments were used to recognize characters in written text [177, 164], categorize images based on what was depicted in them [226], or verify whether specific relationships were present from natural language snippets [178, 305]. In these situations, it was often assumed that an objective ground truth existed (often in the form of an expert judgment), and the crowd 'sensors' offered a way to approximate this truth albeit with some noise to contend with. Thus, the earliest attempts at engaging with the uncertainty in the form of judgments that did not agree, focused on identifying which sensors (human adjudicators) were more reliable (accurate). Some tools were developed to de-noise the human judgment data post-hoc, focusing on improving overall quality of data filtered judgments by using voting mechanisms [269] over individual responses or, later, softer approaches that only re-weighted them [65]. Other tools focused on the initial collection of the judgments, making use of training and testing procedures to weed out low-performing human adjudicators [292, 74], or incorporating 'gold' cases with known answers to continuously evaluate the reliability of them [200] on the job.

As these tasks scaled up, though, deeper inspection into the quality of annotations revealed that in many cases disagreeing human judgments were not the fault of crowd workers being unreliable non-experts, but rather reflected the fact some cases were just more challenging to the point where even experts could not confidently determine what the "ground truth" was—the case was just *ambiguous*. Tasks like text recognition or object classification might seem simple and straightforward, but as one looks into the specific sets of cases, you will eventually get cases where it's impossible to tell whether that was a '6' or a '0' or whether that blob in the distance is a person or just an image artifact. With this, tools had

to contend with the fact that, indeed, sometimes given the information and context available in the case, no ground truth existed. The answer was just indeterminable. Of course, that didn't mean nothing could be done: Tools started incorporating the measurement and monitoring of uncertainty—in the form of inter-annotator agreement [115] on a case when they were otherwise high quality, it may be too ambiguous, in which case sometimes the decision was made by the requesters or experts to exclude it from the set of cases altogether.

With the overall successful utilization of scale-based judgments on these objective judgment tasks, we also started seeing the expanded application of these tools to more complex and nuanced tasks. Now scale-based judgments were being used to measure vaguely defined or subjective concepts, such as the sentiment of a piece of text [30], the toxicity of a comment [10], the amount of effort to perform a task [116], or the quality at which a (generated) image adequately reflected a caption [272]. Unlike earlier tasks, these new judgment tasks present even more complexity when it comes to uncertainty. Rather than simply *missing* "ground truth" because experts fail to confidently determine it, in these settings it's often the case that the entire concept of an 'expert' does not make much sense—what's to say one judgment of the adequacy level of a generated image is better than another? Indeed, as we build these new datasets of judgments, these judgments are serving to *establish* the "ground truth" rather than being a low-cost method to *approximate* it. It may be tempting to continue working with these new sources of uncertainty using existing tools, but we run the risk of either artificially creating clarity on a problem when there isn't [106], or removing subjectivity in an attempt to avoid uncertainty so thoroughly that the problem we are collecting judgments on no longer resembles the original problem. At the end of the day, the various sources contributing to uncertainty have not disappeared—individual adjudicators are still making mistakes, cases can still be ambiguous, and the truth may not clear—so we need new tools to better engage with new complexities in uncertainty.

1.1.2 Understanding and Quantifying Uncertainty

Whether it's errors, ambiguous instances, inconsistent scale interpretations, or simply annotators disagreeing, various difference sources can contribute to the final uncertainty in

human judgments. Thus before we can really discuss the tools to capture and mitigate some of these sources of uncertainty, we need to introduce some ways that have been used to think about how to categorize and quantify the sources of uncertainty.

One classic framework, utilized by fields like biostatistics [132] but also gaining traction in AI and ML [130], is to understand uncertainty based on what we could have done to address it. This framework categorizes uncertainty into two main types: **epistemic** uncertainty, where we *could have* theoretically reduced the uncertainty if we had better measurement instruments (*e.g.*, a more precise scale or ruler), a made fewer simplifications for convenience (*e.g.*, accounting for some factor that we omitted because its effect was small), or understood the property itself better (*e.g.*, knowing that something is a factor in the outcome in the first place); and **aleatoric** uncertainty, where we *cannot* reduce the uncertainty because it an inherent component of stochasticity involved in the process that generated the data in the first place (*e.g.*, a random dice roll). Under this interpretation, we can see that components of scale-based judgment tools like quality control mechanisms mainly engaged with **epistemic** uncertainty, while agreement metrics and thresholds for identifying ambiguous cases engaged with **aleatoric** uncertainty. However, while this framework presents practical insights into how we might quantify and address uncertainty, some have also criticized the utility stability of this distinction in practice [91]. As our understanding of a problem improves, what was once considered irreducible aleatoric uncertainty could end up actually being the result of epistemic sources of uncertainty we had not yet identified. For example, in scale-based judgments on toxicity, we might observe disagreements attributable to differences in personal preference towards the use of profanity—perhaps an irreducible aleatoric uncertainty. However, digging deeper we might conclude that the personal preference may be affected by cultural background of those human adjudicators, and (should we want to) we may be able to better model it through first collecting demographics that we did not before—thus making the problem epistemic.

As seen above, defining a comprehensive theoretical framework for quantifying uncertainty is still an open problem and ultimately what makes one framework more useful than another may depend on how the quantified uncertainty is then used to produce actionable insights for each specific problem. In this work, I will not attempt to create or prescribe

any particular theoretical framework for quantifying uncertainty, but instead look at some possible ways to distinguish sources of uncertainty that can lead to actionable insights for scale-based human judgment tools. One of the lenses I'll use to look at uncertainty is through the distinction of two types of uncertainty—**ambiguity** and **disagreement**—the distinction of which I will introduce below:

Ambiguity derives from the idea that, in scale-based judgment tasks, there will be individual cases being judged where the case does not include enough information or context for a judgment to be confidently established. Earlier we have discussed this idea in the context of experts and establishing ground truth for annotation. To make the idea broadly applicable to situations where there are no experts, I will define the idea of ambiguity as follows:

> **Ambiguity** is uncertainty observed in judgments that manifests through each individual adjudicator judging a case.

For example, an annotator judging a classification task for whether a person is present in an image encountering a dark image where they couldn't confidently tell if a person was present would reflect *ambiguity*—the particular case was ambiguous to this particular annotator. It's important to note that, ambiguity doesn't always have to result from a property that's solely related to the instance itself, rather it can also reflect a mismatch between level of resolution the human adjudicator can provide and the options of the scale: A reviewer who is asked to give a 1–5 star rating judgment of a restaurant can encounter ambiguity if they did not have a strong opinion of a mediocre restaurant. Similarly, annotator on a classification task might also encounter ambiguity if they found that their judgment did not correspond to any of the provided label options or if they believe multiple labels apply in a single-label task [207].

Disagreement derives from the idea that, in scale-based judgment tasks involving multiple adjudicators, different adjudicators may still produce different judgments that are incompatible with each other. In existing tools and workflows where adjudicators provide no additional information for their judgment, disagreement is often used as the only observable measurement of uncertainty, with different hypotheses of what factors actually contribute

to the observed disagreement. Here, though, I will define disagreement as:

> **Disagreement** is an emergent uncertainty resulting from interpreting the judgments by multiple different adjudicators on a case in aggregate.

For example, annotators who disagree on whether the presence of profanity alone makes a piece of text toxic may give the same comment different toxicity ratings. We note that the quality of the our observation of disagreement depends on how well a tool or process for human judgment determine incompatibility. In existing single-choice tools the observation of disagreement may be limited to a binary case of whether the judgments match exactly, while tools that have better expressivity (like ranges or multi-choice sets) may be able to allow measurement of cases where there is partial agreement between human adjudicators. Like ambiguity, there can be different reasons for disagreement, ranging from different interpretations of scale options, different lines of reasoning, different background knowledge, or simply different individual preference.

As we can see from the examples above, while this is not necessarily comprehensive, **ambiguity** and **disagreement** can represent very distinct sources of uncertainty in group judgments that can provide actionable insight into different potential avenues for interventions. In the case of *ambiguity*, it makes more sense to attempt to resolve the uncertainty by focusing on individual judgments, such as adding more context so each adjudicator has more information to work with. On the other hand, when there is *disagreement*, it makes more sense to resolve the disagreement by identifying where different adjudicators diverge, such as by having adjudicators deliberate with with each other.

With this in mind, in the next sections, I will introduce a set of new tools and workflows for working with scale-based human judgments aimed at addressing the limitations of existing tools by focusing on ambiguity and disagreement.

1.2 Building Better Tools for Scale-based Human Judgments

Now that I have introduced the background around uncertainty and scale-based human judgment tools, I will lay out the vision of how we can build better tools for scale-based human judgments given the new complexities involved in the problems we're collecting

judgments on. In this section, I will dive into two sub-problems that can be resolved by better tools: (1) how uncertainty caused by limitations of prior tools can be mitigated while making sure remaining uncertainty is still captured; and (2) how to address uncertainty, focusing on the case of *disagreement*, once it is observed.

1.2.1 Mitigating and Capturing Uncertainty in Initial Judgments

Mitigating uncertainty by calibrating understanding of scales: One of the strengths of scale-based human judgments is the simplicity of being able to interpret each judgment individually with respect to the scale. However, this can be a problem if different adjudicators don't have a shared understanding of the scale. Before we can build tools to effectively collect scale-based human judgments, we need to first ensuring that there is a **consistent** understanding of the scale across human adjudicators. Traditionally, this consistency has been achieved through creating guidelines and training or selecting experts, however these approaches often don't scale when the task is complex or when the task has subjective components that we don't want to prescribe a criteria for. For example, in cases where the goal of the judgment task is to assess the opinions or preferences of communities or end users [13], training or selecting for "experts" does not make sense. Instead, for complex and nuanced tasks, we propose that, rather than try to achieve a perfect set of criteria, past cases that have already been judged can be used as a way to ground a scale that may otherwise be unfamiliar to adjudicators. In the later chapters, I will present how this idea of prior cases serving as anchors can manifest in various different task modalities from scalar ratings to categorical classification.

Enabling adjudicators to express uncertainty in initial judgments: Making sure that scales are consistent is the first step towards uncertainty tools that can scale to the new space of complex and nuanced tasks. However, even when scales are consistent, adjudicators will still face ambiguous individual cases. For example, take the task of rating items on a scale. Traditional tools often don't attempt to capture this ambiguity directly, instead asking adjudicators to give their best-effort precise answer and infer uncertainty from the resulting disagreement in the collected judgments [288]. However, measuring

uncertainty at this phase can already be too late—we no longer have the context around which the original adjudicator made their judgment, meaning that we need to make post-hoc assumptions about whether disagreeing judgments are due to adjudicators being uncertain about their answer and randomly choosing an acceptable judgment or if adjudicators simply do not agree with each other. On the other hand, some prior work has also recognized this limitation, and instead opt to have the adjudicators self-report uncertainty by estimating their own confidence [55]. While self-reported uncertainty is better than no information about uncertainty, asking annotators to assess their own level of confidence itself can be unreliable [157]. Therefore, we should make sure that tools for human judgments that not only enables adjudicators to express uncertainty, but also do so without imposing too much additional effort for estimating their own uncertainty. In later chapters, I will introduce how we created more expressive and intuitive interfaces for capturing uncertainty in various task modalities.

1.2.2 Addressing Uncertainty After Judgments: The Case of Disagreement

In subsection 1.1.1, we introduced why effectively capturing uncertainty in initial judgments can be important. But what about the uncertainty that we observe once the judgments are made? To arrive at a solution for working with uncertainty, we also need to consider how we can address uncertainty after a judgment. One common source of uncertainty in judgments results from the disagreement between adjudicators. The idea of deliberation has been shown by prior work to be potentially effective in resolving disagreement [237]. However, there are different ways to set up deliberation workflows which can affect how effective they are in practice. Like the case of initial judgments, I argue that 3 important aspects that can improve the effectiveness of deliberation: (1) creating shared understanding through context; (2) allowing for more expressivity during deliberation; and (3) training adjudicators on how to deliberate.

Just like how creating a shared understanding of scales is important for reducing uncertainty in initial judgments, having a shared context means that adjudicators have common ground to establish their arguments from. When tasks are defined through guidelines and

criteria, I argue that it's important for deliberation tools to ensure that the guidelines are consistently understood and followed. On the other hand, when tasks don't have clear guidelines, past judgments and precedents can instead serve as context to establish common ground.

Similar to the initial annotation, the ability for adjudicators to express complex arguments is also crucial for effective deliberation. Early crowd deliberation systems often focus on reducing task complexity at the cost of expressivity. For example, systems like MicroTalk [77] utilized a shortened form of deliberation by only having one round of a 'reconsider' workflow where justifications for the disagreement are shown and annotators re-evaluate their answer. However, one-shot justification workflows like this don't allow adjudicators enough expressivity to provide a well supported argument. Thus, I argue that like with judgments, deliberation workflows should provide ways for adjudicators to express their argument comprehensively as needed such as enabling .

Finally, I will also note that the process of deliberation is complex and deliberation that can constructively resolve disagreements also rely on effective arguments being made and understood [237]. To facilitate this, I argue that it can be important to provide training to layperson adjudicators before they engage in deliberation, especially focusing on teaching them how to assess others' arguments and form their own.

1.3 Meta-Processes for Crowd Uncertainty

Uncertainty can appear in multiple different settings for human judgment tools and we discussed several cases of this in subsection 1.1.1, ranging from errors to genuine disagreements. While we need new tools to capture and address new sources of uncertainty from increasingly challenging tasks, we should also recognize that individually the new tools we build won't be a one-size-fits all solution. Beyond individual tools and workflows outlined in the previous section, I argue that to fully support uncertainty in human judgments, what we ultimately need are meta-processes to coordinate how we collect human judgments and apply interventions for reducing uncertainty observed. Specifically, an **uncertainty-aware** meta-process for human judgments should involve the following general steps: (1) We need to make sure that tools for collecting judgments allow us to **capture** uncertainty of human

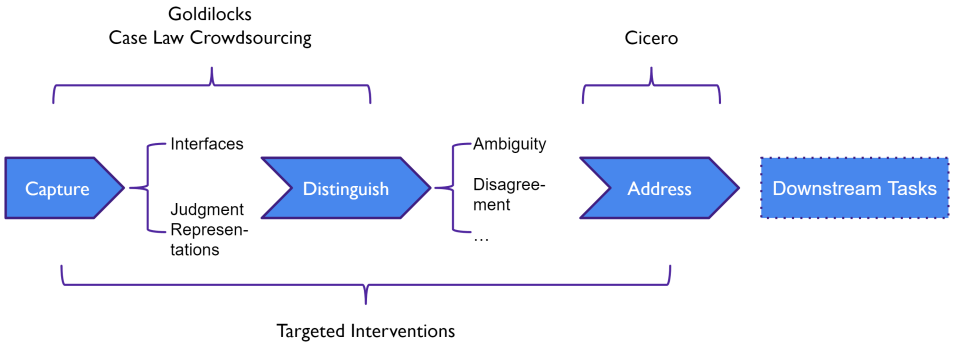


Figure 1.1: A diagram that illustrates meta-processes for crowd uncertainty and how various works presented in this thesis contribute to building such a workflow. These higher workflows inovlve both tools to capture and distinguish uncertainty as well as workflows to address uncertainty.

adjudicators during the initial judgment process; (2) We need to use these measurements to **distinguish** and quantify different sources of uncertainty; (3) This understanding of uncertainty should inform how we choose interventions to **address** sources of uncertainty in a targeted way; (4) Uncertainty should be **made available** to downstream consumers of collected judgments.

In subsection 1.2.1, we discussed how new designs for annotation tools can improve how we may **capture** uncertainty for some types of nuanced and subjective tasks. More generally, though, as we tackle new tasks and modalities of human judgments we need to also consider how to enable adjudicators to express and communicate their uncertainty as part of their judgments [274].

In subsection 1.1.2, we also discussed how different approaches to quantifying uncertainty can lead to insights about the potential actions that can be taken. As our understanding of uncertainty in various tasks evolve, meta-processes can take advantage of this understanding to make **distinctions** that attribute the captured uncertainty to its source.

On the intervention side, work by both us and others have explored interventions for uncertainty often times have concluded that effectiveness of interventions can depend on

a variety of factors [50, 237]. For example, one such finding was that deliberation was ineffective at resolving disagreements when there was uncertainty about cases in the form of ambiguity. A meta-process for collecting human judgments can use the quantified information about sources of uncertainty to decide what the best course of intervention is before **addressing** it—in this case, making sure that ambiguity is addressed before before attempting deliberation.

Finally, such meta-processes preserve information about uncertainty throughout the human judgment process. This means when data regarding human judgments is **made available** to downstream applications, such as to train machine learning models or build intelligent automation tools, consumers of the data are no longer required to make assumptions about why uncertainty is present. By keeping track of our understanding of uncertainty, new meta-processes for human judgments can provide the opportunity to improve the transparency around the data that grounds automated models [24, 262], and allow us to diagnose when potential issues like biases crop up.

1.4 Thesis Contributions

The focus of my dissertation is on how to design tools and crowd workflows so that groups and individuals can better understand and address uncertainty present in their judgments. I contribute research mainly around 3 aspects of this goal: improving the consistency in scale interpretations through the use of precedents and anchors, creating novel interfaces and representations of uncertainty to capture uncertainty at the point of judgment, and empirical observations about different characteristics of uncertainty for several scale-based judgment tasks. In this section we will give an overview of these contributions.

1.4.1 Improving Consistency through Precedents and Anchors

With the shift towards more subjective and nuanced scale-based judgment tasks, existing tools and workflows for collecting human judgments often do not offer sufficient means for groups of adjudicators to form a shared understanding of the various choices on the scale, leading to uncertainty in the form of inconsistent interpretations of the same scale choice. In this work, I demonstrate how we can improve the consistency of scale interpretation across

adjudicators through the use of prior annotated examples to ground understanding. In chapter 3, I focus on improving consistency in a continuous scalar rating annotation setting. I present the tool, Goldilocks, where we utilize past annotated cases as examples presented through both local and global anchors, to ground the inter-annotator understanding of otherwise abstractly defined scale levels. I demonstrate that the use of examples along with traditional text anchors leads to increased consistency across crowd annotators, with an especially prominent effect for tasks that involve more subjectivity. In chapter 4, I turn to focus on improving consistency of categorical judgments when decision bounds are unclear and difficult to specify through guidelines. I present the idea of case law crowdsourcing, where past annotated cases are made available to crowd adjudicators, who use them to assemble their judgment on new cases in the form of sets of precedent cases supporting their judgment. I demonstrate that the judgments made through precedent sets resulted in increased consistency on new cases. Finally, in chapter 5, I found that qualitatively when crowd workers engaged in synchronous deliberation, they also tended to invoke prior cases (in the form of discussions and chains of reasoning they have encountered) as means to support their arguments and build a consistent justification. Thus throughout this work, I make the case that the use of past judgments, in various forms across various judgment modalities, can be an effective way to improve consistency when tasks are complex or subjective and traditional approaches like guidelines are difficult to create (complex) or undesirable (subjective).

1.4.2 Novel Interfaces and Uncertainty Representations

Another limitation I identified with existing approaches in this work is that traditional tools to collect initial judgments often don't enable efficient ways for crowd adjudicators to indicate their uncertainty—either by ignoring uncertainty entirely or requiring adjudicators to estimate their own uncertainty with no assistance. I present solutions to this limitation in the form of novel interfaces and associated representations of judgments that improve the ability for human adjudicators to express their uncertainty. In chapter 3, I introduce the idea of utilizing ranges to establish intuitive representations of uncertainty in scalar

rating tasks. By using ranges to represent individual judgments, we can easily recover information about ambiguity and disagreement. Additionally, to support crowd annotators in creating ranges our system, Goldilocks, breaks down the rating creation into to steps, supporting the establishment of lower and upper bounds separately and assisting the creation of these bounds through enabling simple comparisons. In chapter 4, I introduce the idea of sets of precedents as a way to establish new judgments in categorical classification tasks. Representing judgments through sets of precedents allows us to capture both the intended judgment and the supporting justification of each adjudicator, and additionally allows us to diagnose situations when adjudicators struggled to make well-justified judgments in regions of the decision space that are under-supported by past cases. Our workflow, case law crowdsourcing, supports the creation of these precedent sets though a case exploration system that recommends past cases that may be relevant while allowing crowd annotators to adjust the criteria used to inform the recommendation. Through these tools and workflows, I make the case that new representations like ranges and precedent sets can provide more understanding of uncertainty to inform downstream interventions and that novel interfaces can support the efficient elicitation of judgments in these new representations.

1.4.3 Different Characteristics of Uncertainty in Scale-based Human Judgment Tasks

Finally, one of the contributions this work also makes is in building empirical understanding of the characteristics of uncertainty in several types of scale-based human judgment tasks, such as: word similarity, toxicity rating, satiety rating, age estimation, interpersonal conflict adjudication *etc.*. In chapter 3, I found that consistency in the interpretation of scales may be affected based on the characteristics of those scales. For judgment tasks that may be subjective and involve unfamiliar scales, such as rating satiety and toxicity, we found that annotators had lower consistency naturally, as opposed to tasks like age estimation where we expect annotators to be more familiar with the scale. This suggests that the consistency of scale interpretations can make a difference in terms of contributing to additional uncertainty. I also find that through the experiments in chapter 6, that when annotating datasets, the type of uncertainty involved in each case can significantly affect how effective

an uncertainty reduction intervention is, with a mis-targeted intervention failing to provide significant reduction in uncertainty as opposed to correctly targeted interventions which can. Additionally, I find that uniformly applying a single intervention regardless of what sources contribute to uncertainty for each instance can be detrimental to reducing overall uncertainty around judgments. Through these findings, I make the case that uncertainty can manifest in complex ways that vary depending on tasks, and processes that build better understanding around the uncertainty in each task are crucial if we want to effectively address uncertainty in human judgments.

1.5 Thesis Overview

In chapter 2, I will first position this work in the context of related research that looks into how uncertainty arises in human judgments, how harms can arise from not properly accounting for uncertainty when building upon human judgment data, and the systems and models already in place to address certain types of uncertainty.

Following this, from chapter 3 to chapter 5 I will introduce several novel designs for tools that aim at understanding and addressing uncertainty in various stages of the human judgment process. Starting with chapter 3, I first introduce a novel design for an annotation tool, **Goldilocks**, that allows capturing uncertainty around scalar rating judgments, while also improving on consistency. Following that, in chapter 4 I introduce the idea of **case law crowdsourcing** where prior decisions on a task – in the form of "precedent" cases – are used to form judgments allowing us to capture and understand uncertainty around categorical classification judgments while also improving on consistency of the judgments. In chapter 5, I present and evaluate **Cicero**, a workflow and tool that focuses on addressing the disagreement that contributes to uncertainty around group judgments on complex tasks by utilizing multi-turn and contextual discussions.

Finally, in chapter 6, I present an end-to-end meta-process that utilizes the insights into sources of uncertainty acquired through uncertainty aware annotation tools to dynamically select **targeted interventions** that address different sources of uncertainty directly.

At the end, I will discuss aspects of the broader implications of this work with respect to the larger ecosystem of human judgments in chapter 7, followed by a recap of the con-

tributions and an overview of ongoing and future work in chapter 8.

Content Warning: Some chapters in this thesis involve tasks that pertain to the annotation of sensitive user-generated content such as toxic and/or offensive online comments and posts or politically sensitive text.
We may include un-edited excerpts of this content as a part of the text or figures used to illustrate aspects of our tool design and/or findings. We will include additional warnings in the text before such content appears, or in the case of figures/tables, as a part of the caption.

Chapter 2
BACKGROUND AND RELATED WORK

Humans have had to contend with uncertainty for as long as we have had to make measurements and form judgments around them. As a result, there is an extensive history of work in the areas of statistics, cognitive psychology, and a variety of social, medical, and political domains, that look into the uncertainty surrounding judgments. Additionally, with the increasing prominence of computing systems and the increased dependence on collecting and creating data about the world, there is also a growing area of work around designing human-computer interfaces and processes to truthfully and accurately elicit and capture the uncertainty judgments from humans.

In this chapter, I will start by introducing some prior work that establishes approaches on how we think about uncertainty as a part of making judgments as humans, including the biases in our mental approaches to uncertainty. I will then look at the various ways that have been proposed by prior work to define, distinguish, and measure uncertainty both in general measurement and specifically for subjective human judgments. Following this, I will discuss the space of research into tools, systems, and processes that have been developed to collect human judgments and reduce uncertainty. Finally, I will examine some of the challenges and issues that are surfacing as a result of the increasing reliance on human judgment datasets collected from crowds, and how work in fields of AI and HCI have approached potential solutions.

2.1 Measuring and Distinguishing Sources of Uncertainty

Uncertainty has been an integral component of measurement for as long as we have had to make observations about the world. In the traditional setting of measurement and metrology, uncertainty is often characterized as a measure of the dispersion of the values that could be reasonably attributed to the property measured [150]. Commonly this may be repre-

sented through a statistical lens [60, 102] like standard deviations, confidence intervals, and in many cases through the statistical distribution itself. This view can be useful for understanding uncertainty that comes from our inability to achieve a deterministic understanding of the systems we are measuring [280].

However, while the statistical understanding of uncertainty provides a good way to generalize measurements across many domains, a drawback of this view of uncertainty – in the form of a single distribution – is that it fails to attribute why the uncertainty we observe arises in the first place [130]. This can be especially limiting when the ability to distinguish the source of uncertainty provides additional insight into how we may (or may not) address it. For example, suppose we were asked to characterize the height of a sampled group of individuals. One contributing factor of uncertainty lies in the natural variations across individuals in their height—the stochastic randomness in individual heights results in a limit on our ability to provide a single deterministic measurement for the group. On the other hand, the precision of the instrument (e.g., a tape measure) we use to collect each individual height in the first place, also results in limits on how well we can characterize the final measurement. Both of the above can result in what we observe as increased uncertainty if we were to characterize these measurements through a distribution over values. However, the distinction between the two sources of uncertainty provides additional insight if we want to address this uncertainty: While better instruments may allow us to reduce the latter type of uncertainty, we shouldn't expect to be able to reduce the former type of uncertainty as it reflects the natural variation in the property we are measuring. Building upon this idea, one view categorizes the sources of uncertainty based on : **aleatoric** (or aleatory), where uncertainty arises from the natural unpredictable variance in the property/phenomena measured; or **epistemic**, where uncertainty arises from a the limitations of our models, tools, and understanding [121]. Under this view of uncertainty, it follows that **epistemic** uncertainty can be targeted and reduced while **aleatoric** uncertainty should mainly be accommodated for.

The distinction of aleatoric and epistemic uncertainty doesn't always make for an insightful categorization of uncertainty [91]. This can be seen more prominently when subjective judgments are involved, where it can be difficult to tell whether subjective disagreements

reflect a kind of aleatoric uncertainty inherent to the subjective nature of the judgments or a kind of epistemic uncertainty resulting from the vague nature of a subjective question lacking a clear criteria. And indeed, beyond this separation, many other ways to understand and categorize sources of uncertainty have been explored and proposed. As one example, Soden *et al.* [250] examines how work in field of human-computer interaction (HCI) has engaged with uncertainty and notes that it can also be useful to understand sources of uncertainty through lenses like understanding the politics (distribution of power) that produce uncertainty, or viewing uncertainty as the evolution of how problems are understood and solutions envisioned. There is continuing work around how to categorize and quantify sources of uncertainty around data, each servicing different types of insights. Depending on the specific characteristics of the human judgments involved, it may be the case that different approaches to distinguish uncertainty should be selected. In this work, we don't seek to comprehensively review these possible categorization approaches as there are many different human judgment tasks and different desired uses for the judgments collected.

2.1.1 Uncertainty Around Human Judgments

In the example presented in the previous section, we touched on how uncertainty can manifest in traditional measurements of objective properties. However, in recent years, there has also been a growing demand for conducting measurements of properties based on human judgments in the form of data annotations on subjective or nuanced tasks. For example, datasets have been collected where human raters evaluated properties like toxicity [294], credibility [25, 192], or the degree which text is emotionally manipulative [128]. This increased utilization of human judgment has presented a new set of challenges regarding uncertainty.

The idea of human judgments as a measurement tool has been around since the inception of tools like Likert scales [201, 172]. However, unlike a fixed measurement tool which can conduct measurements consistently and at known levels of confidence, the uncertainty surrounding human judgments can vary based on the item being judged and the human adjudicators themselves [288]. Additionally, even expert human adjudicators often disagree

with each other [236] or even with oneself upon revisiting [106]. All of these aspects of human judgment contribute to the final uncertainty we observe in the collected data. In some cases, this uncertainty itself presents as a strength of crowdsourced human judgments, with systems increasingly recognizing uncertainty in these judgments may not be reducible and instead embracing its use in downstream tasks [144, 79, 80]. We have seen that diverse sets of human adjudicators can bring their different backgrounds, perspectives, and schools of thinking [270, 145] into the judgment they produce, contributing to an aggregated level of wisdom beyond each individual [6].

2.1.2 Annotator Positionality and Human Biases

However, along with the strengths of human judgments also comes limitations and pitfalls if they aren't applied correctly. While the uncertainty surrounding human judgments can certainly reflect the diversity of perspectives across groups of adjudicators, it's also important to recognize that data created from human judgments is also often affected by human cognitive biases [119] and the politics of those making the judgments [238]. Unfortunately, the rising demand for data in fields like machine learning has sometimes meant that limited effort goes in to understanding and addressing human biases in the datasets being used [230].

Cognitive biases have long been studied as a component of human judgment [119, 271, 196] and with the increasing use of crowdsourcing as a means to generate or annotate datasets, many of the same cognitive biases can crop up in the resulting data if unchecked during the design of the crowd tasks [83]. In the past, researchers have explored the effects of cognitive bias throughout various stages of crowdsourcing, from the recruitment of participants [14] to the instructions that crowdworkers use to complete their task [204]. In many cases, cognitive biases also mean that we can't reliably depend on human adjudicators to self-correct, with individuals often unable to accurately assess their own performance [157] and groups prone to fixate on certain suboptimal ideas [4]. Recent work on cognitive biases in crowdsourcing have presented checklists that assist task designers to think about potential cognitive biases in task design [78], though the same work has shown that cognitive biases remain under-addressed in practice today.

In addition to including cognitive biases, the datasets constructed from the decisions of humans will also reflect the positions of those involved [238, 69]. In recent years, we have seen this limitation increasingly acknowledged across various areas where the positionality of individuals or groups become embedded in various aspects of data. For example, the selection of what is included in datasets can encode biases of those making decisions and this can be seen across many domains from computer vision [70, 59] to work on fact checking [5] where prior work has shown that political affinity can affect how human adjudicators select posts to be checked. Similarly, annotator beliefs and identities also factor in to how they judge content, such as what is toxic [234] or what constitutes hate speech [232, 284]. This can be affected by various aspects of the identities of annotators [69], ranging from their cultural background [139], beliefs [69], and generally who they are as individuals [229].

2.2 Reducing Uncertainty

While it is important to understand what sources can contribute to uncertain human judgments, in many cases where human judgments are collected, simply knowing that the final answer was uncertain is often an unsatisfactory result. In fact, many sources of uncertainty are produced by aspects of the judgment process that we have control over—from how we recruit adjudicators to how we design and convey the problems to be judged.

In the following subsections, I will discuss how prior work has proposed approaches to reduce uncertainty through some factors that we do have control over. As most scale-based judgments today are collected through crowdsourced tasks—short microtasks completed by non-expert adjudicators, I will first start by looking at 3 aspects related to reducing uncertainty in a crowdsourcing setting: (1) reducing the effect of *errors* through **quality control**; (2) improving how tasks are conveyed to non-expert crowdworkers through **improving tasks and instructions**; and (3) resolving disagreements between crowd participants through **deliberation**. Then, I will discuss some prior work that looks at how uncertainty is reduced in other settings where human judgments are utilized, focusing on two scenarios: high stakes situations that involve expert judgments, and low stakes subjective judgments and opinions.

2.2.1 Automated Quality Control for Crowdsourcing

Because crowdsourcing systems and workflows rely on work done by non-expert crowd workers, quality control mechanisms have always been a central concern in designing these systems [87, 180]. Early crowdsourced annotation was presented as a way reduce the demand for expert annotators when a large amount of data was needed. The tasks that crowd annotation was applied to focused on those where ground truth—the answer that expert annotators would produce—could be established. Thus, automated quality control systems focused on reducing errors introduced by noisy or "low-quality" workers. The simplest approaches being voting mechanisms, such as majority vote [248] or later tournament voting [259], that were used to annotate natural language datasets. However, voting assumes that crowd workers would have similar performance characteristics and that the signal around correct answers is above the noise introduced by worker error (*i.e.*, that the majority answer is probably the right one). Subsequent work recognized that, even within a similarly recruited pool of crowd workers, the ability of each worker to arrive at the shared ground truth may differ [290, 287]. As a result, differences between workers were modeled through latent parameters, such as worker quality, which were learned through maximum likelihood estimates using processes like that proposed by Dawid and Skene [65]. Later automated quality control systems iterated upon this line of work utilizing more complex models that accounted for aspects like the particular type of annotation task and the difficulty of individual items in addition to parameters about worker quality [186, 180]. In addition to quality estimation, methods like Bayesian truth serum [215] allowed crowd task designers to create processes that did not penalize workers even when they disagreed with the majority.

However, due to the formulation of this problem as a matter of *quality control*, all of these methods assume that some correct answer exists for each case being judged. As more human judgment tasks have been deployed through crowdsourced platforms, we have found that this assumption rarely holds true in practice, especially across a sufficiently large set of cases where cases that cannot be decided are bound to appear. Instead, more modern views into quality control often focus on selecting for better crowd workers based on their performance

on problems where we know ground truth exists, utilizing continuous evaluation through gold-standard questions [118, 74, 200].

Undoubtedly, quality control is something that still needs to be considered for anyone deploying tasks on crowdsourcing platforms today [3]. This is also reflected in this work, where we also employ quality control mechanisms to select for honest participants. However, managing and mitigating error from crowd platforms is an artifact of the medium in which we explore crowd uncertainty, and thus these approaches will not be a focus of this work.

2.2.2 Instructions and Training

With the diminishing returns provided by the noise view into judgment uncertainty, the design of the task itself became an new focus as an important factor that affected the quality of the annotations. Prior work has found that, when tasks were poorly designed or under-specified, workers often became confused about what was requested and produced additional errors [293]. Rubrics have been proposed as a solution to this problem by providing a shared set of clear actionable instructions that is available to all workers [298]. Indeed, when implemented well, rubrics provide an easy way to calibrate and unify how different workers approach a problem (or in the case of our focus, a scale-based human judgment task). However, comprehensive rubrics can be very difficult to create in many situations, especially when the judgment tasks are complex and difficult to specify [166]. Within the area of crowdsourcing, prior work has explored avenues to reduce the effort required to make good instructions. Some have proposed that crowd annotators can be consulted to suggest areas for improvement [35], while others have proposed using the amount of uncertainty observed after annotation to discover potentially under-specified edge cases [213]

Alongside the improvement on instructions, training of workers also became an important aspect of improving the quality (and thus reducing the uncertainty) around answers. Methods such as gated instructions [178] have been proposed which combine quality control mechanisms (in the form of gating) with training procedures to calibrate worker understanding. Training is also not necessarily limited to being a procedure conducted at the start of a crowd task. For example, Dow *et al.* [76] conducted experiments demonstrating

that timely, task-specific feedback during the task helps crowd workers learn, persevere, and produce better results. Even quality control mechanisms can be repurposed to provide training as well, such as exposing answers to gold-standard test questions after they have been used for quality control [118].

Sometimes training doesn't all need to be provided by the task requester. Beyond individual training, feedback from peers can also help train workers: Ho *et al.* [120] show that peer communication improves work quality while Kobayashi *et al.* [154] demonstrate that reviewing can help workers self-correct. Additionally, Zhu *et al.* [310] noted that workers who review others' work perform better on subsequent tasks.

Instructions and training can both be effective ways to reduce uncertainty in human judgment settings too by improving the consistency of task understanding across crowd adjudicators. However, in many cases, creating high quality instructions and training requires significant effort and expertise [35]. Additionally, any issues in instructions and training can easily manifest as problems in the data produced at the end [99, 204], meaning that clarifying tasks via complex instructions and training may not be possible for many subjective or nuanced tasks. In our work, we will look at alternatives to traditional instructions and training that allow us to preserve nuance and subjectivity in human judgment tasks.

2.2.3 Surfacing and Resolving Disagreement

Not all uncertainty stems from confusion about the the tasks themselves. In some cases, even when there is little ambiguity in the cases being judged, crowd adjudicators may still disagree on what the answer should be. This disagreement can be the result of different ways crowd adjudicators reason about the problem [270]. However, without more understanding of the disagreement, there is little hope of being able to resolve it. Researchers have demonstrated that having annotators to submit "rationales" along with their answers, such as through highlighting portions of text [300, 187] or an image [73], and training classifiers using these rationales can result in better classification performance. This suggests that the context and reasoning encoded through rationales contains additional information about why disagreement occurs and that if utilized properly may also lead to ways to resolve

disagreement.

Extending from the idea of rationales for training models, Drapeau *et al.* [77] present the idea of using peer rationales to surface and resolve disagreements during the annotation process itself (as opposed to post-hoc). By acquiring rationales in the form of justifications and then presenting them to peer annotators when disagreements are encountered, peers may be convinced to change their answer if they see a rationale that addresses their disagreement. More broadly, Wiebe *et al.* [292] have also shown that in small group in-person settings, getting annotators to reconsider their positions and discuss them with other workers can be beneficial to the quality of annotations. ConsiderIt [156] applies similar ideas to political judgments, using the creation of pro/con lists to encourage participants to think about alternative positions, and potentially leading to fewer direct disagreements.

On the other hand, a different line of work poses that disagreement perhaps should not be resolved by the annotators directly. In structured labeling [158], researchers propose a workflow where annotators can encode disagreements about classification rationale by creating evolving taxonomies organized by potential uncertainties in criteria. In the follow up work by Chang *et al.* [46], this approach is further developed to enable groups of annotators to build shared taxonomies around rationale-based clusters that encode but not resolve disagreements. With this approach, the responsibility to resolve disagreement is instead passed on to a task requester, who acts as the ultimate source of truth in removing the uncertainty in disagreements. While the idea of a final dictator can be desirable for some classification tasks, it can also be inadequate if the requester in question is not the authority that should be making the final judgment.

In cases where it is undesirable to have a requester dictate how disagreement is resolved, we look back to rationales for a solution. While softer approaches like reflecting on peer rationales [77] can potentially reduce disagreements, they aren't effective against all types of disagreement. Rationales are limited in how much context they can provide, and when the source of disagreement is not covered in the rationale, it may not help prevent disagreement at all. To solve this, there has been work on explicitly using full deliberation processes to actively resolve disagreements when they are observed [237, 50]. These works have shown that for complex but unambiguous tasks, a full deliberation process can be effective

at reducing uncertainty by resolving disagreements explicitly. In our work, we explore deliberation as one way of reducing uncertainty in human judgments.

2.2.4 Managing and Reducing Uncertainty in High-Stakes Judgments

Of course, processes for performing judgments under uncertainty exist in many fields beyond the space of crowdsourcing annotation [243, 235, 254]. Among these, the task of managing and reducing uncertainty can be especially important when judgments need to be made under high-stakes situations among experts. To account for this, processes have also correspondingly been developed in these areas fields to manage and reduce uncertainty.

As one example, within the realm of the judicial system (in this example specifically focusing on that of the United States), uncertainty can be present in many forms throughout the judgment process with many mechanisms to address it. In order for a jury to make an informed decision, they need to consider the evidence that contextualizes a case, which can often come with uncertainty both resulting from the ambiguity of the evidence itself and the positionality of the opposing sides presenting the evidence. To reduce this processes around discovery and presentation have been formulated to address the introduction of uncertainty around context [175]. Additionally, as the final judgment is produced by a group of individual jurors, uncertainty can arise from disagreements in interpretations of the evidence, with deliberation used as a means for the jury to make sense of what has been presented and the corresponding implications [263]. These processes can be very labor and attention intensive, often spanning multiple days of concentrated work. Of course, judicial systems, even with their complex processes for managing uncertainty, are still not a perfect solution to the problem of reducing uncertainty. For example, juries also suffer from human biases during their deliberation [43] and it can be difficult to keep track of details in the deliberation process [267]. Additionally, some have argued that institutions like the supreme court can act as final adjudicator, resolving disagreements through closed processes [249]

Similar to the legal space, in the medical domain where doctors need to make diagnosis and treatment judgments under uncertainty, deliberation is often also used as a tool to resolve uncertainty in the form of disagreements in interpretation [236]. Though, unlike

social settings, the medical domain also involves diagnostic testing, which come with more traditional metrological guarantees about the uncertainty around outputs [132].

Many of the strategies used by other fields have provided inspiration for the tools like deliberation processes that presented in this work.

2.2.5 Working with Uncertainty in Subjective Judgments

While crowdsourcing is often used to collect subjective judgments today, methods for working with uncertainty in subjective judgments has existed before their application in crowdsourced settings [201, 268, 42]. In fact many of the mechanisms used in crowdsourcing are concepts borrowed from classical situations of working with uncertainty in human judgments.

Survey mechanisms for measuring subjective responses, such has Likert scales [172], were introduced as ways to understand subjective judgments. Along with them, methods have been produced to mitigate uncertainty of these responses. More generally, attention check questions have been proposed for surveys as quality control mechanisms [242]. On the other hand, methods such as the calibrated sigma method [285] have also been proposed to remedy the uncertainty resulting from uncalibrated results when comparing between-group Likert responses. Additionally, extensions of the Bayesian truth serum into subjective tasks [215, 214] has also improved the reliability of responses and measurements of uncertainty surrounding them by removing incentives that bias decisions towards consensus, which if not accounted for may mask the presence of uncertainty.

On the theoretical side, the Dempster-Schafer theory [240] and others like fuzzy logic [299] provided frameworks for people to conduct reasoning about uncertainty without resorting to concrete probabilistic estimates. All of these approaches served as inspiration for our work, where we address judgments involving subjectivity which can be easily affected by biases.

2.3 Uncertainty in Downstream Tasks

Human judgments are used in many different downstream applications and as a result the uncertainty around human judgment data creates implications for how this data can be

used in downstream tasks. In this section, I will first discuss the challenges of uncertainty in downstream tasks, focusing on the topics of **evaluation** of AI systems and creating data for **training** systems on subjective judgment tasks. Then, I will discuss some solutions that have been proposed to manage uncertainty in downstream tasks, both **automated** solutions that focus on creating better models and **human-in-the-loop** solutions that involve collaboration between automated systems and humans under uncertainty.

2.3.1 Human Evaluation of AI Systems

The evaluation of generative AI systems has become an increasingly important challenge in recent years. Compared to AI systems that select an answer among a set of choices, generative systems produce free-form output that can't be easily verified against a reference answer. Traditional approaches of evaluating this kind of generative AI system output often focused their effort on enabling rapid evaluation at low cost. As a result, many metrics have been proposed to use human-annotated data to construct benchmarks that are then evaluated by machines (*e.g.*, BLEU, METEOR [40, 2, 68]) through measures of similarity. However, as the performance of models has increased, the higher quality of the output of models often means that automated evaluation approaches can struggle to tell apart similarly high performance models, with some automated approaches penalizing models that produced output that would otherwise be acceptable by a human but did not contain components of the reference answer. As a result, human rating has become an increasingly oft used tool to evaluate performance of models on generative tasks [312] (*e.g.*, summarization [85], translation [171]) in fields like natural language processing. Humans can more easily judge characteristics like fluency, relevance, and conciseness, which cannot be easily and reliably assessed with automated metrics [108]. Human rating has also been used to evaluate the output of chatbots [239, 170], to judge search results [129], or assess clustering quality [301].

Increasingly, human ratings (both comparative and absolute) are becoming an integral aspect in facilitating comparisons between models through evaluation leaderboards and shared tasks [252, 148], where consistency and robustness of comparative results are cru-

cial. However, we also see that human raters can often disagree on these evaluations [210] and the type of evaluation used can affect results too [252] as seen in shared task evaluations switching from relative rating to absolute Likert scales. This has led to traditional evaluations sometimes presenting over-estimates of actual performance that are only observed after within-annotator uncertainty is accounted for [106]. Therefore, creating benchmarks that can account for the various sources of uncertainty as well as adapt to the ever-changing set of tasks and systems is important if we want to accurately understand the performance of AI systems.

2.3.2 Ground Truth from Subjective Tasks

Another important application of human judgment is to establish a source of 'ground truth' as a way to define tasks. For example, human judgments have been used as a way to build knowledge bases of human common sense reasoning [233], to model how humans make judgments on morality [140], to simulate toxic situations [98], or to define concepts like natural language inference [207] and word senses [144]. In all these cases, it's important to understand and document the uncertainty surrounding the human judgments as they are ultimately the foundation of the task definition, rather than necessarily what was intended by those collecting the data [90]. Recent work has shed light into how insufficient understanding of the uncertainty in these datasets can cause significant issues with downstream models trained on them [99, 230]. Along these lines, some work has also called out the need to document aspects of these datasets [97, 18, 211] including providing dis-aggregated data. On the model side, recent work has also explored how data with uncertainty may be used to train more robust models [216, 89] that can respond identify when no certain answer is possible [218]. Thus capturing and presenting the uncertainty around both individual judgments and group collective judgments can be crucial for building downstream AI systems.

In other areas, the use of crowd judgments for establishing 'ground truth' can be useful as a way of creating more legitimate and transparent processes used to define otherwise subjective standards. For example, prior work in online community moderation has shown

that the use of digital juries to make judgments on content and moderation actions [86, 202, 123] can provide insight into the community's values – both that of individual jurors and any disagreements across the jury group. Indeed, on the data side, some have proposed that the presentation of uncertainty itself can be a way of increasing transparency [24, 13] into how judgments on these traditionally subjective tasks are made.

2.3.3 Incorporating Uncertainty in Machine Learning

With the increasing recognition of problems caused by mistreatment of uncertainty in datasets [262], recent work on the modeling side has proposed that models should be trained with the uncertainty in the original judgments in mind. One area of work explores the idea of training models by taking into account the uncertainty around labels rather than simply the aggregated answer, such as in the case of image classification [297] or ranking tasks [296]. Similarly, encoding human judgment uncertainty through soft labels (which represent probabilistic distributions over the label space) has also been proposed as alternatives for training [89, 295], with some recent proposed work that utilizes uncertainty information through soft labels on a per-annotator level [57]. Some have also proposed that models should be trained directly on dis-aggregated data, treating learning from different annotators as separate learning sub-tasks [63, 211]. Others have also shown that label-level disagreement information for labels can be beneficial to training deep learning models [282], with uncertainty information serving to improve robustness of models under adversarial conditions [216].

The idea of utilizing dis-aggregated training also has implications in the space of ensuring fairness of the resulting models. One instance of this idea is reflected in the recently proposed concept of jury learning, where demographic information surrounding annotators can be used to reconfigure the composition of what judgments are used to inform machine learning models [105]. Reconfiguring how uncertain judgments are used ties into a broader line of work of value-sensitive algorithm design [311] seeking to improve the state of the world rather than continue to perpetuate historic biases.

In our work, we also seek to tie in to this body of work in downstream applications by

making available not just measurements of uncertainty, but also richer insights into what contributes to uncertain human judgments.

2.3.4 Communicating Uncertainty to Facilitate Human-AI Collaboration

While building standalone models that help automate traditional human judgments can certainly be important, in many cases AI systems are often used to support human decision-making processes via human-in-the-loop (HITL) processes [109, 264]. Work in human-AI collaborative systems have shown that simply improving the performance of the AI often is not sufficient for such collaborations to be effective—in order for AI systems to be effective partners that support humans, they need to account for the mental model of human partners [16]. Existing work around explainability mechanisms has introduced various mechanisms for AI models to present information about their proposed judgments, often including aspects that visualize and communicate uncertainty, such as confidence scores [117]. However, in many cases the systems that are being built today still fail to achieve real collaboration in the form of complementary performance [17]. It is often the case that information about the AI's uncertainty presented through traditional metrics such as confidence scores doesn't provide much benefit for human decision makers, who struggle to interpret what this uncertainty means. Only recently has it been demonstrated that complementary performance is indeed possible, but only in cases where the cost for verification is much lower than that of completing the task altogether [275].

In collaboration between humans, we often utilize information conveyed about uncertainty to build mental models of when we can rely on our partners. However, the current space of tools used to communicate uncertainty of AI models is much more limited, leading to challenges in building trust and reliance on AI [84]. This increased prominence of hybrid human-AI collaborative processes has led to increased attention on how models represent and consequently communicate uncertainty to human decision-makers, with works in visualization exploring how to most effectively communicate uncertainty to human collaborators [110]. Yet, because existing models are still built to model uncertainty through probabilistic metrics like confidence scores and distributions, uncertainty information is of-

ten still difficult to convey concretely to human partners in a way that can positively improve collaboration outcomes [307]. In my work, we look into some alternative representations of judgments that also encode uncertainty, which may provide a new avenue for thinking about AI systems that produce outputs that also natively encode uncertainty in an interpretable way.

As we have seen from the instances in the sections above, there is a significant and increasing demand for ways to better capture uncertainty and learn uncertainty from downstream applications that utilize human judgment data. In our work, we explore how better tooling around human judgment processes can allow us to capture and maintain understanding around the sources of uncertainty throughout the human judgment process, making this information available to downstream applications. We also propose new representations of judgments that encode uncertainty in ways intuitive to human annotators, which may prove useful for bridging the communication gap between human-AI collaborative decision-making.

Chapter 3

GOLDILOCKS: UNCERTAINTY IN SCALAR RATING JUDGMENTS

In this chapter, we present Goldilocks, a novel crowd rating elicitation technique for collecting calibrated scalar annotations that also distinguishes inherent ambiguity from inter-annotator disagreement. We introduce two main ideas: grounding absolute rating scales with examples and using a two-step bounding process to establish a range for an item's placement. We test our designs in three domains: judging toxicity of online comments, estimating satiety of food depicted in images, and estimating age based on portraits. We show that (1) Goldilocks can improve consistency in domains where interpretation of the scale is not universal, and that (2) representing items with ranges lets us simultaneously capture different sources of uncertainty leading to better estimates of pairwise relationship distributions.

3.1 Introduction

Much of modern machine learning is built on the foundation of human-annotated data. As the application of these models has expanded into more socially embedded and contextually nuanced domains [9, 192, 198], collecting high quality, consistent, and robust data from human annotators has become an increasingly important yet challenging task [25]. As one example, the ability to gather human evaluations of the toxicity of a piece of text is a necessary precursor to being able to build toxicity models to support online communities [294] as well as capture and mitigate harmful outputs generated by large language models [98].

However, traditional rating methods commonly used today, like absolute or comparative rating, can produce inconsistencies in ratings across annotators and even with a single annotator's ratings [10, 229]. This is due to issues such as lack of a common interpretation of the scale in the case of absolute rating, as well as lack of global context in the case of comparative rating [285, 288, 56]. Additionally, while current rating methods can capture

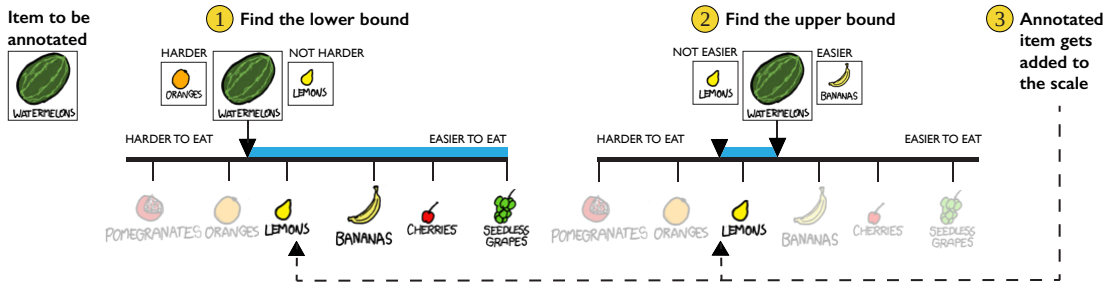


Figure 3.1: The Goldilocks annotation process involves placing items onto a continuous scale that is populated with items that have previously been annotated. The process is broken down into three parts. (1) Find the lower bound by moving the left handle of the slider towards the right and away from its initial position on the far left of the scale. Continue until encountering an item on the scale that is either greater than or indistinguishable from the item to be annotated. (2) Find the upper bound in the same way but moving the right handle towards the left. Continue until encountering an item on the scale that is less than or indistinguishable from the item to be annotated or until the two handles are on top of each other, representing complete certainty. (3) Finally, the lower and upper bounds of the item get added to the scale to join the existing items. Thus, an annotator will be able to see and compare against their own prior annotated items as they annotate more items. Images of fruit are taken from XKCD: `https://xkcd.com/388/`

uncertainty in the ratings, it is difficult to dissect whether the uncertainty is a result of inherent ambiguity in the item—where certain items cannot be confidently distinguished from each other [80]—or from disagreement between annotators on where the item should be placed. Distinguishing these sources of uncertainty offers the potential of better capturing biases between annotators. It also allows us to develop more calibrated models that only make high-confidence distinctions between items when a human would have as well [113].

In this paper, we propose a new design for collecting scalar annotations called Goldilocks[1] that combines the ability to make direct comparisons between items with the simplicity of

[1]Somewhat like Goldilocks in "Goldilocks and the Three Bears", annotators must make use of *multiple* comparisons.

a continuous absolute rating scale (Figure 3.1). To accomplish this, Goldilocks introduces two main ideas—(1) *Calibration using Prior Annotations*: we provide previously annotated items as anchors to ground interpretations of the scale both within and across annotators. (2) *Item-level Resolution Elicitation using Ranges*: we use a two-step process to collect lower and upper bounds for each item instead of a single placement. Goldilocks combines strengths from both absolute and comparative ratings as annotators make multiple comparative judgments while placing an item on an absolute scale. In addition, by directly eliciting an annotator's own judgment of an item's inherent ambiguity instead of relying on aggregating inter-annotator agreement, Goldilocks can separate agreement from perceived ambiguity.

To understand the effectiveness of these designs, we conducted three studies comparing aspects of the Goldilocks annotation process against traditional methods. In the first experiment, we evaluated whether anchoring scales with a shared set of previously annotated items can improve consistency of item placement across annotators. In the second experiment, we examined whether including an annotator's own prior annotations as anchors improves self-consistency. Our final experiment evaluated how well ranges captured using Goldilocks can recover the distribution of pairwise relationships as measured by traditional absolute and comparative rating. Each of our experiments were conducted in three domains representative of the subjective or ambiguous rating tasks that can be challenging for traditional methods: judging TOXICITY of online comments (short text), estimating SATIETY of food depicted in images (visual), and estimating AGE from portrait photos (visual).

From the experiments examining anchors, we found that the addition of shared example anchors to ground rating scales improves rating consistency between annotators in domains where shared understanding of the scale is low. We also found indications that showing one's prior annotations in a session as additional anchors may improve self-consistency on examples where there is high initial uncertainty. From the experiment examining ranges, we found that our two-step range annotation process allows us to infer pairwise relationship distributions that are more robust—simultaneously reflecting both uncertainty of single annotators and disagreement between annotators—compared to alternatives with a single value. Finally, we found that the size of range annotations provides an interpretation of

uncertainty that is distinct from the uncertainty modeled via inter-annotator disagreement.

We conclude with a discussion of the limitations and opportunities for Goldilocks. Regarding efficiency, while our approach is more costly than performing just one of absolute or comparative rating, our method is cheaper than performing both, which would be necessary to recover the richer data that Goldilocks generates. We discuss cases where a deeper understanding of uncertainty can be important for generating more trustworthy model predictions. We also discuss what we envision as a scaled-up Goldilocks workflow: utilizing iterative improvement through multiple annotation sessions with designs for bootstrapping the initial set of anchors along with interesting problems to be explored in each of these aspects.

3.2 Related Work

In this section we review prior work on absolute and comparative rating designs.

One of the most common designs for collecting human ratings today is through *absolute rating* scales, often in the form of Likert or semantic differential scales [172, 201]. When a consistent interpretation of the scale can be established across annotators, designs based on absolute rating can offer many benefits such as being very efficient (only requiring a single annotation per item) and providing easily interpretable ratings that are globally contextualized (rather than depending on other items). However, many annotation domains do not have commonly accepted scales, meaning that divergent interpretations of a scale based on abstract text descriptions can become a source of disagreement and inconsistency across annotators [285]. Even within an annotator's own annotations, the lack of a well defined scale means that to maintain consistent ratings, they must refer to their own memory of their past decisions which can be unreliable [36]. Accounting for these inconsistencies requires additional effort—either through additional calibration [96] or just identifying and reporting them [100]. Absolute scales can also be locally unreliable [288]—because items are only ever compared against the scale's anchors, pairwise comparisons between two items with similar values can only be rigorously done if the measurement resolution (uncertainty around the values) is also accounted for.

As many consistency problems in absolute rating systems result from the lack of direct

comparisons between actual items, a natural solution is to look towards the other major alternative—*comparative ratings* [268]. In comparative rating systems, items are compared against against one another directly, circumventing the need for a scale as a proxy and providing highly reliable measurements of local relationships. This kind of comparison can also be more intuitive for annotators leading to comparative systems sometimes suggested as a more accurate alternative for ranking items [149, 170]. However, collecting comparative ratings can be considerably more costly (on the order of N comparisons per item) unless sampling and ranking aggregation methods or partial comparisons, which trade off additional uncertainty, are used [142, 149]. The focus on local comparisons makes it easy for an annotator to inadvertently produce annotations that are not globally self-consistent, requiring post-hoc corrective action that may not reflect an annotator's actual judgment. Abandoning global context also means that if a rating score (rather than ranking) is desired, a numeric mapping like Elo rating needs to be done [56], which often come with assumptions about uniform spacing between items.

Past work has explored hybrid approaches that combine aspects of comparative and absolute annotation. For example, Sakaguchi et al. [227] present EASL, a hybrid approach where items are rated using continuous absolute scales but similar items are grouped together for annotation allowing for some degree of comparison and contextualization. While similar in motivation, our work differs in that we make comparison an integral part of the annotation process rather than an optional source of context, allowing us to provide more consistency by grounding comparison against global anchors and capture uncertainty intuitively by using comparisons to establish bounds.

Beyond the individual drawbacks mentioned above, neither of the two traditional annotation methods supports effective separation of the sources of uncertainty as a part of the the annotation process [130]. These sources include both aleatoric uncertainty, or irreducible ambiguity inherent to the item being rated, and epistemic uncertainty, or disagreement on the placement of the item. Absolute rating forces annotators to resolve inherent ambiguity into a precise placement causing both sources of uncertainty to be mixed. Meanwhile, comparative rating only provides an indirect view into inherent ambiguity through the size of equivalence sets. Separating the two sources of uncertainty is especially desirable as it can

be an important tool for understanding properties of the items being annotated separate from biases or divergent interpretations among annotators.

3.3 Design

Absolute rating can suffer from inconsistent scale interpretations while comparative rating lacks global context. Our design for the Goldilocks annotation system takes a hybrid approach, with the specific goals of: (1) improving consistency (between annotators and over time within annotator), and (2) enabling intuitive indication of uncertainty with respect to the scale for each example being labeled.

In this section, we will describe the designs that address each of the goals above followed by additional aspects of operating the complete annotation workflow. At the end, we will discuss specific details of the design decisions we made for our implementation separate from the overall design of the Goldilocks annotation process.

3.3.1 Grounding with Prior Examples

We base the main interactions in Goldilocks around an absolute rating design. To mitigate the aforementioned drawbacks of absolute rating, Goldilocks uses prior examples in addition to abstract descriptions to ground the scale, making it possible to make pairwise comparisons while still using absolute rating interactions. Prior work has shown that human judgments measured explicitly with comparisons can be easier than direct labels for some tasks [245, 308, 279], and *fixed* reference anchors have been used in other procedures to provide a more concrete grounding of scales [277]. Similar ideas that use comparisons against samples to contextualize abstract scales also exist in other fields like cognitive psychology [256].

Goldilocks uses a set of previously-annotated examples to add two additional pieces of information to the absolute rating scale—**global grounding** and **local comparisons**, as shown in Figure 3.2. With **global grounding**, a small set of representative examples are selected and placed as anchors along the rating scale, similar to existing text-based anchors for levels in traditional absolute rating. Using concrete examples allows annotators to quickly understand and estimate where each item could fit on the scale. Since there can be many previously-annotated examples, we make sure to only visualize a smaller subset

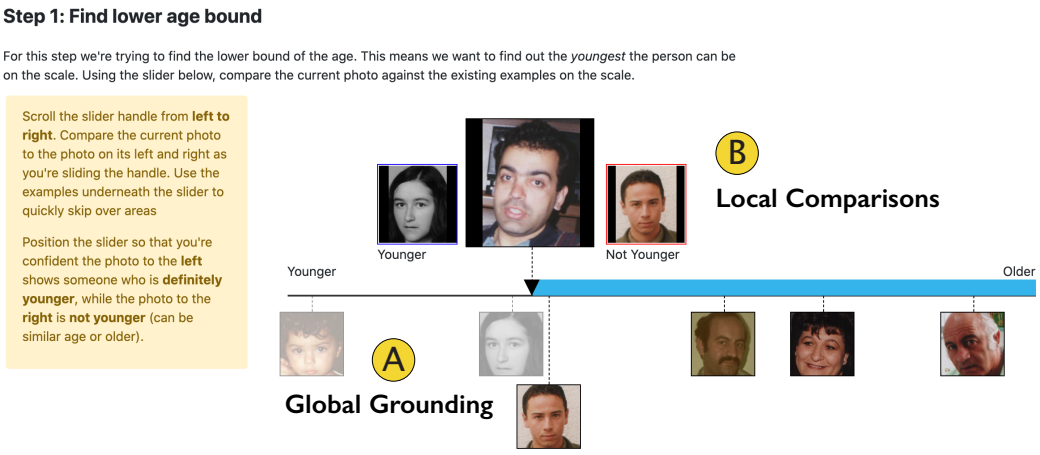


Figure 3.2: A screenshot showing the comparisons that annotators can make while placing the upper or lower bound of an item on the scale in the Goldilocks annotation process. To support grounding with examples, Goldilocks provides: (A) global grounding by selecting 5–7 previously annotated items that are maximally spread out on the scale and placing them as anchors to support coarse and fast global adjustment. (B) Local comparisons of previously annotated items directly to the left and right of the slider handle are shown as an annotator scrubs the handle across the slider. Local items that are not one of the global examples are inserted as anchors. Together, this allows annotators to make fine-grained local adjustments.

of examples (around 5 to 7, similar to typical numbers of Likert levels) that are maximally spread out along the scale. In practice, there are many ways to select these examples. The specific selection process we used is outlined in 3.3.4.

While global grounding is useful for making coarse placements, it alone is insufficient for narrowing down specific placement of items. To help the annotators find specific placements, Goldilocks also surfaces **local comparisons** by showing the immediate neighborhood above and below a position on the scale. As annotators scrub along a continuous scale, we show side-by-side comparisons between the current indicated position and the closest items above and below this position. Placements of these neighbors are also indicated on the scale itself, allowing for annotators to adjust proportional distance to each neighbor based on their evaluation of the item being placed. These designs together allow for a more consistent and concrete instantiation of the scale across multiple annotators.

Finally, Goldilocks addresses local self-consistency by supporting dynamic augmentation of the anchor examples used to ground the rating scale: as annotators progress in an annotation session, their own annotations for earlier items are also incorporated into the set of references alongside any pre-seeded ones (Step 3 of Figure 3.1). These personal annotations will then also take part in both global grounding and local comparisons, making it possible to directly compare new items against past annotations produced in the same session.

One potential limitation for any annotation process involving examples is how to start the annotation when no past examples are available. Goldilocks accounts for this with a separate procedure to curate an initial seed set that is deployed when past examples do not exist. We will dive into more detail about the selection of this initial seed set of items to jumpstart annotation in Section 3.3.3. In the discussion section, we will also discuss avenues of addressing other challenges in example-based grounding such as scaling up annotation with iterative improvement and addressing density as the scale becomes populated with more annotated examples.

3.3.2 Two-Step Range Annotation

Not all items can be meaningfully distinguished from all other items by an annotator. Instead of forcing the breaking of ties, most designs for side-by-side comparisons allow annotators to indicate "indistinguishable" or "tied" pairs [165]—however, there is no such elicitation process for traditional absolute rating designs. With Goldilocks, we propose a new process that allows annotators to indicate "indistinguishable" pairwise relationships on an absolute rating scale. To achieve this, we take inspiration from prior work [81], where annotators were asked to select *all* potentially relevant labels for an item instead of a single best label option. We extend this into the continuous scale domain by introducing the concept of eliciting "range" labels—where upper and lower bounds establish a subsection of the scale representing where an item *could* be placed. Our range-based approach is also reminiscent of methods like best-worst scaling [149] in comparative rating, which can efficiently capture pairwise relationships across many items.

Prior designs have explored alternatives to eliciting uncertainty for scalar annotations, such as in the form of a weighted distribution across *surrounding* anchor labels [55]. However, estimating distributions in this way can be challenging for humans, as an annotator has little guidance on how to allocate weight to the anchoring labels they find reasonable. In Goldilocks, we can take advantage of the comparisons afforded by grounding examples to contextualize distributions intuitively. Specifically, we break down the process of eliciting ranges into two steps: finding the lower bound and then finding the upper bound (Steps 1 and 2 in Figure 3.1). In the first step, an annotator can utilize the past example anchors to quickly search for where to place the lower bound of an item using comparisons to work up the scale and finding the position where they can no longer confidently decide that the closest reference should be lower on the scale than the annotated item. Similarly, in the second step, an annotator establishes the upper bound working down from the scale and stopping when they can no longer identify a reference item as higher than the annotated item.

Positions of anchor items on the scale are themselves internally represented by ranges. During each step, the anchors are visualized using the corresponding opposing bound: when

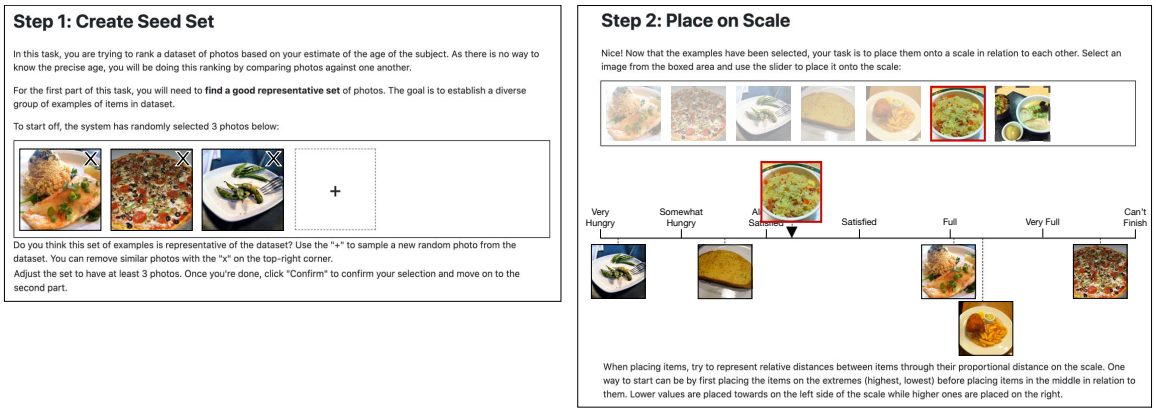


Figure 3.3: Screenshots illustrating the two steps in the cold start process for Goldilocks. Step 1 (Left): A seed set can be created by using the cold start interface to randomly draw examples and drop existing ones to create an adequately sized representative set of examples. Step 2 (Right): The items from the seed set are placed onto a scale by adjusting their position relative to each other, forming the initial values that can be used to bootstrap the annotation tasks in Goldilocks. These initial items can later be reintroduced in the Goldilocks annotation process once other items have been annotated, in order to collect ranges.

finding the *lower* bound for an item, anchor items are placed on the scale according to their *upper* bound values and vice versa for the upper bound (shown in Figure 3.1). This two-step process allows an annotator to easily establish a range that is intuitive and meaningful—it represents the range where the annotator is no longer able to confidently distinguish items.

3.3.3 Cold Start Process

Annotation of any item in the Goldilocks process requires there to be previously-annotated items using the same scale in order to populate the grounding examples and comparisons. However, if prior annotations do not exist yet, they must be created in what we call the cold start process.

The cold start process (shown in Figure 3.3) consists of two steps—representative example selection and placement on a scale. In the example selection step, Goldilocks draws

a certain amount of un-annotated examples randomly from the set of data to be annotated. An annotator can then adjust this set by requesting to draw additional random examples or dropping existing examples. The goal is to adjust this set to be more representative such that there are at least a certain number of examples in the set (defined based on task) and that the examples are maximally different from each other. A similar sample and replace approach was used in Alloy [48] to bootstrap good seed sets for clustering. In the placement step, the annotator successively places all the examples onto an absolute rating scale by comparing them against each other, with the ability to adjust the position of any item on the scale. The scale can be blank at the outset or be initialized with text anchors as shown in Figure 3.3.

The cold start process can be completed with recruited annotators, where the resulting placements are aggregated across them to create the set of seed examples that become the first set of Goldilocks example anchors. Alternatively, the cold start process can be completed by the task designer or by domain experts, making it a way for requesters to specify a scale without having to design a set of training instances. In this case, the steps in the cold start process are used to assist the exploration of the dataset. Once additional items have been annotated using Goldilocks, the set of anchor examples can be augmented with this newly annotated data. If desired, the initial seed examples can be re-annotated by removing them from the scale and re-introducing them as new items to be placed in an iterative improvement fashion.

3.3.4 Implementation Details

We outline specific details about our implementation of Goldilocks that we use for experiments. We implemented Goldilocks based on a custom slider component using JavaScript, HTML, and CSS. Global grounding examples were incorporated as part of the scale via fixed anchor tick markers below the scale. Examples were then rendered in a fixed size box attached to each tick mark. Images were scaled to cover the box, and short text was presented as as scrollable content within each box (Fig 3.2). The interface selects global grounding examples by sorting the set of potential examples and progressively selecting

examples that are at least a certain minimum distance from each other. As an annotators scrolls the slider handle, we dynamically search for immediate neighbor examples above and below the slider position and render them as additional anchors placed among the global grounding examples. Neighbor examples are also placed next to the item being annotated to facilitate comparison. Vertical positioning of the rendered anchor examples is dynamically adjusted so that they never visually overlap with each other.

As our experiments were conducted on the Amazon Mechanical Turk crowdsourcing platform, we also implemented a gated training [178] phase for each of the annotation experiments. This phase focuses on training the crowd workers to use the annotation interface rather than annotating a specific task domain, so we used a common training example based on age estimation across all domains. Workers are presented with a prompt describing the task and interface, including specific actions that can be performed using the interface. As workers complete each annotation step for the training task, we check their partial answers against the reference and provide just-in-time feedback if they make a mistake. Once the worker accurately completes the training task, they will progress into the actual annotation task and given the specific instructions for the domain they are annotating. We implemented some basic quality control measures to prevent gaming of the task such as requiring workers to have interacted with the slider before they are allowed to proceed onto the next item.

3.4 Experiments

In order to answer the research questions behind our Goldilocks designs, we conducted annotation experiments using data from 3 domains on the Amazon Mechanical Turk (AMT) platform and using interfaces that isolate specific aspects of Goldilocks for experimentation. Specifically, we tested the following hypotheses:

- RQ1: Does grounding with examples improve consistency?
  - H1-a: Using example-based anchors reduces the amount of disagreement between annotators on ratings of items compared with using semantic text descriptions as anchors.

  - H1-b: Including an annotator's own annotations from the session as additional anchors results in improved self-consistency reflected by less disagreement with their past placement when placing items again.

- RQ2: Does the range-based process create robust output for understanding relationships between items?
  - H2-a: Range annotation captures item resolution and thus can more accurately model distributions of pairwise relationships (more than, less than, indistinguishable) compared to distributions produced by comparing single value annotation output.
  - H2-b: Resolution of items captured using range annotation are better for modeling pairwise relationships than resolution captured through inter-annotator (dis)agreement.

- RQ3: Does the uncertainty about items captured through the size of the ranges correlate with uncertainty captured in the form of inter-annotator disagreement in traditional semantic scale absolute ratings?

3.4.1 Annotation Task Design

We describe in more detail the task design we used in our annotation experiments, including interfaces derived from Goldilocks and ones from traditional annotation. Unique crowd workers were recruited to use one of the following interfaces to provide annotations for a group of examples:

- **Single Value with Semantic Anchors (SV-SA)**: In each step, annotators are are asked to find a slider position that represents the placement of one item in the annotation sequence using a semantic scale as reference (Figure 3.4 top).

- **Single Value with Example Anchors (SV-EA)**: In each step, annotators are asked to find a slider position that represents the placement of one item in the annotation sequence using a scale anchored by other example item instances (Figure 3.4

bottom). Depending on the experiment and condition, the annotator's past placements in earlier steps may become additional anchors for steps in the future.

- **Pairwise**: Annotators were asked to compare all pairs of items. For each step in the annotation sequence, an annotator was presented with 1 reference item and a list of items it has not been compared to yet. For each item, the annotator was asked to judge the relationship of that item compared with the reference item ($>$,$<$,$\approx$).

- **Range with Hybrid Anchors (R-HA)**: This represents the full proposed Goldilocks design. Annotators are given both semantic labels and example instances as reference anchors. For each item, an annotator is first asked to place a lower bound marker for the item followed by placing an upper bound marker. Ranges annotated in earlier steps are incorporated as additional anchors.

Our first study (3.4.5) examines whether example anchors (**SV-EA**) improve agreement between annotators compated to semantic anchors (**SV-SA**). Following that, our second study (3.4.6) examines whether including an annotator's past placements improves within annotator consistency when using the **SV-EA** annotation design. Finally, in our last study (3.4.7), we collect ground truth pairwise relationships directly using the **Pairwise** interface, and compare how well we can recover the distribution of these relationships using data from the traditional single-value semantic anchor approache (with **SV-SA**) with that of the full Goldilocks range annotation design (**R-HA**).

In all cases, annotators were first given a brief gated "interface training" instructional stage where they are guided to annotate a single item (based on an age estimation domain) using the annotation interface they were assigned. Instructions are provided during the process to guide them through using the interface and feedback is given if the annotator makes a mistake in the annotation. Once an annotator completes the annotation process without mistake, they are given details about the actual task domain they are annotating. Each annotator is then prompted to annotate a sequence of items using the assigned condition's interface.

3.4.2 Annotation Domains and Datasets

We selected the following 3 annotation domains to to conduct annotation tasks: TOXICITY, SATIETY and AGE. These domains were selected to represent common types of rating tasks that have subjective aspects where a Goldilocks style approach to annotation could be desirable. These tasks also span two different modalities, short text and image, which closely align with rating tasks commonly conducted.

Toxicity

For this task domain, annotators judge the degree of toxicity in a short online comment, estimating how strongly the author of the comment intended to offend. Research has demonstrated that human judgments of online toxicity vary considerably from rater to rater due to subjectivity of the task [229]. The TOXICITY domain represents a short text annotation task where annotators compare pieces of text that only consist of a couple of sentences. Similar tasks include judging fluency of text generation or judging text sentiment. To produce the annotation dataset for this domain, we sampled a 50:50 label-balanced subset of 100 comments from the Jigsaw comment toxicity classification challenge dataset [294] behind the Perspective API[2] which contains Wikipedia comments and binary labels of toxicity. Only comments that had between 4 and 280 characters (after markup removal) were sampled. When presenting the task to crowd workers, we borrow Perspective API's definition of a toxic comment: 'a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion'. We also contrastively define healthy comments as those 'relevant to the discussion' and further note that comments 'can express disagreement'.

Satiety

For this task domain, annotators judge how filling (satiable) is the food depicted in an image, taking into account the type of food and the portion size. The SATIETY domain represents an annotation task that contains uncertainty in the visual modality. Prior research has shown that while pairwise comparisons of food for expected satiety can result in robust ratings,

[2] `https://www.perspectiveapi.com`

personal familiarity also resulted in biases [37]. We produced the annotation dataset by selecting a subset of food types from the Food-101 dataset [31] and then sampling images for each selected food type up to a total of 80. One round of manual inspection was also done to verify food was clearly discernible in all images.

Age

For this task domain, annotators estimate the age of the subject depicted in a photo. The AGE domain is another annotation task in the visual modality that contains uncertainty, however age is grounded to a concrete scale that we expect most people to be already familiar with. We produced the annotation dataset by sampling a subset of 100 portrait images from the FG-NET face dataset [92].

3.4.3 Anchors for each Domain

To maintain consistency across experiments, we defined a set of text-based semantic differential scale anchors and a set of example anchors for each domain that was held constant across experiments. For the semantic scale anchors, we used text descriptions similar to 7-point Likert or semantic differential scales. Example anchors consisted of 7 roughly evenly spaced in-domain items each associated with a position on the scale.

For the TOXICITY domain, we used the following text descriptions for semantic scale levels: "1 - Not Toxic at All", "4 - Somewhat Toxic" and "7 - Extremely Toxic". Other levels (2, 3, 5, 6) on the scale were presented as a number without any associated description. The 7 example anchors were manually picked from a set of annotated examples produced from a pilot run of the cold start process with crowd annotators.

For the SATIETY domain, we used the following text descriptions for semantic scale levels: "1 - Very Hungry', "2 - Somewhat Hungry", "3 - Almost Satisfied", "4 - Satisfied', "5 - Full", "6 - Very Full", and "7 - Can't Finish". The 7 example anchors were produced by the authors producing gold annotations directly using the cold start process interface.

For the AGE domain, we used text scale levels based on numeric age values ranging from "0" to "60+" incrementing in steps of 10. The 7 example anchors were picked by finding

all images corresponding to each semantic age level and then drawing a random one at each level and assigning its value to be the ground truth age.

3.4.4 Crowd Annotator Recruitment and Compensation

We recruited annotators for our experiments from the Amazon Mechanical Turk crowdsourcing platform from the United States with the qualification of approval rate no lower than 90% and over 1000 approved HITs completed in the past. Across all studies, annotators were only allowed to participate in annotation if they had both not used the corresponding interface and not annotated the domain before. Overall, we recruited 655 unique workers across all 3 studies with an additional 44 unique workers who only participated in the pairwise annotation used to establish the ground truth for Study 3. For all annotation tasks, we set a base pay of \$0.10 which was given if the worker completed the training phase. Remaining compensation was distributed in the form of a bonus based on the interface being used and the number of items annotated.

Participants assigned to the **Single Value** tasks (both with **Semantic** and **Example** anchors) were given a per-item bonus of \$0.03 (for annotating a group of 10 or 20 items). Participants assigned to the **Range** tasks were given a per-item bonus of \$0.05 (a total of 10 items). Participants assigned to the **Pairwise** annotation tasks were given a per-comparison bonus of \$0.01 (a total of 45 comparisons). We set pay based on our estimate of time needed taken from pilot studies and used completion bonuses to correct for any discrepancies. Based on condition, a final completion bonus of \$1.00, \$0.50, or \$1.00 for each of the previously mentioned interfaces respectively was provided. We distributed the final bonus in 2 batches as the initial completion bonus values we set for the tasks resulted in a measured hourly pay that was lower than desired. The final hourly rate measured between \$9.70 and \$10.90 across the various domains and interfaces when assuming the median work time for each interface.

Manual quality checks were conducted on cases with a large number of similarly annotated values across different items (e.g., consistently placing at 0 or 1) as well as abnormally short work time, resulting in removal of 5 workers (and re-collection of corresponding anno-

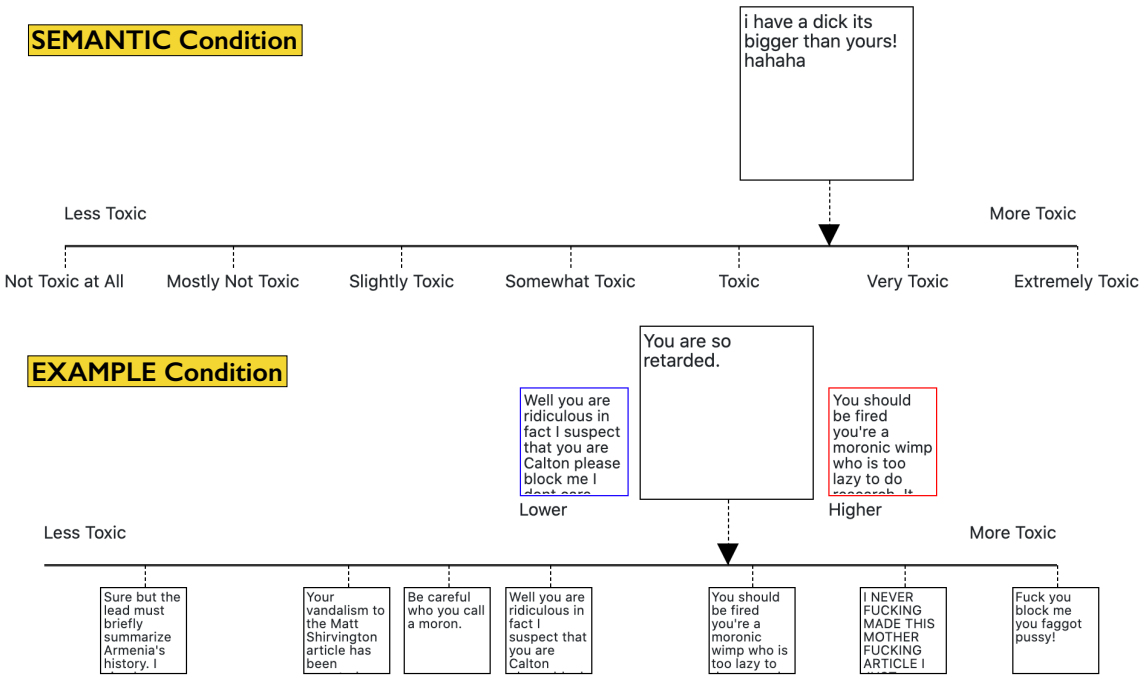


Figure 3.4: Screenshot showing the two interfaces conditions (top: SEMANTIC, and bottom:EXAMPLE) used to evaluate consistency consistency between annotators. Examples shown in figure are from the toxicity domain pilot tasks. (Content warning: toxic comments including offensive and swear words are shown in their original form as a part of this figure.)

tations) across all experiments. Removed workers were included in the counts of recruited workers above. Within the removed workers, those intentionally spamming across their entire sequence of annotations (choosing the exact same placement for all items) only received the base pay for the task.

3.4.5 Study 1: Evaluating Consistency Between Annotators

We first explore whether example-based grounding presented in Goldilocks can improve consistency between different annotators (H1-a). For this experiment, we assigned each annotator to one of two conditions: SEMANTIC, where they were given 7-point text-based

semantic anchors and presented with the **SV-SA** interface; or EXAMPLE, where they are given 7 example instances placed onto the scale using the **SV-EA** interface. For each domain, the anchors used are detailed in 3.4.3. We drew example anchor instances for the TOXICITY and SATIETY domains from past pilots of semantic differential scale annotation on a disjoint set of items, using average rating to establish their initial placement. For the AGE domain, example instances were selected from a separate set of images drawn from the same dataset using the included ground truth age labels for initial placement.

After the training, each annotator was tasked with annotating a sequence of 10 items using the interface of the condition they were assigned. To create sequences, each domain's dataset was shuffled once and then partitioned into equal-sized disjoint sets. Each sequence for each domain was annotated by 10 workers in each of the two conditions. Annotators' placements of items on the scale was mapped as a continuous numeric value within the range [0, 1]. For the TOXICITY domain, the first and last items in each sequence were set to the same item to pilot measurement of within-annotator consistency, so only the 8 remaining annotations were used for analysis in this experiment.

Results

To evaluate the amount of consistency between annotators for each annotated data point, we computed the standard error across annotators as a proxy for the amount of disagreement. We note that the standard error values are comparable across conditions as the range of values on the scale and number of annotators was fixed between all conditions. We also evaluated the significance of any difference by conducting a two-tailed paired t-test on the standard error of each annotated item across each pair of conditions (SEMANTIC versus EXAMPLE) in each domain. A summary of the results are shown in Table 3.1.

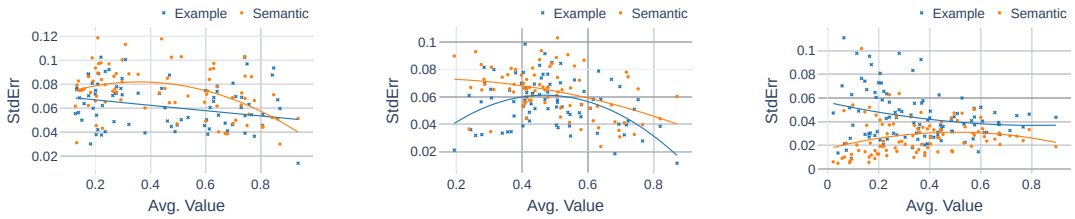
We observed a statistically significant decrease in value disagreement across annotators for the TOXICITY and SATIETY domains, providing support for hypothesis H1-a. However, we observed a statistically significant increase in disagreement across annotators for the AGE domain, which contradicts H1-a. We then plotted the disagreement (standard error) in both conditions for each item against the mean value across both conditions in each domain

Table 3.1: Results for the experiment measuring consistency between annotators comparing between SEMANTIC and EXAMPLE conditions. Average disagreement (Avg. Dis.) is calculated as the standard error (over 10 annotators) for each instance averaged across all annotated instances. Significance testing done as a paired t-test across conditions for disagreement. We also examine how much of the 0-1 scale is being used by annotators on average in each condition by averaging each annotator's minimum and maximum rating values.

Domain	**Condition**	**Avg. Dis.**	**Significance**	**Scale Util. ($\overline{\textbf{Min}}$, $\overline{\textbf{Max}}$)**
TOXICITY	SEMANTIC	0.07348	P <0.001	0.773 (0.103, 0.876)
	EXAMPLE	**0.06379**	Very Significant	0.794 (0.104, 0.899)
SATIETY	SEMANTIC	0.06373	P < 0.005	0.603 (0.230, 0.833)
	EXAMPLE	**0.05548**	Significant	0.635 (0.166, 0.801)
AGE	SEMANTIC	**0.02765**	P < 0.001	0.696 (0.054, 0.751)
	EXAMPLE	0.04443	Very Significant	0.593 (0.072, 0.665)

to understand the behavioral differences we see with the age domain as shown in Figure 3.5.

We find that the pattern for disagreement in the SEMANTIC condition is consistent with behavior observed in prior work [288] for similar domains with subjectivity and uncertainty. However, we note that overall disagreement between annotators was lower in the AGE domain compared to the other two domains. We also noted that scale utilization was similar in both conditions for the TOXICITY and SATIETY domains, exhibiting a slightly increase in utilization of the full scale in the EXAMPLE condition. Prior work in psychology has shown that increased spacing of items has relatively minimal effect on accurate placement when items are discriminable [255] so we don't expect this slight increase in scale utilization to affect disagreement levels. However, opposite to the other domains, the utilization of the scale in the AGE domain was 10% *lower* for the EXAMPLE condition. We hypothesize that unlike the TOXICITY and SATIETY domains, estimating age from appearance is a domain



(a) TOXICITY domain. (b) SATIETY domain. (c) AGE domain.

Figure 3.5: Scatter plots of disagreement between workers (as measured by standard error) for each item plotted against the mean annotated value of each item. Trendlines represent a fit with a degree 2 polynomial.

where a numeric age scale is actually more consistently understood by human annotators, thus example anchors provide no further benefit to annotators in understanding the scale. The scatter plots in Figure 3.5c indicate that uncertainty for younger subjects was much higher in the EXAMPLE condition. Combined with the lower scale utilization we observed for EXAMPLE, we hypothesize that uncertainty about judging exact age is higher for older subjects. As we only show example-based anchors in the EXAMPLE condition, this increased uncertainty about the reference images depicting older subjects may have resulted in more hesitation to use the higher values on the scale. This suggests that: (1) comparisons with anchor examples mostly benefit cases where shared understanding of the scale is low, and (2) example-based anchoring should be used in *addition* to semantic anchors as *only* using example anchors can be detrimental to consistency if the domain is one where the semantic scale has a high degree of shared understanding already. Drawing from this experiment, our full Goldilocks annotation process uses both example-based anchors and semantic anchors to frame the scale.

3.4.6 Study 2: Evaluating Consistency Over Time Within Annotator

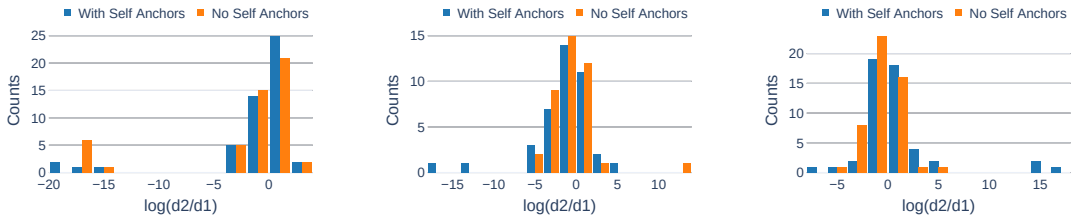
For our second experiment, we explored the effect on self-consistency resulting from including an annotator's own past annotations as additional reference examples augmenting an

initial seed set (H1-b). The example-based **SV-EA** interface was used for this experiment, with each annotator was assigned one of the two conditions: CONTROL, where only the seed set examples was used for reference anchors; or AUGMENT, where an annotator's own past annotations in the same session were included along the seed examples as references. Since we are interested mainly in the effect on self-consistency, we reduced the initial set of seed examples to just 3 examples for each domain drawn as a subset of the 7 example instances used in the EXAMPLE condition of the previous experiment. We took the items corresponding to the lowest, highest, and median ratings.

The items in each domain were shuffled and then partitioned into sequences of size 20, resulting in 5 sequences for the TOXICITY and AGE domains and 4 sequences for the SATIETY domain. Each annotator was given interface training and then subsequently tasked with annotating one of the sequences (of 20 items). To probe for changes in the rating of an item, we replaced the 10th and 20th items in each sequence above with repeats of the first item, which we will refer to as the probe item. When the probe item is annotated in the AUGMENT condition, the annotator's own past annotation for the probe item will be withheld from the set of reference items. We measure Δ_1 as the size of the value change between the first and second annotation attempts of the probe item and Δ_2 as the size of the value change between the second and third annotation attempts of the probe item.

Results

From Table 3.2 we can see that for most domain condition pairs, the absolute amount of an annotator's disagreement with their past rating tends to exhibit a natural decrease as they get familiarized with the scale. Since the magnitude of initial self-disagreement for the probe item varies for each annotator, comparing absolute change in self-disagreement can be misleading as the same *proportional* change in self-disagreement will reflect as a larger *absolute* change. To account for these factors, we instead look to the self-disagreement *ratio* (Δ_2/Δ_1) as a measurement for the proportional decrease (or increase) in self-disagreement. Ratios below 1 indicate that self-disagreement has decreased while those above 1 indicate an increase. In Figure 3.6, we show a histogram of this ratio on a log-scale for each condition



(a) TOXICITY domain. (b) SATIETY domain. (c) AGE domain.

Figure 3.6: Histogram of distance ratios between first re-annotation and second re-annotation of the probe item on a log scale. Negative values indicate more decrease in disagreement with the annotator's own answers while positive values indicate more increase in disagreement with the annotator's own answers. Ratios were smoothed using Laplace smoothing with $\epsilon = 10^{-8}$.

in this study.

Our first step is understanding whether self consistency improves over time simply from doing the task and being exposed to more examples. We conducted a sign test for each of the task domains and find that in the TOXICITY domain, self consistency does improve over time (P ¡ 0.005) for both CONTROL and AUGMENT conditions. Self consistency was not found to have a significant across-the-board improvement in any of the other domains. Comparing across the two conditions, we did not measure significant effect on self-disagreement ratio in any of the 3 domains.

We then hypothesized that effect on self-consistency may not be uniform across all probe items—if an annotator already has low self-disagreement in the first re-annotation round (Δ_1), it likely implies there is little uncertainty about the placement of the item and thus we shouldn't expect further improvements. Considering this, we now look at only the top 30% 'most uncertain' annotation sessions for each domain and condition combination, as sorted by decreasing Δ_1. This set consists of 15 sessions for the TOXICITY and AGE domains and 12 for the SATIETY domain. In this high-disagreement subset of sessions, we find that augmenting reference examples (AUGMENT) with past annotations in the session

Table 3.2: Table breakdown of the change in rating for the probe item (compared to its last most recent rating) when re-annotated for the first time (Δ_1) and when re-annotated the second time (Δ_2). The "Top Avg. Δ" columns represent the averages when only considering the instances where Δ_1 was among the top 30% most uncertain.

Domain	**Condition**	**Avg. Δ_1**	**Avg. Δ_2**	**Top Avg. Δ_1**	**Top Avg. Δ_2**
TOXICITY	No Self (CONTROL)	0.105	0.062	0.244	0.077
	With Self (AUGMENT)	0.133	0.090	0.347	0.095
SATIETY	No Self (CONTROL)	0.140	0.086	0.308	0.171
	With Self (AUGMENT)	0.126	0.052	0.286	0.027
AGE	No Self (CONTROL)	0.110	0.063	0.280	0.133
	With Self (AUGMENT)	0.063	0.066	0.157	0.067

does result in a larger proportional reduction in self-disagreement (reflected through self-disagreement ratios) when compared to CONTROL for both the TOXICITY and SATIETY domains. For the SATIETY domain, median proportional decrease in self-disagreement was 0.076 (92% reduction in self disagreement) for the AUGMENT condition compared to 0.263 (74% reduction) for the CONTROL. The median ratios were 0.190 (81% reduction) and 0.310 (69% reduction) respectively for the TOXICITY domain. However, the limited amount of data points in these groups means we do not have statistical power to claim significance. Overall, we don't find sufficient support for H1-b, but we note a pattern of improvement in self-consistency for items with high initial self-disagreement when including an annotator's own prior annotations as additional references. Similar to the previous section, we were unable to observe benefit of augmenting reference examples on the AGE domain, likely due to the already limited utility of reference examples in this domain.

3.4.7 Study 3: Evaluating Range Annotation

For the final experiment, we explored how robustly ranges produced by the two-step annotation process in Goldilocks reflect properties of relationships between items. In this experiment, annotators were asked to annotate a sequence of items using the full Goldilocks two-step annotation process (using the **R-HA** design shown in Figure 3.1). The annotation experiments were conducted on the TOXICITY and SATIETY domains with sequences generated by shuffling each dataset and partitioning the dataset into groups of size 10, resulting in 10 and 8 groups respectively for the two domains. We then recruited 5 annotators to annotate each sequence in each of the domains.

At the start of the task, each annotator was first trained on how to use the two-step annotation system described earlier in Section 3.3.2 by annotating a sample task with guidance given during each step. After the annotator completes the training example item, they then proceed to annotate the assigned sequence of 10 task items. To seed the initial reference examples, we used the same reference anchors as used in the first experiment. We also included each annotator's own annotations as anchors during their annotation in a similar way as the AUGMENT condition in the second experiment.

Establishing Pairwise Relationship Distributions

In order to measure ground truth distributions over the pairwise relationships, we recruited separate workers and used the **Pairwise** design to directly collect pairwise judgments on relationships ($>$,$<$,$\approx$) between all pairs of items in each group. Distributions across the 3 relationship types were then created by counting the proportion of annotators indicating each type of relationship across for each pair of items. These distributions reflect the degree of disagreement among annotators for the pairwise relationship.

We then considered how one would recover similar distributions across relationships for pairs of items using the traditional approach of single-value absolute rating scales based on semantic anchors, creating two alternative baselines. Since the traditional approach cannot simultaneously elicit item ambiguity and agreement, producing a similar distribution would involve a tradeoff.

For the **Direct** baseline, we assume that there is no item-level ambiguity, meaning that even local pairwise comparisons can be made by directly comparing the raw values from the absolute rating. For example, we count an annotator as indicating a ">" relationship on a pair (a, b) if their single rating scores indicate $r_a > r_b$. One can generally expect this to be reliable when a and b are far apart on the scale but it can be much less reliable for close neighbors.

For the **Infer** baseline, we assume that all disagreement observed between annotators reflects the ambiguity of the item. In this case, we aggregate the individual ratings into a *single* 95% confidence interval for each item by measuring the mean and standard error between these samples. We then infer the relationship between of a pair of items by comparing the confidence intervals, treating overlapping intervals as indicating a relationship of 'indistinguishable ($\approx$)'. In this case, the distribution across relationships for a pair would see all the probability mass allocated to the single relationship measurement produced by the comparison.

Finally, with Goldilocks annotation, we have range evaluations on a per-annotator granularity. For each annotator, we can use their range labels to find the relationship between two items, treating overlaps as indicating $\approx$. We can then produce a distribution by counting the proportion of annotators indicating each relationship. With Goldilocks we don't need to make tradeoffs between measuring item ambiguity and agreement.

Results: Recovering relationships between items

To compare and quantify how robustly each of these methods recovers relationships, we measured the Wasserstein distance between relationship distributions for each of the 3 approaches in 3.4.7 and the ground truth relationship distributions collected through pairwise comparative rating. Table 3.3 shows that among the 3 methods to produce distributions over pairwise relationships, recovering distributions using range labels most accurately agrees with the ground truth distribution, supporting H2-a.

We found that using inter-annotator agreement to infer the inherent ambiguity (referred to by prior works as "resolution") of items results in an over-estimate of the amount of

Table 3.3: Comparing the quality of the pairwise relationship distributions as recovered by (1) **range**s collected in Goldilocks, (2) **direct**ly comparing raw values picked by each annotator, and (3) indirectly using ranges **infer**red from the 95% confidence intervals. Details in 3.4.7. Wasserstein distance to the ground truth distribution (collected directly using pairwise comparisons) was computed for each case. Goldilocks ranges produce distributions the closest (least distance) to the ground truth.

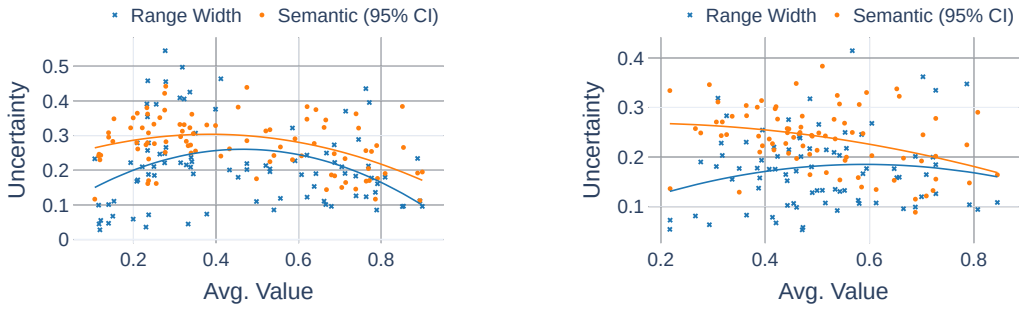
Domain	Avg. WD (Range)	Avg. WD (Direct)	Avg. WD (Infer)
TOXICITY	**0.332314**	0.366944	0.450556
SATIETY	**0.424352**	0.449444	0.597222

ambiguity. In the TOXICITY domain, 43.5% of the relationships that were distinguished in the ground truth distribution collected directly through pairwise comparisons were inferred to be "indistinguishable", with this ratio as high as 68.1% in the SATIETY domain. In contrast, ranges over-estimate ambiguity (under-estimating resolution) only about half as often, with 22.1% and 30.9% respectively. This supports the idea that ranges are a better model of resolution (H2-b).

Results: Comparing aggregation uncertainty with range sizes

Finally, we explored differences in the type of uncertainty measured through Goldilocks annotation ranges sizes with uncertainty measured by confidence intervals of annotations using semantic scales. We hypothesize that since ranges focus on capturing resolution (distinguishability against peers) of items, the resulting uncertainty represented by the size of ranges will be different than uncertainty represented by inter-annotator disagreement metrics, though the two may still be related.

First we look at the behavior of the two kinds of uncertainty measurements across the range of values on the scale. Figure 3.7 plots the two kinds of uncertainty: average size of ranges and 95% confidence intervals for semantic scale annotation values. We find



(a) TOXICITY domain.

(b) SATIETY domain.

Figure 3.7: Comparison of uncertainty measured as range sizes from Goldilocks annotation with uncertainty measured through standard error confidence intervals from traditional single-value semantic scale annotation. Trendlines represent a fit with degree 2 polynomials.

that overall range sizes represent uncertainty lower than that measured by 95% confidence intervals from aggregating semantic scale annotation (P ¡ 0.001). This makes intuitive sense as we would expect item level resolution to be a tighter uncertainty. We also find that in the TOXICITY domain, both types of uncertainty behave similarly with respect to extreme values on the scale corresponding to lower values of uncertainty in definitions. In the SATIETY domain, however, we found that lower values (corresponding to foods depictions that are less satiating) corresponded to larger uncertainty in the form of disagreement but not with range sizes. We think this may result from higher disagreement about what foods are not satiating among different annotators but with annotators each confident about their own determination of satiety (high resolution/distinguishablity of items).

Looking at correlation between the values produced by the two types of uncertainty, we observe only very weak correlation between range sizes and confidence intervals (scaled standard error) for both TOXICITY and SATIETY domains with $R^2 < 0.01$ in both domains. This indicates that the uncertainty we measure with ranges does not have significant correlation with inter-annotator disagreement measures like standard error (RQ3). We note that with range annotations, inter-annotator disagreement measures can be further computed

for the range bounds themselves to evaluate disagreement separately from item uncertainty (resolution) captured by ranges. However, as single-value semantic scalar annotations can't facilitate separation of the two uncertainty types, we are unable to make direct comparisons.

3.5 Discussion

In the prior sections, we demonstrate that the ideas of grounding absolute rating scales with examples and explicitly capturing item-level measurement resolution can be beneficial for more consistent and robust annotation of subjective domains lacking shared understanding of absolute ratings scales. In this section, we will discuss some of the other considerations in adapting Goldilocks as a full annotation technique, including examining the annotation efficiency (in terms of work time) of Goldilocks compared to hybrid application of traditional methods and envisioning how Goldilocks may be scaled up to multiple annotation sessions using iterative-improvement processes. We will also discuss limitations of the Goldilocks process and potential avenues for future work.

3.5.1 Annotation Efficiency and Cost of Range Annotations

One of the main advantages of the Goldilocks annotation process is the ability to capture item-level ambiguity and disagreement between annotators simultaneously through the use of range annotations. However, separating these sources of uncertainty comes at an extra cost for the data collection process—even though range bounds in Goldilocks can be collected with low overhead compared to traditional absolute rating, the tasks can be more work for the requester to set up. This presents a tradeoff for practitioners when deciding whether the higher quality of data is worth the cost. Prior work simulating data annotation tasks inspired by measuring objective properties has shown that, given a fixed budget, some learning algorithms actually benefit more from a larger amount of lower-quality annotations on novel examples rather than higher-quality annotations on fewer items [173]. Indeed, for these tasks where disagreement is likely caused by noisy perception, it's likely that a practitioner will see relatively little benefit by separating item-level ambiguity from annotator disagreement. However, with the rising demand for training data in domains that involves subjectivity or

nuance, understanding and accounting for sources of uncertainty and limitations within the data itself has become increasingly important for building models that are *trustworthy* rather than just more *performant* [24]. Separating disagreement from inherent ambiguity using range-based annotation can also offer better transparency about the annotation process and data produced, allowing for the potential to diagnose model limitations and human biases even into the future. In these cases, the higher cost of setting up Goldilocks annotations can be justified by the richer information that can be derived from range-based rating data.

Of course, Goldilocks is not the only approach to capture both item-level ambiguity and disagreement. It is possible to use traditional absolute and comparative rating to separately collect scalar annotations and pairwise comparisons to recreate absolute rating estimates and pairwise relationship distributions. We also wanted to understand whether Goldilocks can provide efficiency benefits when compared to hypothetical hybrid approches using only a combination of traditional annotation interfaces. We look at the work time taken by crowd workers in our various experiments to extrapolate the effort necessary for such an approach. Assuming a task group size of 10 items, we find that the Goldilocks two-step workflow results in a median work time (including both training and annotation) of 429.5s per worker per task group on the SATIETY domain and 592.5s per worker per task group on the TOXICITY domain. Collecting only single value rating annotations with Likert-style anchors takes a median work time of 307.5s per worker per task group on the SATIETY domain and 238s per worker per task group in the TOXICITY domain. Finally, comparative rating on a group of size 10 implies 45 pairwise comparisons to capture full pairwise relationships, which takes a median time of 513.5s per worker per group and 502s per worker per group for the two domains respectively. Thus we expect that at the same level of redundancy for annotations, Goldilocks can be 20-48% more efficient through the use of our two-step range-based annotation that collects ratings and relationship distributions together. Consistency improvements of Goldilocks may be able to push efficiency further in practice by requiring a lower amount of redundancy to achieve the same level of agreement.

3.5.2 Goldilocks and Iterative Improvement

So far in this paper we have examined the ideas presented in Goldilocks only for single annotation sessions where we didn't need to update the anchor examples beyond incorporating an annotator's own ratings. In order to scale up to larger datasets, it becomes necessary to perform annotations over multiple sessions which involve using aggregation approaches to iteratively construct an updated set of anchors. To achieve this we envision a process based on the idea of iterative improvement [176].

In each round of iteration, a group of annotators individually annotate a subset of the dataset, sharing a 'seed' set of anchor examples used to ground the interpretation of the scale, with their own annotations also incorporated as they progress along the annotation session. Once all annotators have completed the session, the annotations collected will be aggregated into a new set of seed examples used to ground the next round of iteration. In addition to progressively annotating new examples, this iterative process may also be used to revise past annotations, such as those created during the cold start process. This can be accomplished by first removing the items to be revised from the set of grounding examples and then re-annotating them as new items in an iteration. This process of periodically aggregating annotations and then re-seeding anchor examples can serve as a method to scale up annotations while ensuring a stable scale as annotators place items.

We believe that this represents a feasible design for scaling up annotation, and we envision further work can be done to explore options for aggregation and re-annotation strategies as well as evaluate their effectiveness. We also see potential for using iterative improvement as way to dynamically re-calibrate the definition of scales to account for distributional shifts over time. For example, a scale that can dynamically adapt to improving quality of machine summarization systems can be adapted as a living benchmark. We think the ideas presented in Goldilocks for single annotation sessions provide a first step into building an effective iterative workflow.

3.5.3 Limitations and Future Work

While Goldilocks provides a path to more consistent scalar annotation that also captures uncertainty, we also recognize that the current design is still subject to some limitations which we believe can be good avenues for future work.

Creating High Quality Seeds in Cold Start

The cold start process in Goldilocks provides a way to generate the initial seed set of grounding examples that enable the comparisons and consistency benefits of Goldilocks. However, the quality of this initial set of seed examples can also influence whether consistency benefits can be realized. We observed some of these limitations when experimenting with example-based anchors in the AGE domain. A good seed set should consist of examples that achieve both good coverage of the scale and have low ambiguity themselves. When the seed set achieves good coverage over the scale, the comparative process can allow seed examples that are distinguishable to quickly be excluded from the range of the annotated item, resulting in measurement resolution that mainly depends on the number of examples in the seed set. However, a set of examples that is not representative of the full range of items to be ranked can lead to issues of scale drift when these examples (that annotators may desire to rate higher or lower than the current implied bounds of the scale) are encountered in the future. The current cold start process provides some mitigation to the issue of representativeness by incorporating a 'resampling and replace' phase to increase the diversity of items in the seed set. However, for sufficiently large datasets this may not be enough to capture rare items that are also outliers for the scale. For future work, we envision enabling the ability for annotators to rescale the visible scale itself through an interaction similar to zooming in or out, allowing the annotation of items that lie outside the current extremes of the scale when they are encountered.

Another current limitation of the cold start process is that the cold start design cannot effectively capture item ambiguity as we only elicit a single label for each reference item. In pilot studies we found it infeasible to introduce ranges into the cold start process as there are no anchors to compare against to effectively determine these ranges. It is possible to have

suboptimal seed sets where the seed items can have high ambiguity themselves, thus acting as a lower bound on range sizes. We hypothesize that the iterative improvement process in 3.5.2 may offer a way to limit the impact of the cold start seeds if we can conduct subsequent annotation rounds where we can instead seed with regular annotated range data, though we leave exploration of this to a future study.

Addressing Long-form Tasks and Context

Some common tasks where crowd scalar ratings are desirable, such as evaluating relevance, conciseness, fluency, or faithfulness of summaries produced by text summarization models, can depend on understanding long-form context (e.g., a news article) or even multiple documents [85]. While we have shown that Goldilocks can support annotation domains based on small amounts of text (1-2 sentences) using a similar interface as the one used for images, long-form text will require a different design for conducting comparisons both with the global scale and local neighborhood.

Additionally, interactions in Goldilocks assume that items can be compared against other items in the same dataset. However, when rating items with context, such as summarization or translation, it is likely that reasonable comparisons can only be made with certain other items sharing the same context (i.e., alternate summaries/translations of the same source). A potential avenue for future work extending Goldilocks may exist in introducing virtual views to the Goldilocks scale that enable contextual comparisons on the scale by only exposing items sharing the same context. Future work on an algorithm for determining optimal global example anchors could also take into account aspects that could make comparison easier, such as similarity to the item being annotated.

Working with Density

One of the strengths of Goldilocks is the ability to use past annotations from any source, including data from existing datasets to establish grounding for a scale. By providing past annotations from a dataset as reference examples, it will be possible to augment the dataset in a way that is consistent with past examples but also doesn't require building complex

rubrics. However, as the set of past annotations increases, it poses potential problems for the local comparison aspect of the Goldilocks annotation process. There are practical limitations on how fine adjustments can be on a slider-based scale, so as regions on the scale become densely populated by examples, it becomes harder to use local comparisons to find precise upper and lower bounds in those regions. Even small adjustments in a dense region can mean moving across many reference points.

One potential solution to the density problem could come from allowing the scale to be itself scaled, similar to that proposed in 3.5.3. Initially the full view of the scale is presented along with global anchors for coarse navigation. As an annotator narrows down on a dense region, they can increase the zoom level of the annotation scale to span just the dense region across the entire width of the scale, increasing the amount of space and in turn reducing interaction issues caused by density. New global anchors can be selected to allow for quick navigation at the new zoom level.

3.6 Conclusion

In this paper, we present and evaluate Goldilocks, a novel technique to elicit scalar annotations using the crowd that improves on consistency and captures pairwise relationships more robustly. We show that by prior examples can be used as anchors to ground otherwise abstract absolute rating scales (such as semantic or Likert scales) leading to more consistent interpretation between workers. We find that including an annotator's past annotations in a session can lead to more self consistency on items that have high initial uncertainty. Finally, we show that introducing range annotation into absolute rating can enable simultaneous elicitation of both perceived ambiguity on a per-annotator scale while also capturing inter-annotator disagreement. This simultaneous measurement enables a better recovery of pairwise relationship distributions.

Chapter 4

CASE LAW CROWDSOURCING: USING PRECEDENTS TO GROUND UNCERTAINTY IN CATEGORICAL JUDGMENTS

In the previous chapter, we presented a tool for conducting rating tasks in the form of continuous scalar rating. However, while scalar ratings are common, the ordering imposed by a rating scale does not translate to other judgment tasks, such as categorical judgments.

In this chapter, we introduce *case law crowdsourcing*, a novel approach to scale up adjudication around complex decision bounds that takes inspiration from the legal concept of case law, or law established by decisions from prior cases as opposed to by statutes. In our workflow, previously decided cases form *precedents* to serve as fine-grained criteria for judging new cases. Then, when provided with a new case, an annotator explores the space of relevant prior cases and collects ones that they believe should be either *positive* precedent or *negative* precedent for the current case. These two sets of cases serve to illuminate the annotator's *decision bound*, resulting in an elicitation of both the reasoning behind their final judgment on the case and their confidence surrounding it. To enable case exploration at scale, we contribute an interactive tool that recommends relevant cases so that adjudicators can quickly understand the local decision bound around a case. In an evaluation using prior human judgments taken from `r/AmITheAsshole`, we find that case law crowdsourcing produces more consistent judgments across annotators compared to traditional methods based on predefined criteria and examples. We also found that we are able to identify situations where underspecified decision bounds were reflected through less confident judgments.

4.1 Introduction

The use of human judgments, both from individuals and groups, has become increasingly prevalent as a way to understand and produce ground truth answers for complex problems. However, as the complexity of the problems increases, the ability to clearly define the criteria

for making the "right" judgment becomes increasingly difficult or even impossible. Whether it's a community of Wikipedians debating about what should be included in an article [131], or a group of moderators making a call on what to do with a post [153, 123], or even a single crowd worker deciding whether an image fits the definition of some class of items [46], it is crucial to establish clarity and consistency in what constitutes the distinctions used to make that judgment—the decision bounds.

Traditionally, the job of discovering, defining, and conveying such distinctions has been reserved for expert task designers. To accomplish this, task designers often start by examining some samples of the problem, followed by using their expertise to come up with generalized procedures and criteria, and finally, conveying them to the adjudicators in the form of task guidelines, instructions, and training mechanisms. Indeed, the result of these procedures can be seen in the complex rules and guidelines established by Wikipedians, moderators and crowdsourcing requesters [39].

However, this has led to a couple of problems when scaling such processes. For one, creating and maintaining high quality guidelines is a costly job that requires a lot of expertise and continuous involvement of experts [213]. This often results in long turnaround times when new examples arise in areas where guidelines may have been absent—as we have witnessed in how platforms have addressed the ever evolving types of misinformation [283]. Secondly, even with a good set of guidelines and criteria, it can be challenging to confirm that they are being applied in a consistently manner by all human adjudicators. Many times, decisions are made without documenting arguments, and even when arguments are documented, there is often limited ability to systematically understand disagreements and dissent.

Thus in this work, we propose a novel approach—*case law crowdsourcing*—that takes inspiration from the concept of case law in legal systems to simultaneously address the challenge of defining complex decision bounds and validating the judgments around them. As a legal mechanism, case law makes use of past decisions–precedents–to define complex nuanced decision bounds through establishing connections and distinctions between details and facts in new cases against that of established ones. In this way, the precedents can serve as an evolving set of guidelines, while the relationships established form a type of support

or argument for the decision. In our proposed workflow, human adjudicators utilize an exploration interface to understand the local decision bound around a case being adjudicated without the need for complex guidelines and training. Adjudicators then use a set-based annotation system to establish commonalities and distinctions between the adjudicated case and any relevant precedent cases. These constructed sets then serve as the argument for the judgment outcome from each adjudicator. By examining across sets produced by multiple adjudicators, we can infer the level of consensus, conflict, and collective uncertainty around each decision.

In this paper we make the following contributions:

- We introduce case law crowdsourcing as a generalizable technique and workflow for rapid adjudication of novel cases through exploration of prior decisions and construction of precedent supporting sets.

- We implement a system for carrying out case law crowdsourcing at scale that enables annotators to quickly home in on a relevant context for a given case and explore the local decision bound.

- We show that our workflow produces more consistent judgments compared to using guidelines for the same task, achieving a standard error disagreement value of 0.227 across annotators compared to 0.406 using traditional guidelines.

- We also show that while the quality of judgments can be affected by the quality of precedents, it is possible to identify situations where past cases are insufficient as precedents by looking at how consistently positive precedent sets are constructed across different annotators.

4.2 Related Work

In this section, we examine four areas that motivate and inform our work: (1) the prevalence of modern tasks/domains that involve complex decision bounds; (2) traditional approaches in crowdsourcing to deal with complex bounds; (3) why it is desirable to have rationales

when tasks are complex; and (4) tools and techniques from legal practitioners carrying out case law.

4.2.1 Complex Decisions in Human Judgment

Humans have been making decisions based on complex criteria for millennia. This has only increased as our lives have become more complicated, intertwined, and information dense. But a number of major shifts due to advances in technology in the last several decades have led to a need to collect human judgments on complex tasks at an unprecedented scale.

One major shift has been the rise of online social media, leading to an explosion in content creation and, in response, a desire for moderation of that content that has risen in tandem. As human expression is multi-faceted and ever-changing, so are the types of content that communities and platforms seek to moderate, including ill-defined concepts such as toxicity [273, 161], misinformation [25], or hate speech [225]. As different communities have differing conceptions of undesired or objectionable expression [44, 45], communities oftentimes recruit volunteer moderators from within their community to manually conduct moderation [185], as they can weigh complex criteria such as the tradeoffs between different values their community holds [286]. When it comes to platforms that host content for millions or billions of people, conducting moderation so that judgments are attuned to cultural or community nuances becomes significantly more challenging, if not utterly intractable [139]. Still, platforms make an effort to be comprehensive and consistent when judging different categories of violative content. They try to achieve this through extensive training of thousands of paid moderators [223] and the maintenance of long internal guidelines and training manuals delineating a laundry list of positive and negative carveouts and examples for each category of content [152]. One leaked set of internal manuals from Facebook contained over 1,400 pages across over 100 manuals [112].

A second major shift has been a rise in the use of machine learning, where powerful models for classifying text, images, and other content can be built that are trained on large human-labeled datasets [244, 141]. While earlier models focused more on classifying more objective and narrowly defined concepts, increasingly models are being built and deployed

for classifying more socially and culturally situated concepts [106]. For instance, in the field of natural language processing, researchers have built classifiers to automatically tag some of the complex concepts that content moderators currently determine manually [294]. To build these models, researchers must curate large datasets that are hand labeled by crowdsourced human annotators in a consistent manner. However, when labeling these concepts, annotators often produce labels that conflict with each other. Sometimes, this is due to annotator positionality, where annotators bring in their own lived experiences and subjective biases when interpreting a concept [69]. Other times, the content itself may be difficult to label in certain ambiguous cases [49], or the guidelines as a whole are incomplete or ambiguous [46]. Still other times, researchers will disagree on what should be the definition for a concept [90] and even domain experts on the task may also disagree on the criteria [25].

4.2.2 Training the Crowd on Complex Decisions

Due to the demand for human judgments on complex decisions at scale, researchers have developed human workflows, computational techniques, and other strategies for training lay annotators on complex decision bounds. In their quest for consistent and high quality judgments, task designers have determined the importance of clear and carefully worded instructions and rubrics [293]. It is also common for task designs to incorporate a training phase for annotators to learn the rubric; researchers have also proposed implicit training via ordering of tasks [166] or adapting training to the annotator [34]. However, task design remains a difficult and time-consuming process, typically requiring some iterations between the task designer authoring the task materials, deployments with crowd annotators, and analysis of resulting annotations, leading to potential revisions to the task, and so on. Some research attempts to improve this process by guiding task designers through rubric iteration with input from the crowd [35, 213]. Beyond rubrics, past research has also focused on the aspect of training for workers via automated methods [289] or through task designs like gated instructions [178]. Additionally, there has also been effort focusing on the sub-goal of constructing or selecting examples as a way to train crowd workers on challenging

classification tasks [118, 247].

In addition to the expert led approaches, other prior works have explored a different direction, focusing on the data as it is being annotated. For example, structured labeling [158, 46] promotes the idea that crowd workers can explore the space of instances in an almost 'unsupervised' way, forming clusters of similar cases that with undetermined labels that may be regrouped by experts in a post-hoc way. Alternatively, others have proposed that hard cases can be detected [217] during annotation and rerouted [281] to other workers as needed.

4.2.3 Exploring Cases and Precedents

Advances in modern computing have also presented new solutions for those learning or practicing law [155]. One major advancement has been the increased capability of computer-assisted legal research. While early computer assistance mainly provided the ability to conduct keyword searches over legal databases [253], modern systems have seen many improvements that provide context aware or semantic search capabilities [179, 266, 127] under the assistance of improvements in machine learning and artificial intelligence.

Beyond the space of computer-assisted legal research, there have also been advances in the general area of open faceted and semantic search [47] that make use of the improved quality of similarity metrics based on embeddings [219]. These systems and approaches provide some of the building blocks we used to implement our prototypes for case law crowdsourcing.

4.2.4 Importance of Justifying Crowd Decisions

Alongside the increased demand for conducting human judgments on complex decisions at scale, there is also a renewed interest in making sure that crowd decisions are produced with quality in mind. one of the ways to accomplish that is by eliciting rationales [162] as an extension of just bare labels. Indeed rationales themselves can be a way of increasing the quality of crowd judgments [62]. For example, some crowdsourcing systems have explored the use of deliberation [77, 50] as a way of utilizing rationales to improve data quality.

However, deliberation can be costly and often needs some conditions to be fulfilled in order to be effective [236]. It can be easy to end up with a costly process [75] that provides limited utility.

Increased quality of data is not the only reason for collecting rationales. The act of collecting rationales can itself often provide a level of procedural legitimacy [86, 202]. For example, in content moderation settings, people want explanations for judgments [137] taken against content and it has been shown that providing these rationales in the form of explanations can lead to fewer reoffenses [138]. Benefits of rationales in decision quality can also be realized in these settings, allowing for consistent and repeatable judgments [123].

4.3 Design

In this section, we present a model of the workflow that we envision for a case law inspired crowdsourcing adjudication process. We first outline the general procedure and intermediate tasks (Figure 4.1) that each individual adjudicator performs to create their judgment on a novel case. Then, we dive into specific components of each intermediate task in the workflow and present the designs we developed to support scaling these tasks so that they can be completed by novice crowd workers with little training. Finally, we talk about how we can utilize the sets of precedent cases selected by crowd adjudicators to understand the judgments of the crowd as well as the bases these judgments are formed.

4.3.1 Case Law Crowdsourcing

The judgment procedure in case law crowdsourcing (Figure 4.1) draws from the idea of using *precedents* as a means to produce consensus decisions in legal systems based on common law. Simlar to common law jurisdiction where a small group of expert lawyers conduct legal research and analysis to produce an argument for their client based on statutes and precedents, in case law crowdsourcing, a crowd adjudicator assembles an argument for their annotation judgment on the new case based on selecting past decided annotations. As such, the case law crowdsourcing workflow consists of the following stages:

1. **Selecting the relevant context**: Annotation instances are decomposed, when pos-

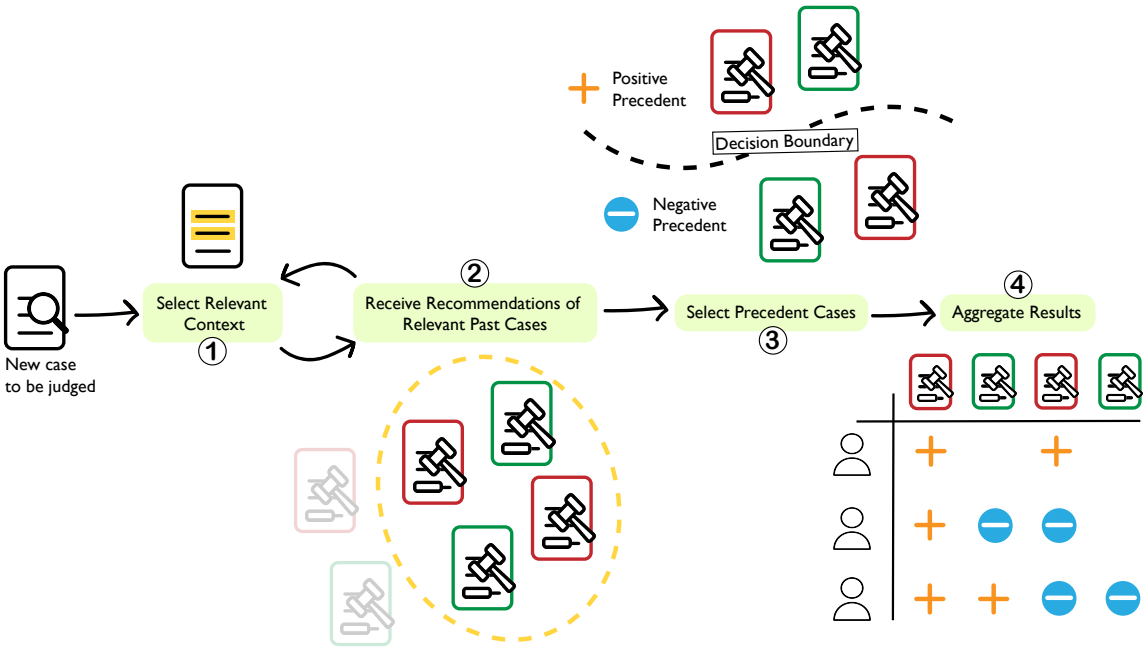


Figure 4.1: A diagram that illustrates the case law crowdsourcing workflow. After adjudicators are presented with a new case to be judged, they will first use the case exploration tool to explore potentially relevant precdents. Adjudictors can tune the recommended precedents by (1) toggle context from the judged case to be included or excluded. Then (2) based on the relevant past cases retrieved, adjudicators select cases (3) to construct two sets of precedents to support their judgment of the current case—positive precedent indicating a similar case or negative precedent indicating a distinct one. Once multiple adjudicators have constructed their judgments, results can be aggregated (4) surfacing any disagreements or ambiguities in the judgments. Details about each stage of the workflow are provided in Section 4.3.1.

sible, in a way that allows adjudicators to select what context in that instance they think is relevant to making their judgment.

2. **Receiving recommendations of relevant past cases**: Based on the relevant facts selected, past cases are found and recommended to adjudicators. Adjudicators use these cases to form an understanding of the decision bound around the judged case.

3. **Selecting precedent cases**: The adjudicator then selects the cases that they would like to use as precedent in anchoring the decision bound around *the current case*. This is done in the form of picking **positive precedents** (i.e. cases sharing similar context and judged following the same principles/reasoning) or **negative precedents** (i.e. cases sharing similar context but judged under different principles/reasoning) from *the current case*.

4. **Aggregating judgments across adjudicators**: Judgments from multiple adjudicators are compared and aggregated. Group judgments can then be derived by looking at which cases were commonly used by many adjudicators as precedents and what the past judgments were.

However, as case law crowdsourcing is built around utilizing non-expert crowd adjudicators to make decisions about lower stakes or subjective annotation judgments while under more resource constraints, aspects of traditional common law jurisdiction have also been simplified or relaxed. Our workflow makes a balance between providing support for judgments through precedent cases while also making sure the intermediate tasks are not too challenging by omitting aspects such as having adjudicators explicitly compose arguments.

4.3.2 Precedent Case Exploration

The first stage of the case law crowdsourcing workflow involves the *research* component. In this stage, adjudicators locate relevant past cases from which they learn the decision bounds as well as select precedents that are adopted to form their judgment. In our workflow, this takes the form of an interactive case exploration tool (Figure 4.2). The case exploration

tool presents the case being judged to the crowd adjudicator, alongside with a list of ranked relevant prior decisions. Past cases that are candidates for precedents are displayed as cards featuring both the case summary and any context. The final judgment of these prior cases is also displayed within the list.

The list of recommended prior cases are selected and ranked by comparing them against the case being judged using a similarity metric. Depending on the domain involved and how instances can be decomposed, the actual selection of this metric may vary. For example, in our experiments, one domain we used involves natural language text in the form of Reddit posts. In this case, each case (in the form of a post) can be decomposed into sentences that describe the case background. Considering this, we selected a similarity metric based on distances in a text embedding space. We used DistilBERT [231] to compute sentence embedding vectors that were then aggregated into an overall document embedding for each case. We then used cosine similarity as the metric to compare cases.

In addition to the list of cases selected by default, the case exploration tool also facilitates active exploration in the space of past cases directed by the crowd adjudicator. Since not all aspects of a case will be relevant for the judgment, similarity scores that use the entire context of a case can result in candidates that have superficial similarities (e.g. mentioning specific names or locations that are not relevant to a decision) but aren't meaningfully related. To address this, the case exploration tool allows adjudicators to "turn off" irrelevant contextual facts by excluding them from the similarity metric. In this case of text based tasks, crowd adjudicators can click on a sentence to toggle whether it is used or ignored for the purposes of similarity comparison. The adjudicator can also progressively expand the space explored around past cases by asking for cases similar to an existing one ("more cases like this") or sampling cases along a particular judgment ("more ¡positive/negative¿ cases"). This exploration process is enabled throughout the judgment process, so as crowd adjudicators assemble the cases that make up their judgment, the can also come back at any time to adjust the criteria and sample more cases in the neighborhood.

4.3.3 Assembling Sets of Positive and Negative Precedent Cases

As crowd adjudicators explore the space of precedent cases, they also work on completing their main task of assembling their judgment. However, unlike traditional approaches in crowdsourced argumentation where adjudicators are asked to provide their judgment along with an argument to support it [77], our workflow takes advantage of the fact that precedent cases have an existing associated judgment so adjudicators need only find the cases to adopt as precedent rather than making their judgment first and writing arguments to support it.

In order to construct their argument (and corresponding judgment), adjudicators are asked to focus on categorizing past cases discovered through the precedent case exploration stage, based on their relevance to the current judgment. The goal of adjudicators in this stage is to find cases based on the following:

- **Positive Precedents**: A past case should be marked as a "positive precedent" if the crowd adjudicator believes that the same principles or evidence leading to the past decision can be adopted for the new case being judged. These are cases that fall on the same side of the decision bound of the judged case. We instruct crowd adjudicators to prioritize finding these types of cases where past decision can be adopted if possible.

- **Negative Precedents**: A past case can be marked as "negative precedent" if the crowd adjudicator believes that despite being in the neighborhood of similar cases, some principle(s) or evidence leading to the past decision does not apply to the current case. Negative precedents don't necessarily have final decisions that would be different from the current case being judged. Instead, a negative precedent indicates that some reasoning or principles leading to the precedent don't apply to the current case.

- Cases can also be left uncategorized. This encompasses cases that are either not relevant, don't share principles/criteria with the current case.

Crowd adjudicators use an "binning" interface (similar to that of organizing a folder) where they can curate these sets of cases either via drag-and-drop or through buttons presented on each case. Each case can only be placed in one of these sets, and once a

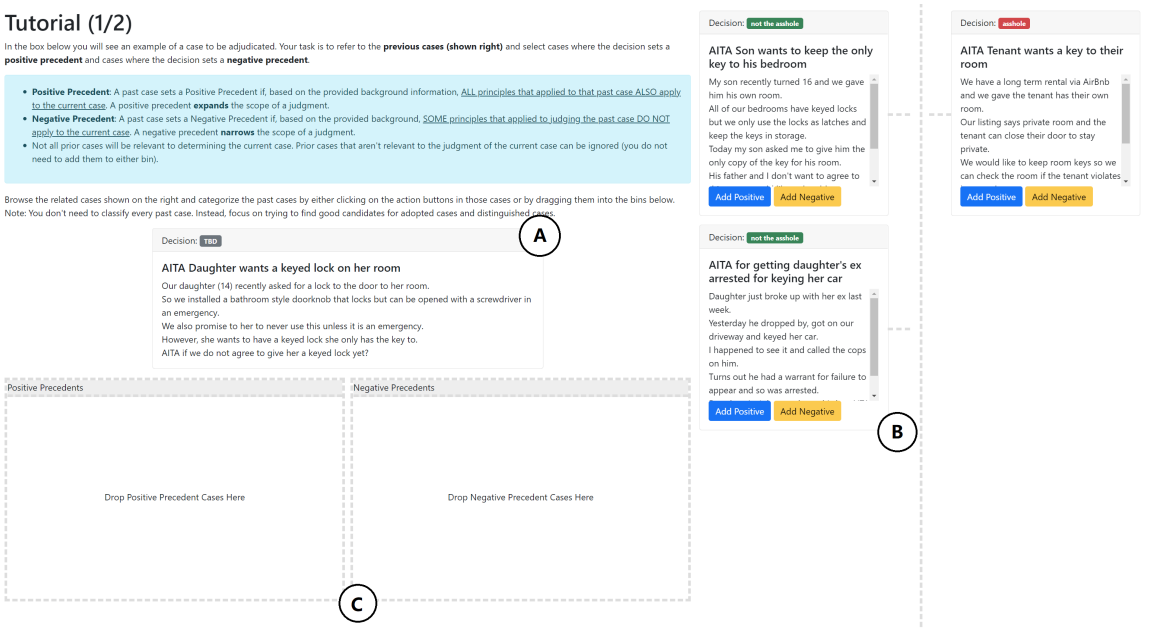


Figure 4.2: A screenshot of the case exploration tool prototype currently showcasing the example used for training. (A) Shows the case that is currently being judged. (B) Shows the candidate cases recommended by the case exploration tool. (C) Shows the area that organizes the positive and negative precedents. Cases can be added/removed either with the action buttons or through drag-and-drop.

case is added to a set as an "positive" or "negative" precedent, it will be omitted from any recommended lists of potential precedent cases.

When assembling these sets, adjudicators are prompted to aim for finding at least one "positive" precedent case. However, as is the case in common law judicial systems, it is possible for adjudicators to fail to find relevant precedents for a case that is novel. In these cases, adjudicators are allowed to submit their judgment even if it consists only of "negative" precedent cases without a positive case.

4.3.4 Interpreting Judgments Individually and in Aggregate

Once individual adjudicators have assembled their sets of precedent cases, the sets of precedents can be interpreted as judgments on the case. This allows our workflow to produce

both traditional classification judgments of a case while also enabling the observation of any disagreements. To produce the final judgment for an adjudicator, we can observe the cases that they selected as positive precedents. As these cases indicate situations that the adjudicator agrees matches that of the current case being judged, we can apply the precedents by adopting the judgment of these positive precedent cases as the judgment that should also be applied to the current case. If an annotator did not find any positive precedents, that indicates that they believe the current case to be in a situation where the decision bounds needed to make a judgment are not clear and thus a judgment can't be reliably made.

In addition to producing these simple final judgments, positive and negative precedents also allow us to interpret the degree of agreement between adjudicators. Disagreement between adjudicators can arise as 2 types: (1) omissions, where a past case selected as a precedent by one adjudicator was not selected at all by another; and (2) conflicts, where a past case was selected as a positive precedent by one adjudicator and a negative precedent by another. The first type of disagreement provides a signal about the ambiguity around the precedent – the more often it is selected the more confident the group is that the past case is sets a good precedent. The second type of disagreement provides a signal about how different annotators may disagree on their interpretation of the case. We note that, the presence of a conflict in the precedents won't always result in observed disagreement of the judgment on the case. For example, one adjudicator might select two past cases with the same judgment as both being positive precedents while another may believe one of these cases relies on principles that don't generalize to the new case. In this situation, even though the final judgment may appear to agree, different adjudicators actually have different interpretations of the decision bounds for the task.

4.3.5 Prototype and Implementation Details

We implemented a prototype of this case law crowdsourcing workflow in the form of a browser-based annotation tool that can be deployed on a crowdsourcing platform like Amazon Mechanical Turk (Figure 4.2). Details about each case being judged and any previously judged cases are managed in a backend service. Crowd workers can make adjustments to

the context used to recommend relevant cases by clicking on the details (Figure 4.2, (A), each sentence is a togglable unit) to toggle them on/off. When the context is adjusted, the toggle state is sent to the backend, which constructs a transient 'synthetic case' with only the enabled details, and computes a document embedding vector for this case. We then compute the cosine similarity score between the new embedding vector and the existing cases to rerank the set of candidate precedents, which are then sent over as the updated recommendations.

4.4 Experiments

We conducted two experiments to evaluate aspects of the judgments produced through case law crowdsourcing. Specifically we were interested in answering the following research questions:

- **RQ1**: Are group judgments more consistently produced when grounded by precedents in case law crowdsourcing?

- **RQ2**: Are we able to distinguish situations where precedents are insufficient to ground complex decisions?

4.4.1 Experiment Setup

To answer these questions we designed two annotation experiments. For our first experiment, we explore the consistency aspect of case law crowdsourcing by recruiting annotators to judge a set of cases under one of two conditions. In the CONTROL condition, annotators are provided with traditional instructions and training based on guidelines and fixed examples. For each case, annotators provided a single judgment in the form of a class label (*not the asshole* or *asshole*). In the CASE-LAW condition, annotators are instead given access to the case exploration tool (Figure 4.2). Using this tool, annotators are asked to construct sets of positive and negative precedents which then form their judgment on the case. We then compare the amount of disagreement between labels produced in each condition to evaluate the level of consistency across annotators.

For our second experiment, we explored whether the case law crowdsourcing is sensitive to identifying situations where good precedent candidates are not available. To construct scenarios where precedents may be insufficient, we used a sentence embedding-based metric to find and exclude the top 10 most similar cases in the neighborhood around each case being judged. We then conducted annotation using the case law crowdsourcing workflow using this new set of cases as potential precedents. We compared the sets of positive and negative precedents produced in this INSUFFICIENT-PRECEDENT condition against the results from the CASE-LAW condition to evaluate whether there is an observable difference when the precedents are insufficient to ground the decision bounds.

4.4.2 Task Domain

For our experiment data, we drew examples from a dataset [199] of posts collected from `r/AmITheAsshole`, a subreddit featuring community posts and judgments about interpersonal conflict scenarios. Posts in this subreddit usually feature a description of a real world scenario or situation where two parties (one of which is often the author) are in conflict over an action or matter. Context and background information that the author believes is relevant is also included, often in great detail, in the body of these posts. Community participants then judge whether the author's actions are considered ethically acceptable given the situation. The task of making judgments on interpersonal conflicts presents a representative instance of the type of domain where case law crowdsourcing processes can provide potential benefits—in this case individual cases are complex and the exact criteria for making judgments is difficult to summarize into a set of comprehensive guidelines. Cases in this dataset were categorized under 1 of 4 judgment types: *not the asshole*, *asshole*, *everyone sucks*, or *no assholes here*.

For our experiments, we focused on subset of instances sampled from instances where consensus judgment was either *not the asshole* or *asshole*. We also filtered out any cases that contained too few details (one or fewer sentences) or other text inconsistencies due to data extraction artifacts. Cases were then truncated to contain 10 sentences of background. From these cases, we randomly sampled 25 instances (divided into 5 groups of 5 cases) to

serve as target cases to be judged by the crowd. Remaining cases in the sampled subset served as the pool from which we drew candidates for precedents in the case exploration tool.

4.4.3 Participants and Recruitment

We recruited crowd annotators through the Amazon Mechanical Turk (AMT) platform. For each group of 5 cases in each condition, we recruited 5 unique annotators to conduct the annotation resulting in a total of 75 annotators. In all conditions, we aimed for a $16/hr pay for the crowd annotators. Participants were paid $2.00 as base payment for completing the training, followed by a bonus after completing the annotations. In the CASE-LAW and INSUFFICIENT-PRECEDENT conditions, participants were paid an additional $6.00 bonus for completing the 30 min annotation task to create the sets of positive and negative precedents through the case law crowdsourcing workflow. In the CONTROL condition, participants were paid a $2.00 bonus for completing the 15 min annotation task to directly select the judgment class (*not the asshole* or *asshole*) based on text guidelines.

4.4.4 Results

RQ1: Consistency

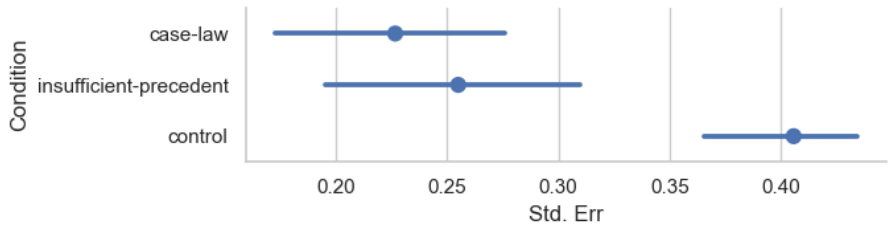


Figure 4.3: Figure shows the consistency of judgments (as measured by standard error) across annotators under all 3 conditions. We observe that conditions based on case law adjudication resulted in higher consistency reflected through lower standard error.

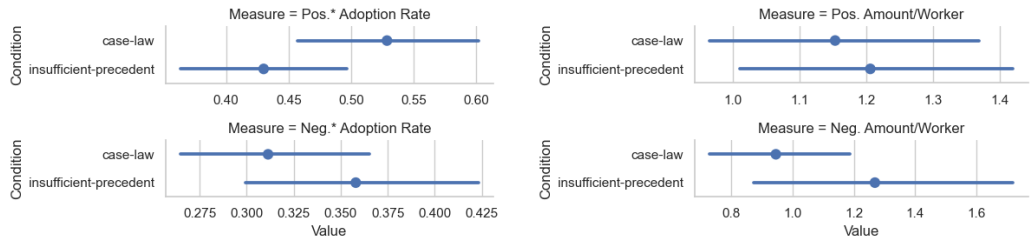
For our first experiment, we examined whether the judgments produced through the

CASE-LAW condition were more consistent compared to those produced from directly collecting judgments through the CONTROL condition. As adjudicators in the CASE-LAW condition produce sets of positive and negative precedents instead of directly producing a judgment, we use the precedents to interpret what the judgment of the annotator would be (as noted in Section 4.3.4). Specifically, for each adjudicator in the CASE-LAW condition, we look at the cases in their set of positive precedents. If the prior judgments of all positive precedents is unanimous, then we adopt that judgment as the annotator's judgment of the new case. If prior judgments of the set of positive precedents is not unanimous, then we interpret the annotator's judgment as *undetermined*. We use the disagreement between annotators to evaluate how consistently judgments can be produced. To evaluate disagreement, we generalized [27] the 3 possible categorical judgment types that can be produced by each annotator on a scale by assigning judgments of *not an asshole* as 1, *undetermined* as 0, and *asshole* as -1, noting that the uncertain *undetermined* case falls in between the two extremes. We found that the mean disagreement as measured through standard error was 0.227 for the CASE-LAW condition and 0.406 for the CONTROL condition (Figure 4.3). As the target cases annotated were the same in both conditions, we use a paired t-test to compare across conditions, observing that this difference is statistically significant at $P < 0.01$.

We note that this observed result falls within our expectations. The task domain of adjudicating interpersonal conflict scenarios is inherently complex and comes with a set of complex criteria that can be challenging to capture with traditional annotation guidelines and examples. While guidelines can be helpful in generalizing decisions across higher-level concepts (such as whether the actions in the scenarios was ethical), the complexity involved in the context and background of each interpersonal conflict case means that most cases will likely also need more specific guidance.

RQ2: Detecting Insufficient Precedents

For our second experiment, we examine whether the potential problem of precedents that are insufficient for informing judgments can be detected through our case law crowdsourcing workflow. As in case law legal systems, while the consistency afforded by adopting prece-



(a) Adoption rate (% annotators who included) of the most common case in the Pos./Neg. sets across annotators.

(b) Amount of cases selected per annotator precdedent Pos./Neg.

Figure 4.4: Comparing properties of the Pos./Neg. precedent sets created by annotators in the CASE-LAW and INSUFFICIENT-PRECEDENT conditions.

dents for new judgments can be desirable, blindly applying a precedent that does not share circumstances with the case being judged can also lead to a consistent but an incorrect or biased judgment. We conducted this experiment to evaluate whether we can identify and correct from situations where precedents start to fail in grounding decisions and it may be desirable to fall back to an alternate process to judge a novel case. Specifically, we look at whether reducing the relevance of precedents around the target cases results in measurable differences in the sets of positive and negative precedents created by crowd adjudicators.

We compared the sets of positive and negative precedents produced between the CASE-LAW condition (where the regular precedents were available), against the sets produced in the INSUFFICIENT-PRECEDENT condition (where precedents relevant to judging target cases were intentionally removed to create an ambiguous condition), focusing on 2 aspects: (1) Is there a difference in the *consistency* of cases selected by different adjudicators, and (2) Is there a difference in the *amount* of cases adjudicators end up selecting for each set of precedents.

As adjudicators can potentially select any case as a precedent, we evaluate the consistency of precedents by focusing on the case that was *most commonly* included as a precedent, looking at the positive and negative precedent sets separately. For these cases, we

then looked at the proportion of adjudicators who did include the case as a precedent. By focusing on the consistency of agreement in the 'most common' precedent, we can account for the inherent base differences in quality of available precedents. Comparing between the CASE-LAW and INSUFFICIENT-PRECEDENT conditions, we found that when looking at the case that was most commonly included in the set of positive precedents, an average of 52.8% of adjudicators would include this case under the CASE-LAW while only a 43.0% of adjudicators included this case in the INSUFFICIENT-PRECEDENT condition (Figure 4.4). Even stratified across each individual case judged, we found around a similar margin of difference between the adoption rate of this most common precedent, with this result being statistically significant under a paired t-test with $P \approx 0.0046 < 0.01$. We did not see a similar gap in the adoption rate of the most common precedent for the negative precedent set, with an average adoption rate of 31.1% and 35.8% for the two conditions respectively (no significant difference).

Looking at the *amount* of cases adjudicators end up selecting for each set of precedents, we also found some differences. On average, adjudicators picked around the same amount of cases for their positive precedents, at 1.15 cases/adjudicator in the CASE-LAW condition and 1.21 cases/adjudicator in the INSUFFICIENT-PRECEDENT condition. However, crowd adjudicators tended to pick more negative precedents when the relevance of precedents was lower: 1.27 case/adjudicator in the INSUFFICIENT-PRECEDENT compared to 0.95 cases/adjudicator in CASE-LAW. However, we note that this difference in *amount* of cases was not statistically significant under a paired t-test, with $P \approx 0.08 > 0.05$.

Based on these observations, we find that the case law crowdsourcing workflow does exhibit different behaviors depending on the relevance of the precedents provided, with a more reliable signal coming from the evaluating the consistency between adjudicators in choosing positive precedents. However, we also note the observation that, regardless of the relevance of precedents, adjudicators ended up picking similar amounts of cases for their positive precedents. This hints that under these settings adjudicators can be biased to aim for a certain amount of positive precedents even when it is not required. This implies that unless care is taken to detect the case of poor precedents, the quality of positive precedent sets constructed individuals can be reflective of the overall quality of candidate cases for

precedents.

4.5 Discussion

In the sections above, we have presented the design of our case law crowdsourcing workflow and conducted evaluations to validate its utility. In this section, we will discuss some of the limitations that arise from using case law crowdsourcing and potential solutions for future work.

4.5.1 Balancing Guidelines and Precedent Cases

While precedents can be useful in defining challenging local decision bounds, guidelines can still be more efficient in situations where the decisions are simpler and can be conveyed easily in a summarized form. Indeed in even in many challenging real world tasks with complex decision bounds, such as content moderation [273, 161], misinformation identification [25], and data annotation [305], not all decisions will be challenging and there will often be sets of cases where the decision bounds that separate them are significantly less complex than those around the edge cases. Additionally, while case law crowdsourcing presents a way to improve consistency when past judgments exist, a certain level of bootstrapping still needs to be performed first to create the first judgments that will become those future precedents.

Considering these aspects, it is important to find a balance between taking advantage of the efficiency provided by traditional instructions and guidelines, and improving the decisions on the edge. One direction to address finding this balance may come from systems that make use of measurements on the consistency of annotations under traditional guidelines [213] to identify problem areas that could then be redirected to a case law crowdsourcing process.

4.5.2 Overruling and Reversing Precedents

One property that comes with adopting the metaphor for precedent cases from the legal space is the eventuality that precedent judgments may one day be overruled or reversed. Indeed, within the space of case law, it is common for new judgments to set precedents

that overrule earlier rulings or for past case judgments to be reversed. In the realm of tools for computer assisted legal research, this problem is often solved through the process of shepardizing [1] where automated tools can assist in determining whether a precedent has been overruled or is still "good law".

However, while changes in precedent decisions can present challenges to cases built on them, the use of precedents also enables the potential for automated tools to address these changes, even for case law crowdsourcing. We note that the sets of negative and positive precedents produced in the case law crowdsourcing process not only serve to inform judgments, but can also be a form of referencing. By constructing and traversing the graph over the chains of references produced through precedent sets, the case law crowdsourcing workflow can also adapt to future instances where precedents can be overruled or reversed, providing additional value. Unlike datasets constructed from direct judgments, which must be re-annotated should criteria change, sets of judgments made in the form of precedent cases enables affected judgments to be easily identified in case of an overruled or reversed case, thus reducing the amount of re-annotation effort needed to maintain consistency.

4.5.3 Resolving Conflicts Between Adjudicators

Finally, while the workflow we present in case law crowdsourcing provides a means for conflicts to be identified through comparing each individual adjudicator's choice of precedents, in many situations it can also be desirable to resolve conflicts and disagreements and produce consensus decisions. However, depending on the task being judged, the most adequate way to address conflicts can differ. For example, in content moderation, prior work has shown that the perceived legitimacy of the resolution process can affect community acceptance of the decision [202] so process legitimacy may be an important consideration in this case. Additionally, who takes part in resolving conflicts can also vary. While in some situations it may be useful to have adjudicators resolve conflicts among themselves [46, 50], for certain annotation tasks, experts like task requesters may have specific goals in mind and may end up having the final say in dictating how conflicts should be resolved [35]. Given the large

[1] `https://www.lexisnexis.com/documents/LawSchoolTutorials/20081015085048_large.pdf`

variability, we don't attempt to prescribe any particular way of resolving conflicts, instead focusing on providing insight into different types of conflicting judgments.

4.6 Conclusion

In this paper, we present a novel approach to human judgment on complex tasks in the form of case law crowdsourcing. Through experiments, we demonstrate that case law crowdsourcing produces more consistent judgments compared to directly collecting judgments while also providing rationale in the form of positive and negative precedents.

Chapter 5

CICERO: ADDRESSING UNCERTAINTY BY RESOLVING DISAGREEMENT THROUGH DELIBERATION

In the previous two chapters, we presented tools and workflows for human judgment with the crowd that allows us to improve the consistency of judgments and capture uncertainty. However, how can we address uncertainty once we know it's there?

In this chapter, we present Cicero, a new workflow that improves crowd accuracy on difficult tasks by engaging workers in multi-turn, contextual discussions through real-time, synchronous argumentation. Our experiments show that compared to previous argumentation systems which only improve the average individual worker accuracy by 6.8 percentage points on the Relation Extraction domain, our workflow achieves 16.7 percentage point improvement. Furthermore, previous argumentation approaches don't apply to tasks with many possible answers; in contrast, Cicero works well in these cases, raising accuracy from 66.7% to 98.8% on the Codenames domain.

5.1 Introduction

Crowdsourcing has been used for a wide variety of tasks, from image labeling to language transcription and translation. Many complex jobs can be decomposed into small micro-tasks [177, 22, 197, 54]. After such decomposition, the primary challenge becomes ensuring that independent individual judgments result in accurate global answers. Approaches ranging from aggregation via majority vote [248] to programmatic filtering via gold-standard questions [200] have all been created to achieve this goal. Further improvements have led to more intelligent aggregation such as expectation maximization (EM) [65, 290, 287]. However, EM may still fall short, especially on hard problems where individual judgments are unreliable. Indeed, some researchers have concluded that crowdsourcing is incapable of achieving perfect accuracy [66].

Yet recently, *argumentation* has been shown to be an effective way to improve the accu-

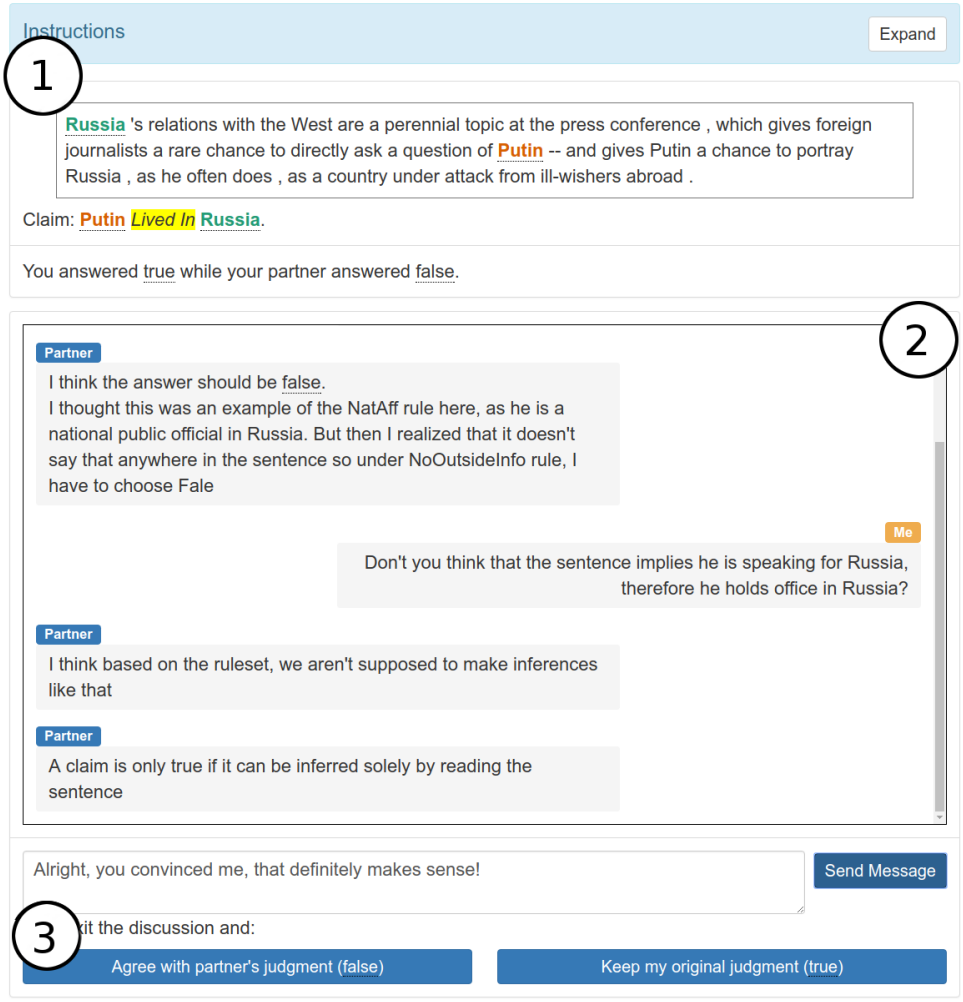


Figure 5.1: Discussion interface for use in CICERO, inspired by instant-messaging clients, showing a fragment of an actual discussion in the Relation Extraction domain. (1) Presents the question (sentence + claim) and both sides' beliefs. (2) Initial discussion is seeded with the workers' justifications. (3) Options added to facilitate termination of a discussion once it has reached the end of its usefulness.

racy of both individual and aggregate judgments. For example, Drapeau *et al.*'s MicroTalk [77] used a pipelined approach of: 1) asking crowd workers to *assess* a question's answer, 2) prompting them to *justify* their reasoning, 3) showing them counterarguments written by other workers, and 4) allowing them to *reconsider* their original answers to improve individual judgments. In principle, this simplified form of argumentation allows a single dissident worker, through force of reason, to steer others to the right answer. Furthermore, the authors showed that argumentation was compatible with EM; combining the two methods resulted in substantial gains in accuracy.

However, while asynchronous argumentation systems like MicroTalk attempt to resolve disagreement, the steering power of a one-round debate is limited. Workers are only shown a pre-collected justification for an opposing answer; they aren't challenged by a specific and personalized argument against the flaws in their original reasoning. There is also no back-and-forth interaction that could illuminate subtle aspects of a problem or resolve a worker's misconceptions—something which may only become apparent after several turns of discussion. Furthermore, since justifications are pre-collected, workers need to write a generic counter argument; while this works for binary answer tasks, it is completely impractical for tasks with many answers; such a counter-argument would typically be prohibitively long, refuting $n - 1$ alternatives.

This paper presents CICERO, a new workflow that engages workers in *multi-turn and contextual* argumentation to improve crowd accuracy on difficult tasks. CICERO selects workers with opposing answers to questions and pairs them into a discussion session using a chat-style interface, in which they can respond to each other's reasoning and debate the best answer (Figure 5.1). During these exchanges, workers are able to write context-dependent counter-arguments addressing their partner's specific claims, cite rules from the training materials to support their answers, point out oversights of other workers, and resolve misconceptions about the rules and task which can impact their future performance on the task. As a result of these effects, workers are more likely to converge to correct answers, improving individual accuracy. Our experiments on two difficult text based task domains, relation extraction and a word association task, show that contextual multi-turn discussion yields vastly improved worker accuracy compared to traditional argumentation

workflows.

In summary, we make the following contributions:

- We propose CICERO, a novel workflow that induces multi-turn and contextual argumentation, facilitating focused discussions about the answers to objective questions.

- We introduce a new type of worker training to ensure that workers understand the process of argumentation (in addition to the task itself) and produce high quality arguments.

- We develop CICERO-SYNC, a synchronous implementation of our workflow using real-time crowdsourcing, and apply it to conduct the following experiments:

  - In the Relation Extraction domain introduced by MICROTALK [77], we show that contextual, multi-turn argumentation results in significantly higher improvement in accuracy: a 16.7 percentage point improvement over individual workers' pre-argumentation accuracy *v.s.* a 6.8 point improvement using MICROTALK's one-shot argumentation. When aggregating the opinions of multiple workers using majority vote or EM, we see 5 percentage points higher aggregate accuracy, accounting for cost.
  - Using a version of the Codenames domain [309], that has many answer choices (making MICROTALK's non-contextual argumentation untenable), we show that CICERO is quite effective, improving individual worker accuracy from 66.7% to a near-perfect 98.8%.
  - We qualitatively analyze the discussion transcripts produced from our experiments with CICERO-SYNC, identifying several characteristics present in contextual, multi-turn argumentation.

5.2 Cicero Design

In this section, we present the CICERO workflow as well as design considerations in a synchronous implementation of the workflow used for our experiments. We first explain the

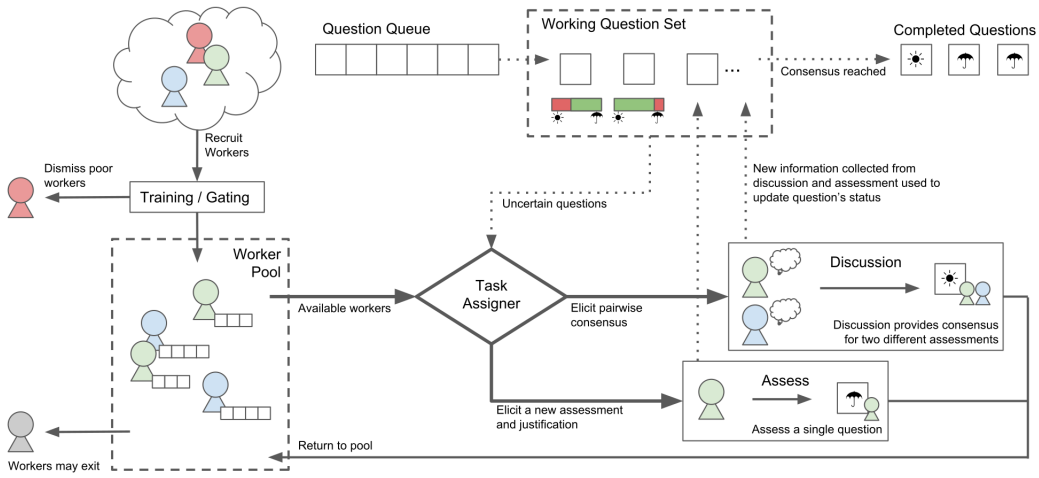


Figure 5.2: CICERO System Diagram. Solid arrows indicate paths for workers through the system. Dotted arrows indicate how questions are allocated through the system.

rationale for contextual, multi-turn discussions and give an overview of our CICERO workflow. We then talk about the decision to implement our workflow in a synchronous system—CICERO-SYNC. Finally, we discuss the design choices we made to (1) create an interface for effective real-time discussion, as well as (2) improve instructions and training for the domains we examined.

5.2.1 Contextual and Multi-Turn Discussion

In natural forms of debate, participants who disagree take turns presenting arguments which can refute or supplement prior arguments. Our CICERO workflow is designed around the concept of emulating this process in a crowd work setting by using paired discussions facilitated by a dynamic matching system. Participants are matched with partners based on their current beliefs and are encouraged to present their arguments over multiple turns.

While real-life debates may include multiple participants each responsible for addressing arguments on different aspects of a problem, in the crowd setting we can utilize the diversity of workers to cover a broad set of views and reasoning; thus, to simplify the process, we

focus on a two-participant discussion model.

5.2.2 Workflow Overview

Since argumentation happens on an ad-hoc basis, it's much more flexible to have our workflow focus on managing transitions between different states a worker may be in instead of defining a single pipeline. Due to this, our design of the CICERO workflow follows an event-based definition model where the automatic task assigner allocates tasks as workers' state changes. Figure 5.2 summarizes how our workflow allocates worker resources and questions in a dynamic way.

Initially, workers are recruited from a crowd work platform (such as Amazon Mechanical Turk) and are immediately assigned to a **training** task. Workers who pass training and the associated gating tests [178] enter the *worker pool* and wait to be assigned work. Then, instead of a fixed workflow, our event-based automatic task assigner decides which type of task and question to assign to a worker subject to a set of constraints. As workers complete their tasks and update the beliefs of questions in the working set, new candidate tasks are dynamically selected and allocated. Cicero's dynamic matching engages workers across diverse pairings, which has been shown to promote better output in large creative tasks such as in Salehi *et al.* [228].

In CICERO, there are two main types of tasks that the automatic assigner may assign to an idle worker: **assess** and **discussion**.

- The **assess** task acquires one worker from the worker pool who is then presented with one question — in our case a single question in the domain — that asks for an answer to a multiple choice question and optionally a free-form justification for their position. This task is a combination of the assess and justify microtasks in MicroTalk [77] as a single task.

- The **discussion** task acquires two workers from the worker pool who are both shown a discussion interface for a question. At the end of a discussion, the justification text may be updated for both workers. This task is a multi-turn, contextual version of the

reconsider task in Microtalk [77], which actively engages both workers. We will cover details on the design of the discussion task in later sections.

The automatic task assigner is defined as a policy that decides which type of task should be allocated when a worker changes their state (such as upon completing a micro-task) and, depending on domain, can be designed to prioritize specific kinds of tasks, particular questions or qualities such as minimizing worker wait time and increasing concurrent work.

In general, the task assigned can be adapted to the goals of the requester. However, there are a few general constraints that the task assigner must follow:

- **Incompatible beliefs**: A discussion may only be assigned to workers if they have incompatible beliefs. Implicitly, this also requires existence of the both beliefs, implying they must have been collected (*e.g.*, via assess tasks).
- **No repeated discussions**: Two workers may only discuss a question if they have never discussed the question with each other before.

These constraints guarantee that the workflow will eventually terminate when there are no more workers who disagree and have never paired with each other.

There are many benefits to dynamically allocating partners. Since pairings aren't fixed, CICERO can automatically adapt to existing workers dropping out and new workers entering the pool. Additionally, in contrast to previous systems [77, 237], CICERO's automatic task assigner sequentially exposes each worker to discussions with multiple partners for a particular question. This allows for the possibility of a minority opinion reaching and convincing the majority. A worker who is convinced by a minority belief is able to spread the new answer as they may now be matched with workers they used to agree with, increasing the size of the minority.

5.2.3 CICERO-SYNC*: A Real-Time Implementation*

While the CICERO workflow does not constrain the type of interaction during a discussion task, we decided to test out the effectiveness of our workflow using synchronous discussions.

A synchronous and real-time discussion environment allows us to mimic real world continuous dialogue spanning many turns thus preserving discussion context in a simple and natural way.

In CICERO-SYNC, workers are held in a waiting room until a partner becomes available. Once workers are matched into a discussion, they will not be assigned other tasks for the duration of that discussion and are expected to give each other their undivided attention. We note that, while useful for experiments, this design has limitations: the synchronous nature of discussions means that some workers will have to wait for a partner to become available and workers need to be online and active within the same time window, both of which imply a higher cost to the requester.

Additionally, there are many practical challenges to implementing and setting up synchronous real-time experiments with crowd workers, including implementing real-time client-server communication and working with APIs for worker recruitment and payment [126]. Fortunately, there have been enough real-time, crowd deployments [22, 26] that many useful lessons have been distilled [125]. We elected to use TurkServer [183], whose tools simplify the interfacing with Amazon Mechanical Turk for worker recruitment and task management and allow us to automatically track worker state as well as building our worker pool (Figure 5.2) using the TurkServer *lobby*.

5.2.4 Discussion Interface

The discussion task is the most important and defining task of the CICERO workflow. We considered multiple different options for the discussion interface focusing on ways to organize discussion structure and facilitate discoverability.

Early proposals included designs that were inspired by the posts-and-replies interfaces in social network timelines and the split-view pros-and-cons interfaces used in ConsiderIt, a political, argumentation system [156]. Our pilot studies showed that these methods were cumbersome and non-intuitive, so we decided on a free-form instant messaging (chat) metaphor for the discussion task (shown in Figure 5.1).

When a pair of workers enter a discussion, they are placed into a familiar instant mes-

saging setting, where they can freely send and receive messages. Each message is tagged with being either from the worker themself ("me") or their unnamed partner ("partner"). An additional "exit" section below the chat interface allows either participant to terminate the discussion if they feel that it is no longer useful. Workers can utilize this exit mechanism to indicate that a consensus was reached or that no agreement is possible between them.

The discussion interface can be easily adapted to specific needs of each experiment domain: In the Relation Extraction domain, the justifications collected from earlier assess or discussion tasks are used to seed the system, which we found to be beneficial in starting a conversation. In the Codenames domain, a drop-down menu below the text input field accommodates switching to alternate answers during the discussion addressing the non-binary nature of the questions.

5.2.5 Optimizing Task Instructions

Good instructions are essential for high inter-annotator agreement [178]. We observed in early pilot experiments that arguments which refer explicitly to parts of task guidelines were more effective at convincing a partner. However, the original task guidelines and training did nothing to encourage this practice. Workers came up with different ways to refer to parts of the instructions or training examples, but this was inconsistent and frequently caused confusion. References to the guidelines were hard to identify making it harder for workers to determine correct invocations of rules in the Relation Extraction domain pilots. Since arguing in synchronous discussion sessions is time-sensitive, creating rules and shorthands that are easy to cite is important for discussion efficiency.

We adjusted the task guidelines for the Relation Extraction domain from those in MicroTalk, re-organizing them into five concrete and easy-to-cite rules as shown in Figure 5.3. Each rule was given a shorthand so that workers can unambiguously refer to a specific rule and aid in identification of proper or improper rule usage during the discussions. We observed that citing behavior became more consistent within discussions with workers frequently utilizing our shorthands in the discussion context. In the Codenames domain, which has simple instructions but a lot of example cases, we designed the instructions

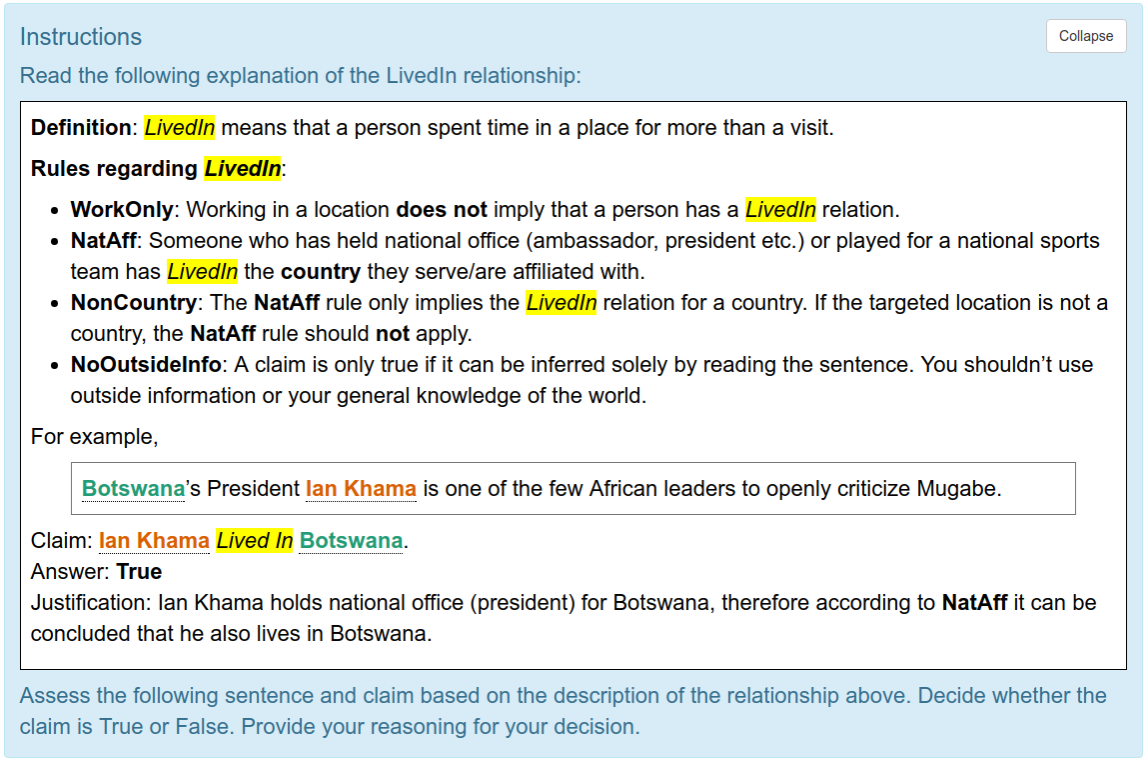


Figure 5.3: Screenshot of our *LivedIn assess* task (Relation Extraction domain) instructions containing 5 citable rules including the definition. Shorthands (in bold) allow for efficient citation of rules during discussion and within justifications (as shown in the example's justification).

to both show the general guidelines and also provide a way for workers to review examples from training if they decide to reference them.

5.2.6 Selecting and Training Effective Workers

In initial pilots with CICERO-SYNC, we noticed that workers were performing inconsistently. Following Drapeau *et al.*, we tried filtering for "discerning workers" using the Flesh-Kincaid score [136] to eliminate workers whose gold-question justifications were poorly written; to our surprise, this did not increase worker quality, but it did substantially reduce the number of possible workers. Filtering workers based only on gold standard question performance was also ineffective as it did not train workers to understand the rules required for our complex tasks.

Instead, we implemented a gating process [178], that can both train and select workers at the same time. Workers are presented with questions laid out in a quiz-like format. Each training question is provided along an introduction of related concepts from the task instructions. The questions are interleaved with the instructions in an interactive tutorial where new questions are presented as new concepts are introduced to reinforce worker understanding. Automated feedback is given when a worker selects an answer. At the end, workers' performance on a set of quiz questions is recorded. If a worker's accuracy on the quiz falls below a certain threshold, the worker will be asked to retry the training section (a limited number of times) with the order of the quiz questions randomized. Workers are dismissed if they exceed the retry limit.

5.2.7 Selecting and Training Effective Argue-ers

In existing argumentation systems [77, 237], training focuses on the target task instructions, however, not all kinds of arguments are productive. Argument forms and norms that contribute to positive discussion have been studied in the education community, termed 'accountable talk' [190]. During our pilot studies, we found that many workers' arguments weren't accountable, and realized that we need to train workers *how to argue* in order to ensure that discussions between workers are productive. To address this, we designed a novel

justification training task incorporated as a part of the training process to train the workers to recognize good justifications and arguments before they interact with a partner.

In this training task, workers encounter a sample assess task, followed by a justification-like task where, instead of a free-form justification, workers are asked to select the best one from a list. We then provide feedback in the form of an argument for why a justification is better or worse with reference to the task rules. By undergoing this training, workers are exposed to both how to think about justifications and what an effective counter-argument can be.

In the Relation Extraction domain, specifically, each incorrect option targets a potential pitfall a worker may make when writing a justification, such as: failure to cite rules, incomplete or incorrect references to the rules, or making extended and inappropriate inferences. In the Codenames domain, questions can have ten or more possible answers, so it's not practical to create and present multiple justifications for all of them. Therefore, the training is adjusted to instead show a reference counter-argument when a worker selects an incorrect answer that refutes the incorrect choice and supports a correct one. Our sample questions are designed to illustrate different argumentation strategies in different situations as the rules in this domain are simpler.

We note that this design of exposing the concept of arguments to workers during training can be generalized to many domains by providing feedback in the form of counter-arguments. By training workers to recognize and analyze arguments (before they enter a live discussion), our justification training promotes more critical discussion.

5.2.8 Worker Retention and Real-Time Quality Control

Due to the synchronous nature of discussions in CICERO-SYNC, workers may become idle for short periods of time when they are waiting to be matched to a partner. To ensure that idle workers in the *worker pool* are available for future matching, we implemented a real-time *lobby* design where workers wait while a task is assigned. This design was mainly inspired by both the default lobby provided in TurkServer [183] and from a worker-progress feedback design developed by Huang *et al.* [125] for low-latency crowdsourcing. While in the lobby,

workers are presented with information on their peers' current status, such as how many workers are currently online and which workers may become available soon. Workers also see statistics on their work, which is tied to bonus payments, and are encouraged to wait. In CICERO-SYNC, the task assigner is configured to immediately assign work as it becomes available. While in the lobby, a worker can voluntarily exit with no penalty if either their total waiting time exceeds a preset threshold or if they have completed a sufficient number of tasks (a single discussion in CICERO-SYNC).

In addition, while our gating process is designed to select workers serious about the task, we do incorporate several techniques to assure that workers stay active when a task gets assigned to them. Individual tasks, such as assess tasks, impose anti-cheating mechanisms to discourage spammers from quickly progressing. These mechanisms include character and word count minimums and disabling of copy-paste for free-form entries. Workers are also encouraged to peer-regulate during discussion — participants can indicate a partner's inactivity upon ending a discussion with no agreement. Paired with corresponding payout incentives, these methods ensure that most workers stay active throughout the duration of an experiment.

5.3 Experiments

We deployed our experiments on our synchronous implementation, CICERO-SYNC, to address the following questions: 1) Does multi-turn discussion improve individual accuracy more compared to existing one-shot reconsider based workflows?; 2) Is multi-turn discussion effective in cases where acquiring justifications to implement one-shot argumentation (reconsider) is impractical?; and 3) Do discussions exhibit multi-turn and contextual properties?

We selected two domains to evaluate the research questions above: a traditional NLP binary answer task, Relation Extraction, for comparing against one-shot argumentation and a multi choice answer task, inspired by the word relation game Codenames, to evaluate CICERO in a non-binary choice domain.

In the following sections, we first introduce the experiment setup and configuration, then we introduce each domain and present our results for experiments on that domain. At

the end, we present a qualitative analysis of discussion characteristics and explore whether discussions can improve future accuracy.

5.3.1 Experiment Setup

CICERO's design enables interleaved assignment of different task types (assessments or discussions) for individual workers. This can be beneficial in reducing worker waiting overhead by assigning individual tasks when paired tasks are not available. However, in order to evaluate the effects of contextual, multi-turn argumentation under a controlled setting, we need to isolate the process of assessment and argumentation. For our experiments, we implemented a "blocking" task assigner that avoids interleaved concurrent tasks and is designed to assign the same type of task to a worker until they have answered all questions of that type.

The *blocking assigner* includes a few extra constraints in addition to those required by the workflow:

- **Gold Standard Assessments**: The task assigner assigns **assess** tasks for gold standard questions to evaluate quality of workers who passed the training and gating quiz phase. Workers are assigned these questions before any other questions. No discussions are ever initiated for these questions; they let us control for worker quality and filter workers that do not pass the gating threshold.
- **Greedy Matching**: The task assigner tries to assign a discussion as soon as such a task is available. In the case of multiple candidates, the task assigner picks one randomly.

Additionally, the *blocking assigner* doesn't allocate any discussions until a worker has finished *Assess*-ing all questions. This allows us to collect the initial answers of a worker before they participate in any argumentation.

We adjusted CICERO-SYNC to include these experimental constraints. The resulting system used in experiments consists of three distinct stages: *Training*, *Assess*, and *Discussion / Reconsider* with workers progressing through each stage sequentially.

We conducted a between subjects experiment with 2 conditions. In the **discussion** condition, workers are matched to partners in synchronous discussion sessions after they complete the *Assess* stage according to the allocation policy described earlier. In the **reconsider** condition, we implemented the adaptive workflow and task interface as described in MicroTalk [77] to represent one-shot argumentation. In this condition, workers are adaptively asked to justify or do reconsider tasks depending on their initial answer: When a worker is the only worker with a particular answer for a question, they will be asked to provide a justification for their answer. Reconsider tasks are only assigned to a worker if there is a previously justified answer opposing their current answer. We evaluated the Relation Extraction domain with this experiment setup.

Additionally, we examined the performance of CICERO on multiple choice questions with many answers through the Codenames domain. It is infeasible to run a **reconsider** condition on this domain (as we detail later), so workers only participate in the **discussion** workflow. We also included an extra *individual assessment* stage to examine whether workers were learning from discussions. For simplicity, we may refer to this as the **codenames** condition.

5.3.2 Recruiting and Incentives

We ran experiments on Amazon Mechanical Turk, using workers who had completed at least 100 tasks with a 95% acceptance rate for both of our experiment domains. We recruited a total of 102 workers across the discussion, reconsider, and codenames conditions ($60, 28, 14$ respectively), with a gating pass ratio of 64%, 43%, 63% for each respective condition. Worker drop-out (post-gating) was 1, 0, 2 for each respective condition.

Within each domain, we calibrated our subtask payments by observing the average worker time for that subtask from a pilot run and allocating an approximately $7 hourly wage. Our training bonus of $1.00 for successfully completing training and the gating quiz is also calibrated using the average time it takes workers to complete the training session.

For the Relation Extraction domain (**discussion** and **reconsider** conditions), workers are paid $0.10 as base payment and $1.00 for passing the *Training* stage. Workers are then

paid a per-question bonus of $0.05 for an assessment and $0.05 for a justification during the *Assess* stage. Depending on the condition, a bonus of $0.50 is paid for participating in a **discussion** task and $0.05 for a **reconsider** task in the last stage. Note that in the **discussion** condition, a justification is always collected for each question during the *Assess* stage so workers always get a $0.10 per-question bonus. These per-question incentives are chosen to match those used in MicroTalk [77].

For the Codenames domain, workers are paid $0.20 as base and $1.00 for passing the *training* stage. Workers are paid a per-question bonus of $0.20 for each correct answer and a per-discussion bonus of $0.50 for participating in a discussion with an extra $0.25 for holding the correct answer at the end of discussion.

While it is possible to design a more complex incentive structure, our main goal for this set of incentives is to discourage cheating behavior and align with that of MicroTalk. We think these incentives are consistent with those used in other, recent crowdsourcing research [178].

5.3.3 Relation Extraction Domain: Binary Answer

In the interest of comparing to previous work, we evaluated our method on a tradition NLP annotation task of *information extraction* (IE) — identifying structured, semantic information (relational tuples, such as would be found in a SQL database) from natural language text [111]. The task is of considerable interest in the NLP community, since most IE approaches use machine learning and many exploit crowdsourced training data [304, 209, 7, 178].

Specifically, we consider the problem of annotating a sentence to indicate whether it encodes the TAC KBP *LivedIn* relation — does a sentence support the conclusion that a person lived in a location? While such a judgment may seem simple, the official LDC annotation guidelines are deceptively complex [260]. For example, one can conclude that a national official lives in her country, but not that a city official lives in her city. Figure 5.3 defines the task, showing the instructions given to our workers.

We created a set of 23 challenging TAC KBP questions drawing from the 20 used in

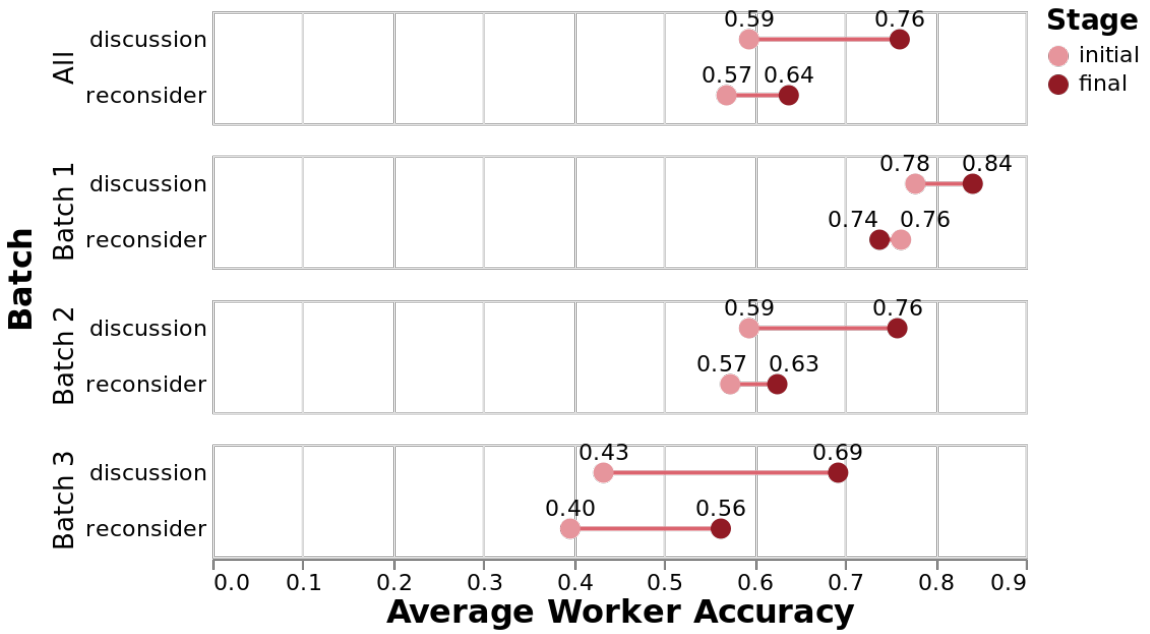


Figure 5.4: Comparison for improvement in average worker accuracy (Relation Extraction domain) for each batch (subset) of questions (Batches 1–3) as well as on the entire set of questions (All).

MicroTalk [77] and adding 3 additional questions from Liu *et al.* [178]. This set was then divided into 3 batches of size 7, 8, and 8 for our discussion experiments. For gold standard questions, we selected 3 simple questions from the TAC KBP set, each of which can be resolved with an invocation of one rule. Upon recruitment, each worker is also presented with a 6 question gating quiz and are allowed 2 attempts to pass the gating threshold. Gating questions were written to be simple and unambiguous, testing whether the worker was diligent and had absorbed the guidelines.

5.3.4 Multi-turn vs. *One-shot Workflows*

Our first experiment compares worker accuracy for the multi-turn, contextual discussion workflow design against that of a one-shot (non-contextual) reconsider workflow on the binary answer Relation Extraction domain (*i.e.*, CICERO *vs.* MICROTALK). We deployed both conditions with the configuration described in the experiment setup with the gating threshold set at 100% — workers needed to answer all gold standard questions correctly to be included. Also, since workers need to complete all assessments before starting discussions which would cause increased waiting time on a large set of questions, we deployed the *discussion* condition experiments in 3 batches ($N = 9, 16, 13$) corresponding to the 3 batches the experiment questions were divided into. In the *reconsider* condition ($N = 12$), workers were put through our implementation of the adaptive workflow from MICROTALK on all questions.

From the plot shown in Figure 5.4 we can see that the *discussion* condition (CICERO) improves average worker accuracy by 16.7 percentage points over the initial accuracy compared to 6.8 for the *reconsider* condition (statistically significant, t-test at $p = 0.0143$).

We performed a t-test on the initial accuracy of workers across both conditions for each batch and found no statistically significant difference ($p = 0.77, 0.78, 0.67$) indicating that workers of similar quality were recruited for each of our batches. On average, workers participated in 7.7 discussions ($\sigma = 4.75$) and were presented with 16.8 reconsider prompts ($\sigma = 3.83$) in the one-shot workflow.

We do note that discussions are more costly, largely due to paying workers for time spent

Candidates	business, card, knot
Positive Clues	suit, tie
Negative Clues	corporation, speed
Explanation	Workers must find the single best candidate word that is related in meaning to some positive clue word, but none of the negative clues. In this example, all three candidates are related to some positive clue: a suit for business, a suit of cards, and to tie a knot. However, business relates to corporation and knot is a unit of speed. Card is the best answer: it's related to a positive clue, while being largely unrelated to any negative clues.
Best Answer	card

Table 5.1: Example of a simple question used for training from the Codenames domain. Real questions have around 7-10 candidate words.

waiting for their partner to respond. Each CICERO-SYNC discussion took an average of 225.3 seconds ($\sigma = 234.8$) of worker time compared to a one-shot reconsider task averaging 13.6 seconds ($\sigma = 15.0$). We believe that an asynchronous implementation of CICERO could reduce overhead and dramatically lower costs.

5.3.5 Codenames Domain: Multiple Choice with Many Answers

Previous work using one-shot argumentation [77, 237] focused mainly on evaluating argumentation in domains that only acquired binary answers such as Relation Extraction or sarcasm detection. These systems ask workers to fully justify their answer, which can be

done by arguing against the opposing answer and for one's own.

However, we observed that this is not sufficient to represent a wide variety of real world tasks. As the number of answer options grows, it becomes increasingly inefficient and even infeasible to ask workers to provide full, well-argued justification for their answers beforehand. Full justifications for multiple choice answers would need to address not only the worker's own answer, but also argue against *all* remaining options, making the justifications long and difficult to understand. Multi-turn discussion can address these scaling issues through back-and-forth dialog through which workers argue only against their partner's specific answer.

Inspired by the popular word association *Codenames* board game, we created a new test domain that requires choosing between numerous possible answers. Similar game-based domains have been adopted to evaluate cooperative work designs such as in DreamTeam [309], which utilized a cooperative version of Codenames, and CrowdIA [169], which used a mystery game. The objective in the game is for each team to identify the tiles assigned to them from a shared list of word tiles. Clue words are given by one team member (the "spymaster") who can see the assignment of word tiles (which ones belong to which team) while other teammates have to find the correct word tiles for their team while avoiding the tiles assigned to the other team.

Our Codenames task domain draws inspiration from the competitive aspect of the game. We observe that late into the game, good word guesses are often informed by both the teammate clues (which should be matched) and opponent clues (which should be avoided). With this observation, we created tasks which consist of a list of candidate words, several positive and several negative clue words. Workers, in the role of a team member, are instructed to find the single best candidate word that is related in meaning to some positive clue word but none of the negative clues. An example of this task can be seen in Table 5.1. Each question contains around 2 positive clues, 2–3 negative clues and 7–10 candidate words. We created 3 gating questions, 7 experiment questions, and 1 question for the *individual assessment* stage for this task. We used a gating threshold of 66.7%. While Codenames is not a typical task for crowd work, as also noted in DreamTeam, we think its aspect of multiple choice answers is representative of a whole class of similar tasks that lack effective

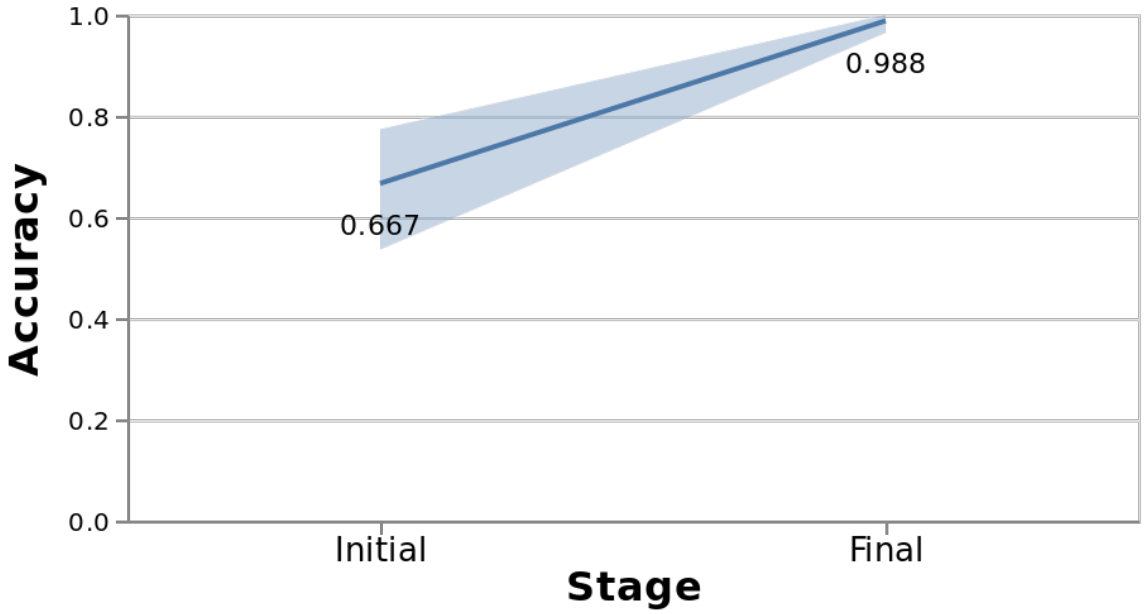


Figure 5.5: Initial and final accuracy of multi-turn argumentation on the Codenames domain with 95% confidence intervals.

one-shot argumentation strategies.

The loose definition of words being "related" in the Codenames domain reduces the amount of worker training required for participation since it utilizes common knowledge of language. However, this may lead to ambiguity in reference answers which would be undesirable. We elected to manually create a set of questions which were validated to have only 1 objectively best answer. The distractors for each question and our reference argument were evaluated with a group of expert pilot testers. We confirmed that all participants agreed with our reference counter-arguments against the distractors and also with our reference answer. In the pilot test, we also noted that this task can be very challenging even for experts as multiple word senses are involved in distractors.

5.3.6 Evaluating on Multiple Choice Tasks with Many Answers

Our second experiment (N = 12) examines the performance of CICERO-SYNC on multiple choice answer tasks from the Codenames domain, a domain that would be very inefficient

for one-shot argumentation (justifications would need to address up to 9 alternatives). We achieved a final average worker accuracy of 98.8% compared to 66.7% initial accuracy (Figure 5.5) – a 32.1 percentage point improvement.

We tested the significance of this improvement through an ANOVA omnibus test with a mixed effects model using worker initial accuracy as a random effect and found that the improvement was statistically significant at ($F(1, 57.586) = 85.608, p = 5.445x10^{-13} < 0.001$). The average duration of each discussion was 123.56 seconds ($\sigma = 64.79$) and each worker had an average of 6.3 discussions ($\sigma = 3.89$).

5.3.7 Discussion Characteristics

We can see from the previous experiments that multi-turn, contextual argumentation is effective at improving worker accuracy across a variety of tasks, but are the discussions actually taking advantage of multi-turn arguments and the context being available? To answer this question, we collected and analyzed the transcripts recorded for each domain: Relation Extraction and Codenames.

We computed statistics on multi-turn engagement by analyzing the number of worker-initiated messages – each of which is considered a turn. We found that in the Relation Extraction domain, discussions averaged 7.5 turns ($\sigma = 6.1$, median of 5) while in the Codenames domain discussions averaged 8.3 turns ($\sigma = 4.23$, median of 7). We also found that in Codenames, the number of turns correlates to convergence on the correct answer ($F(1, 31) = 7.2509, p < 0.05$) while we found no significant relation between turns and convergence ($p > 0.1$) in the Relation Extraction domain. We note that in Relation Extraction, discussions are seeded with workers' justifications from the assess task (equivalent to 2 non-contextual turns, which should be added to the average numbers above for comparison purposes) whereas discussions in the Codenames domain use actual contextual turns to communicate this information. Compared to workers in Relation Extraction conditions, workers in the Codenames discussions sometimes utilized extra turns to reason about alternative choices neither worker picked when entering the discussion.

Additionally, we noticed several patterns in the discussion text that appeared in both

	Relation Extraction	Codenames
Refute	42%	59%
Query	25%	35%
Counter	34%	14%
Previous	16%	10%

Table 5.2: Proportion of each pattern appearing in discussions that converged to the correct answer for each domain. Refute and Query suggest utility of multi-turn interactions while Counter and Previous mainly suggest utility of context.

domains. We further examined these patterns by coding the the discussion transcripts (147 from Relation Extraction and 38 from Codenames). We surveyed the discussions looking only at patterns specific to argumentation and came up with 8 patterns related to argumentation techniques and 6 reasons workers changed their answer.

We then narrowed down the argumentation patterns by removing any that were highly correlated or any that had just 1–2 examples and finalized the following 4 prominent patterns as codes:

- **Refute**: Argue by directly giving a reason for why the partner's specific *answer* is believed to be incorrect. Examples: "Small [partner choice] is the opposite of large [negative clue] and will not work"; "Louisana [sic] isn't a country, therefore NonCountry applies."

- **Query**: Ask the partner to explain their answer, a part of their answer or ask for a clarification in their explanation. Examples: "Why do you think it should be bill?"; "How would bridge work?"

- **Counter**: Pose a counter-argument to a partner *in response to* their explanation. Example: A: "Erdogan's government is nationally affiliated with Turkey." B: "[...] The sentence could be interpreted as one of Turkey's allies is helping them with the EU thing."

- **Previous**: Explicitly state that knowledge/line of reasoning acquired from a previous discussion is being used. Example: "I had window at first too, but someone else had bridge, but they thought bridge because of the card game bridge, and that made sense to me";

We found that workers used these contextual patterns frequently during their discussions for both domains with 77.6% and 86.8% of all discussions utilizing at least one pattern in the Relation Extraction and Codenames domains respectively. We can also see that distribution of patterns across the two domains (Table 5.2) on discussions converging to the correct answer indicates that the utility of each pattern may be different in different domains. We hypothesize that the higher frequency of **Counter** and lower frequency of **Query** in Relation Extraction is likely due to the justification seeding which reduced need for workers to ask for explanations but encouraged more counter-arguments.

We also condensed the reasons for workers changing their answer down to 3 basic categories: learning about the *task* (rules), agreeing on meaning of concepts in a *question*, and being *convinced* by an argument. After coding the discussions, we found that the distribution of the reason for changing answers was $18\%, 3\%, 79\%$ for Relation Extraction domain and $17\%, 28\%, 55\%$ for Codenames, across each category (task, question, convinced) respectively showing that discussions could help workers understand the task.

We also observed that 70% of all discussions and 75% of discussions converging to the right answer used our rule shorthands when referring to the rules instead of describing them. However, we note that simply citing shorthands doesn't correlate with convergence of a discussion ($p > 0.1$).

5.3.8 Do Workers Learn Through Discussion?

While we didn't design discussions to be used as a way of training workers, many reported that they "understood the task much better" after discussions in pilot experiment feedback so we explored the effects of discussions on workers' future accuracy. We tested a worker's performance by adding post-test questions after they finished their corresponding experiment condition. We selected 4 questions for the Relation Extraction domain and 1 for the

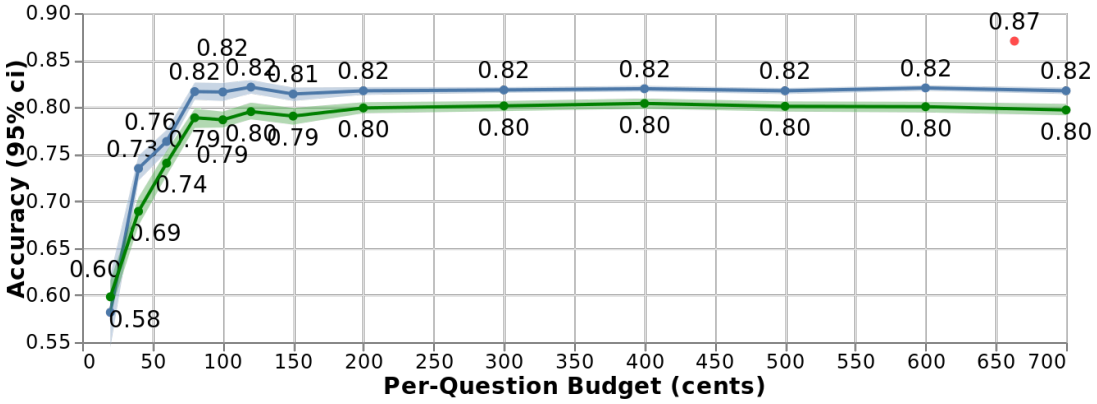


Figure 5.6: Scaling of majority vote (green) and EM-aggregated performance (blue) for one-shot argumentation (Microtalk) on the Relation Extraction domain, computed by simulation (100 simulations per budget) excluding training cost. While expensive due to the use of real-time crowdsourcing, EM-aggregated performance of CICERO-SYNC (shown as a red dot) is higher.

Codenames domain, all of comparable difficulty to the main questions, to be individually evaluated.

Average accuracy on the individual evaluation sections trended higher for the discussion condition: accuracies were 66.7%, 69.3%, and 73.9% for the *baseline* (no argumentation), *reconsider* and *discussion* conditions respectively in the Relation Extraction domain and 46.7% and 52.0% for the *baseline* and *discussion* conditions in the Codenames domain. However, ANOVA on all conditions for each domain shows no statistically significant interaction ($F(1, 49.1) = 0.013, p > 0.1$ and $F(2, 58.3) = 1.03, p > 0.1$ for Codenames and Relation Extraction respectively) between the experiment condition and the accuracy on the individual evaluation questions. We conjecture that need for argumentation may be reduced as workers better learn the guidelines through peer interaction [160, 76], but the difficult questions will likely always warrant some debate.

5.4 Discussion

While each **discussion** task in CICERO-SYNC required more worker time, we found significantly higher gains to individual worker accuracy compared to the **reconsider** condition from MICROTALK. We believe that much of the increase in work time stems from our decision to use synchronous, real-time crowdsourcing in CICERO-SYNC, leading to higher per-argument-task costs. Under a synchronous environment, workers must wait for other workers' actions during and in-between discussions. Since our experiments are focused on *evaluating* the multi-turn argumentation workflow, synchronized discussions allowed us to better collect data in a controlled way. Many efficiency optimizations, that we did not explore, could be implemented to run the CICERO workflow at scale in a more cost effective way. Specifically, an asynchronous implementation of CICERO would eliminate the need for workers to wait for each other, reducing costs. However, if the synchronous implementation were run at larger scale on a much larger set of problems, there would be proportionately less overhead. A semi-asynchronous workflow can be created using notifications and reminder emails [237]. Larger asynchronous group discussions can also be made possible through summarizing discussions [302] thus reducing the cost of new participants getting up to speed.

Argumentation, whether one-shot or multi-turn, may not be appropriate for many tasks, even those requiring high-effort [52]. For example, if one is merely labeling training data for supervised machine learning (a common application), then it may be more cost effective to eschew most forms of quality control (majority vote, EM or argumentation) and instead collect a larger amount of noisy data [173]. However, if one needs data of the highest possible accuracy, then argumentation — specifically contextual, multi-turn argumentation — is the best option. We simulated the effects of recruiting more workers according to the policy described in [77] at higher budgets. Figure 5.6 shows performance for one-shot argumentation after aggregating answers across all workers using EM along with the aggregated CICERO-SYNC results. We observe that accuracy plateaus for one-shot argumentation, confirming previous reports [66, 77], and that CICERO achieves 5 percentage points higher aggregated accuracy compared to previous work, even when accounting for the higher cost

of multi-turn discussions.

In the end, the most cost effective crowd technique depends on both problem difficulty and quality requirements. High-cost methods, like argumentation, should be reserved for the most difficult tasks, such as developing challenging machine learning *test* sets, or tasks comprising a high-stakes decision, where a corresponding explanation is desirable.

5.5 Conclusion & Future Work

In this paper, we explored the potential for multi-turn, contextual argumentation as a next step for improving crowdsourcing accuracy. We presented CICERO, a novel workflow that engages workers in *multi-turn, contextual* argumentation (discussion) to improve crowd accuracy on difficult tasks. We implemented this workflow using a synchronous, real-time design for discussions tasks and created the CICERO-SYNC system. Since the quality of a discussion depends on its participants, we also designed and implemented gated instructions and a novel justification training task for CICERO-SYNC to ensure competent discussions through improving workers' ability to recognize and synthesize good arguments.

We demonstrate that our implementation of CICERO-SYNC, the synchronous version of the CICERO workflow, is able to achieve two things:

- Higher improvement in accuracy compared to a state-of-art, one-shot argumentation system on a difficult NLP annotation task: a 16.7 percentage point improvement over individual workers' pre-argumentation accuracy *v.s.* a 6.8 point improvement using one-shot argumentation and 5 percentage points higher aggregate accuracy when aggregating the opinions of multiple workers using majority vote or EM, accounting for cost.
- Very high accuracy in a non-binary choice answer task that would be impractical with one-shot argumentation: 98.8% accuracy (a 32.1 percentage point improvement over the initial accuracy.)

Both these accuracies are much higher than can be achieved without argumentation. Traditional majority vote and EM without argumentation approaches plateau at 65% on

similar questions [77]. Additionally, we observed several interesting patterns of discourse that are enabled by multi-turn, contextual argumentation and note that many successful discussions utilize these patterns.

There are many future directions for improving the argumentation workflow and system implementation. Currently, the cost of argumentation is still relatively high but cost may be reduced further as discussed earlier.

There are also details in the interactions that could be examined in future work. While we kept workers anonymous between discussions, benefits of assigning pseudonyms as a persistent identity [237] in repeated sessions may be worth considering. Additionally, the idea of utilizing worker produced highlights to refer to the task guidelines and question in [237] could be incorporated in a future iteration to extend our concept of rule shorthands.

We also envision that better models of discussions could allow a future system to only pair arguments where the outcome reduces uncertainty. Furthermore, there is potential in incorporating natural language processing techniques to identify and support positive behavior patterns during argumentation and opportunities for learning from misconceptions surfaced during discussion to improve training and task instructions [35].

Finally, we believe argumentation techniques can be extended to a wider range of tasks and meta-tasks, including issues like micro-task organization studied in Turkomatic [159] and flash teams [220], as well as offer new avenues for human-machine collaboration.

Chapter 6

JUDGMENT SIEVE: BUILDING DYNAMIC WORKFLOWS TO ADDRESS UNCERTAINTY WITH TARGETED INTERVENTIONS

In the previous chapters, we presented tools for capturing uncertainty in various judgment modalities as well as a workflow to address uncertainty in the form of disagreements. However, in practice uncertainty can arise from multiple sources, such as ambiguity due to limited context, or disagreements due to different perspectives or an underspecified task, and sometimes both at the same time. Simply applying one intervention to reduce uncertainty could result in ineffective or counter-productive scenarios where adjudicators are forced to deliberate on cases they agree is ambiguous. Thus, rather than create one-size-fits-all interventions, if we make use of uncertainty-aware judgment tools tools to distinguish the source of uncertainty, we can target an intervention to solve the most prevalent source of uncertainty.

In this chapter, we introduce a new approach to reduce uncertainty in tasks involving group judgment in a targeted manner—by utilizing measurements that separate different sources of uncertainty during an initial round of judgment elicitation, we can then select a targeted intervention adding context or deliberation to most effectively reduce uncertainty on each item being judged. We test our approach on two tasks: rating word pair similarity and toxicity of online comments, showing that targeted interventions reduced the uncertainty score of the targeted source for the most uncertain cases. In the top 10% cases, we saw an ambiguity reduction of 21.4% and 25.7%, and a disagreement reduction of 22.2% and 11.2% for the two tasks respectively. We also found that our simulated dynamic approach reduced the average uncertainty scores for both sources as opposed to uniform approaches where reductions in average uncertainty from one source came with an increase for the other.

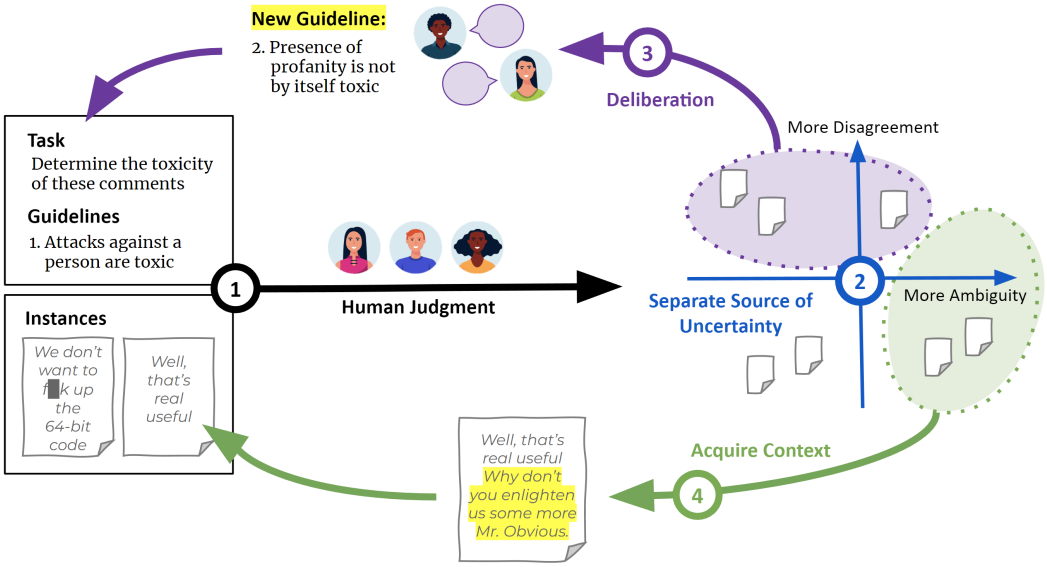


Figure 6.1: A high-level overview of the workflow: (1) Human judgments are collected using an annotation tool that quantifies distinct sources of uncertainty; (2) For each instance, scores that correspond to sources of uncertainty (i.e., *ambiguity* and *disagreement*) are computed; (3) Instances with more disagreement are given the DELIBERATION intervention to resolve disagreements, producing new guidelines; (4) Context is collected for instances with more ambiguity and incorporated into the instance.

6.1 Introduction

Uncertainty is an unavoidable challenge in many tasks that involve making judgments on items. In particular, judgments that involve groups of people must grapple with uncertainty often, as uncertainty in the group setting can arise from both uncertainty experienced by individuals in the group as well as uncertainty at the group level. Individuals in the group may each feel some level of uncertainty due to the *ambiguity* of the item they are judging, making it hard to personally decide on a judgment. At the same time, even if individuals are certain in their personal judgments, group uncertainty can still arise due to *disagreement* between members of the group, which can come from differences in perspectives of the group members that have not been addressed via a specification in the task instructions.

For example, a group of community moderators determining whether a post should be taken down for being harmful may face uncertainty due to ambiguity in the language used and the poster's intent [208]. At the same time, differences in background and culture [139] mean that members of the community may often disagree on what even is harmful [15] and what actions should be taken as a consequence [13]. Similarly, in the education setting, teaching assistants and instructors are often faced with uncertainty when grading open ended assignments. While the goal of grading is to evaluate a student's level of understanding, a poorly designed assignment question may lead to an answer that does not clearly demonstrate the student's understanding one way or the other. Separately, if a shared grading rubric doesn't specify what to do in a particular case, different graders may end up relying on personal judgment, creating disagreement and inconsistencies [246].

Failure to account for uncertainty during the process of human annotation can lead to unreliable and inconsistent measurements [288] even in domains involving expert judgments [58]. Additionally, biases resulting from different backgrounds and perspectives of individuals in the group can also create biased group judgments if not properly accounted for [232, 234]. Due to these observations, in many areas, approaches and processes have been developed to measure uncertainty in data in order to discard unreliable judgments [24, 106, 211] or to create systems that can make use of disaggregated data [89, 167, 282]. While accounting for existing uncertainty is important for building more robust processes, when it comes to actually making decisions, in many cases intervention to reduce uncertainty becomes necessary.

One intervention to reduce uncertainty tackles the ambiguity in the item being judged by providing additional context to help make a decision. For example, providing information such as the parent post of a comment has been shown to sometimes result in an opposite judgment of the toxicity [208]. However, context is complex, and collecting the full scope of relevant context for all cases can be difficult [251]. Thus, it is often infeasible to collect context for all instances ahead of time. Another way to reduce uncertainty tackles the disagreement between individual judgments in the group. Several methods have been proposed to address disagreement by adding greater specification to the task, such as through the use of anchoring examples [49] to ground task understanding and using measured uncertainty

to find unclear guidelines [181]. Other methods tackle the underlying issue of differing perspectives, where deliberation has been shown to improve consensus [237, 50, 86]. However, these interventions also come at a high cost, often requiring a synchronous collaboration process.

Not only are these interventions costly but applying an intervention meant to address one form of uncertainty when the cause of uncertainty lies elsewhere may lead to wasted effort. For instance, prior work has found that resolution of disagreements via deliberation can fail when the context is ambiguous or missing [237]. On the other hand, more context may not be helpful if group members are certain about their judgment but still need to resolve disagreements. Instead of applying all of these costly interventions in every case that presents uncertainty, if we can measure and distinguish the sources contributing to group uncertainty for each case, then it would be possible to select a more targeted and effective intervention on a per-case level.

In this paper, we present a new workflow, Judgment Sieve, for efficiently reducing uncertainty in group judgments. Judgment Sieve involves a decision process that selects a targeted intervention based on the types of uncertainty observed during the initial annotation of each item (Figure 6.1). When individual uncertainty is detected, we focus on acquiring more context to reduce ambiguity in the item; when disagreement between annotators is detected, we focus on engaging annotators in deliberation to reconcile their diverging perspectives and better specify the task instructions.

We make the following contributions in this paper:

- We present Judgment Sieve, a workflow for reducing uncertainty in group judgment scenarios by utilizing measurements related to ambiguity and disagreement for each instance. We also provide a prototype implementation of this workflow for scalar rating tasks.

- We conduct annotations on two scalar rating task domains: word pair similarity rating (*wordsim*) and toxicity rating (*toxicity*), and verify that adding context and deliberation are **effective interventions for reducing ambiguity and disagreement respectively**.

  - In the top 10% most ambiguous cases, we observed a 21.4% (*wordsim*) and 25.7% (*toxicity*) reduction in ambiguity by introducing context.
  - Similarly for the top 10% highest disagreement cases, we saw a 22.2% (*wordsim*) and 11.2% (*toxicity*) reduction in disagreement by introducing guidelines created from deliberation.

- However, we also observed that a broad application of interventions over all items can increase uncertainty in some circumstance, where adding context increased disagreement by 2.06% (*wordsim*) and 3.54% (*toxicity*).

- We conduct a simulation experiment to evaluate the targeted intervention aspect of Judgment Sieve which selects an intervention based on the type of uncertainty measured in the initial annotation. We find that targeted selection of interventions applied to the most uncertain examples resulted in reductions in the overall means of both uncertainty sources as compared to a uniform approach where reductions in one source of uncertainty came with an increase in the other. Though, we do note that when including instances where our dynamic approach did not assign any intervention, this reduction was not statistically significant.

6.2 Related Work

There is an increasing recognition in the spaces of machine learning and social computing that accounting for and addressing uncertainty in crowd judgments is an important problem to tackle. In this section we will review this body of prior work, focusing on: (1) establishing the distinction between error and uncertainty; (2) understanding how (aggregate) measures of uncertainty have been utilized in existing systems and workflows; (3) exploring some theoretical frameworks around distinguishing sources of uncertainty; and (4) discussing prior work around context and deliberation and how they informed the design of the interventions we will be using to address the sources of uncertainty.

6.2.1 Error v.s. Uncertainty

In the past, much work around reducing observed uncertainty in group judgments has been focused on mitigating *errors*, especially in contexts where such judgments are elicited from the crowd [248, 135]. In crowdsourcing settings, the common assumption is that crowd workers conducting judgments are often non-experts with relatively limited training and experience. As a result, the problem of uneven quality and reliability of participants can still be an important concern today [118]. Many approaches to reducing *error* often focus on adjusting the task design itself to provide clear executable instructions to crowd workers and reducing the opportunities for making mistakes [293, 203]. Additionally, procedures and workflows have been developed to improve the efficiency of training [93], selecting [178], and maintaining a high-quality set of attentive crowd annotators [21].

In addition to work on the design side, automated approaches and models have also been developed to correct for errors utilizing the responses of others in the group [65] or via measurements of worker quality through the use of gold standard questions [292]. However, one of the limitations of these approaches is that they often come with the assumption that ground truth can be established with certainty given *only the information in the task presently*, an assumption that is increasingly mismatched for the types of annotation work being conducted today [99]. Prior work has found that in many human judgment tasks, the information provided by the task guidelines and individual instances is not sufficient even for experts to make judgments with certainty [207, 82, 11]. Sometimes this is the result of the task criteria being too vague or potentially self-contradictory [288, 207, 167]. In other cases, individual cases might just be unclear [23] or have plausible alternative contexts under the assumption of which an alternative judgment may be reached [232].

Finally, we also note that not all errors are unintentional aspects of the task. In crowd work settings, we can also experience errors that are the result of workers who may not make an honest attempt at the task [205, 278]—sometimes referred to as spamming workers. However, spamming activity is often motivated by adversarial financial incentives and can be coordinated [72], which means it may be more effective to instead adjust task and incentive designs to discourage spam in the first place. Additionally, as with all human judgments,

crowd work is subject to biases as a result of the recruitment pool [71] and, more generally, human cognitive biases [119].

While addressing *errors* is still an important and integral part of any crowd-based workflow, the increasingly important challenge falls upon how to account for these sources of *uncertainty* not caused by annotator errors but rather arise from the annotators and task. In our work, we mainly focus on creating a workflow that engages with these sources of uncertainty rather than looking into mechanisms to control or reduce errors. This means in practical applications of the Judgment Sieve workflow, we expect each component to also utilize existing designs proposed by prior work to address errors (such as gated instructions [178] and mechanisms to check for attention).

6.2.2 Accounting for Uncertainty in Human Judgments

As mentioned in the previous section, uncertainty is present in many situations where human judgments are involved. As a result, existing literature has also presented a variety of different ways to engage with and account for *uncertainty* in these human judgments.

The most straightforward way that is used to account for uncertainty today, is by using it as a filter—if the annotators don't agree, then the judgment could not be made reliably. Following this approach, common solutions make use of measurement to evaluate the level of disagreement, such as inter-rater reliability [115], after judgments are made and discard data that falls above a disagreement threshold or acquire more judgments until sufficient agreement is reached [226, 305]. Often times, achieving a certain level of annotator agreement is used as a certificate of the quality of a dataset [248]. However, as has been observed, annotated instances naturally lie on a spectrum of uncertainty [262, 288] so dropping examples may lead to a biased sampling of instances. Many are calling for more visibility into how datasets are constructed [97] beyond that of just agreement metrics.

Beyond selecting for instances, uncertainty has also been proposed as a way of reflecting on unclear guidelines or under-specified tasks. For example, work in crowdsourcing has utilized uncertainty as a way to identify guidelines that are unclear [182, 181, 35]. Alternatively, uncertainty has also been harnessed to create guidelines and taxonomies when edge

cases are present [46].

One final approach to working with uncertainty focuses on incorporating uncertainty as a part of downstream models that utilize human judgments. For example, recent systems have been introduced that learn from labels with uncertainty [297, 296]. In the space of machine learning, there is an increasing body of work that seeks to harness dis-aggregated labels as additional training data for models [133, 89] and uncertainty aware models have been shown to achieve higher performance. Not simply limited to training data, uncertainty in labels has also seen use in creating more robust evaluation [106]. However, even when built to utilize uncertainty information, automated systems are not sufficiently flexible and still fail to address uncertainty in the same socially cognizant way humans can [28].

More recently, there is an increased recognition that not all sources of uncertainty should be treated the same as, even within a single dataset, different instances or cases can have different types of uncertainty. Instead some have proposed that we should categorize and quantify uncertainty in order to optimally address it.

6.2.3 Quantifying Sources of Uncertainty

Traditionally, quantification of uncertainty was often done through a statistical lens, by presenting and inferring a probability distribution [60] from the judgments collected. Along this view, recent works have made efforts to quantify uncertainty through capturing distributions over responses. However, capturing answer distributions can be costly [55] and distributions themselves offer little insight into the sources that may have contributed to what is eventually observed statistically—requiring further analysis and data collection to discern [288].

More recently, there has also been work that looks at teasing apart these differences by drawing from classical formulations of uncertainty [121, 151]. One such framework proposes that uncertainty around human judgments can be viewed as arising from two main sources: *aleatoric* (or aleatory) uncertainty—where uncertainty arises from the natural unpredictable variance in the property/phenomena measured, and *epistemic* uncertainty—where uncertainty arises from a the limitations of our models, tools, and understanding [130, 280].

However, others have also critiqued the practical utility of this formulation of uncertainty, as our evolving understanding of the problem can often mean what was once seen as irreducible uncertainty is actually a result of factors not yet known [91].

In this work, we make use of one approach for distinguishing sources of uncertainty: through the lens of *ambiguity* and *disagreement*. This approach allows us to distinguish the uncertainty introduced by. We note that more generally, there can be many different meaningful ways to quantify and distinguish sources of uncertainty that may suit different end goals for addressing uncertainty [250]. While we primarily focused on *ambiguity* and *disagreement* in the evaluation of our workflow, it should be noted that our workflow could be adapted to a different framework for quantifying uncertainty by utilizing different interventions targeted for such alternative distinctions.

6.2.4 Providing Context to Disambiguate

One source of uncertainty in human judgments is often attributed to the *ambiguous* nature of what is being judged. Some have attributed this kind of ambiguity as the result of a fundamental lack of sufficient *context* surrounding the case to be judged. Additional context is commonly used to reduce uncertainty by adding clarity to the instances being decided on. In toxicity rating tasks, adding context about parent posts can affect the outcome [208, 284] while context is also often necessary for investigating online abuse [189]. Many tasks in natural language processing have also seen context added to improve performance [124]. Indeed, the idea of establishing context has also been crucial for human judgment outside that of annotation or crowdsourcing tasks. Classical areas, such as in the legal space where judgments are involved often have complex specialized procedures involving experts to establish context around a case [175]. In the field of education, more effective student performance assessments also involve considering context in the form of evidence of understanding shown through students' responses rather than just the answers [184].

More generally, though, knowing what context to acquire ahead of time can be difficult. In the domain of content moderation, context may be sought out directly as a part of the moderation process. For example, Wikipedia moderators dealing with problematic behavior

in talk pages, may need to investigate the potential use of sock-puppet accounts. This often involves analyzing additional context like a user's past behaviors on the platform, metadata associated with their posts, and interactions with related content, beyond the text directly involved in the moderation decision [251]. In other cases, communities may choose to investigate cases based on their expertise and spotting discrepancies, drawing historical context such as in the case of Civil War portrait identification [193]. Additional context can also target issues with the instance, such as cases where objects in images may be occluded [168], context may take the form of additional images to be collected.

6.2.5 Resolving Disagreement through Rubrics and Deliberation

Another common source of uncertainty in human judgments can be attributed to underlying *disagreements* between individual adjudicators of a case. Many sources can contribute to these disagreements, ranging from inconsistent interpretation of the task criteria to the diverse backgrounds of adjudicators resulting in different perspectives.

For tasks involving crowd annotation, rubrics [94] have been proposed as an effective tool to convey requirements and resolve confusion about aspects of the task. However, rubrics that can cover all the edge cases can be hard to create even by experts, so prior work has utilized crowd participants to help create rubrics [181, 212, 35] by finding areas of high disagreement and asking for suggested guidelines. Rubrics and guidelines can also be implicit, such as in the form of examples [166] or anchors that allow comparison with prior cases [49].

Beyond challenges in creating rubrics, rubrics and guidelines also have limitations when applied. Even when expert-created guidelines are used, adjustments and refinements may be necessary after judgments are made to address issues with the original guidelines [257]. Additionally, beyond layperson crowds, groups of experts judging instances may also just disagree on what the criteria should be [236]. In certain higher-stakes domains like education [246], medical diagnosis [29], or legal judgments [265], it is often the case where the expertise of the humans results in existing guidelines not being applied exactly, instead often conditionally overridden or even contested and overturned. In socially embedded do-

mains, like content moderation, guidelines can also fall out of alignment as distributional properties of the data or adjudicators shifts, such as when social norms shift on online platforms [261, 101] or in broader society. In these situations, past judgments and the criteria that they used may be contested by future adjudicators.

Finally, unclear tasks are not the sole source of disagreement. Even when task goals and guidelines are clear, annotators with different perspective may still disagree about how to judge an item based on different reasoning perspectives [145]. Prior work to automate and scale up deliberation through crowdsourcing has shown that simple reflection-based approaches can be effective at resolving disagreements [77, 156]. More recent work utilizes synchronous deliberation [50, 237] to provide contextual deliberation where those participating in deliberation can quickly form targeted arguments for the particular points of disagreement. Some have also examined the trade-offs between various forms of deliberation design choices (such as the participants, deliberative process, communication medium, etc.) and found that effective deliberation involves building an environment that best matches the task [64]. Of course, more broadly, the successful use of deliberation to resolve disagreement can also depend on other factors. For example, it can be important to make sure deliberation participants are trained to argue effectively [50] and communicate in a way that is collaborative and inclusive [41], especially given the conflicting nature of deliberation. Additionally, the dynamics of deliberation (an intellective task) as groups also means that it can be important to make sure that those matched into deliberation teams are compatible [291].

In our work, we draw from existing literature on disagreement resolution to build out one of our interventions—deliberation. However, designing the right deliberation can be challenging and depends on the participants and tasks, so our application of a simpler form of deliberation may very likely be less effective than customizing deliberation for the tasks involved.

6.3 Design

In this section we will describe our design of the Judgment Sieve workflow. Our workflow consists of the following procedure (as also illustrated in Fig. 6.1):

1. Collect judgments on each instance from individuals in the group using a process that allows measurement of both individual ambiguity and group disagreement for each instance.

2. Compute two scores for each instance: Ambiguity (M_{a}) and Disagreement (M_{d}), based on the measurements in the previous step.

3. For each instance, based on its ambiguity and disagreement measurements, assign potential interventions:

   - If ambiguity score is above a set threshold, the instance is assigned the CONTEXT intervention. Under this intervention, additional context is gathered for the instance and incorporated into it.
   - If disagreement score is above a set threshold, the instance is assigned the DELIBERATION intervention. Under this intervention, a new group is recruited to re-annotate the instance and then conduct deliberation focusing on their disagreements on the judgment for the instance. At the end of the deliberation for each instance, the group will then collectively produce a suggestion for a new general guideline that they think best resolves the disagreement.

4. Incorporate the new information produced from the interventions. Additional context acquired is included as part of the corresponding instance. Additional guidelines produced are included into the judgment task definitions.

5. Repeat the process as necessary until an acceptable level of uncertainty is reached.

In the remaining parts of this section, we will go into more detail about each aspect of the workflow design.

6.3.1 Measuring Sources of Uncertainty

In order to select the right intervention, we first need an approach to understand what sources may be contributing to the group's current uncertainty on each instance. In our

(a) When annotators find instances *ambiguous*, wider ranges—reflecting more rating levels they find acceptable—are produced.

(b) When annotators *disagree* about the rating, we will see their ranges be placed in different locations on the scale, resulting in less overlap.

Figure 6.2: Illustration showing the how the two sources of uncertainty—ambiguity and disagreement—can manifest in the form of range measurements produced by a range-based rating annotation tool like Goldilocks [49].

workflow, we focus on two distinct sources of uncertainty: the amount of ambiguity inherent to each judgment (M_{a}) and the amount of disagreement between judgments (M_{d}). On a high level, one way to think about the distinction between these two sources of uncertainty is through who contributes to the uncertainty: Ambiguity reflects each individual annotator's certainty about their judgment of the item directly collected through our annotation interface (Fig. 6.3); Disagreement is an emergent property that results from aggregating judgments across the individual annotators.

In our experiments, we look at a common application of our workflow in the context of making rating judgments on a continuous scale. Before we can apply targeted interventions, we first need an annotation approach that allows us to distinguish different source of uncertainty. Most common annotation tools that focus on rating judgments measure un-

certainty through aggregation, utilizing disagreements as a proxy for uncertainty [288, 88]. However, in order to apply effective interventions, we need an approach that is able to distinguish the sources of uncertainty. Some prior work have incorporated means for individual annotators to indicate their confidence through directly providing estimates of their own uncertainty [55], however, humans are generally not good at making these types of assessments [271]. For our specific application of scalar rating, we make use of an annotation method introduced by prior work, Goldilocks [49], which proposes a way to separately collect measurements on the sources of uncertainty evaluated by each annotator individually during their annotation process. Goldilocks achieves this by adapting rating judgments as a range annotation task where instead of single ratings, raters produce a range ($[l_i^{(\mathrm{x})}, u_i^{(\mathrm{x})}]$) that reflects values that they find acceptable to place the item.

Using the range annotations collected through this approach, we can define two metrics that quantify different sources of uncertainty for each instance (x). We first look at ambiguity—the situation where an individual annotator is unsure about the rating of the instance being judged. With the range-based annotation procedure, we can see that this kind ambiguity would be reflected through the size of the range produced, with "wider" ranges corresponding to more ambiguity around the rating (Figure 6.2a). Thus we can define an ambiguity score for each instance to be the average size of all ranges collected from the group of annotators participating in the judgment process.

$$\begin{aligned} \text{Ambiguity}(x, i) &= u_i^{(\mathrm{x})} - l_i^{(\mathrm{x})} \\ M_{\mathrm{a}}(x) &= \frac{1}{|N|} \sum_{i \in N} \text{Ambiguity}(x, i) \end{aligned}$$

As for (dis)-agreement between participants, we can see that when range-based annotation is used, the more annotators agree, the more likely it is that the ranges they produce will overlap. So a natural metric can be formed by looking at the amount—in this case the *ratio*—of an annotators range that overlaps with that of another (Figure 6.2b). However, unlike with range sizes which relate to the fixed scale, simply computing the overlap would result in a metric that is also affected by the size of the ranges (or in our case, the *ambigu-*

ity). We can see that as the absolute size of any of the ranges increases (reflecting higher *ambiguity*), the likelihood of that range to overlap with another also increases, resulting in a higher overlap ratio. To account for this and derive a metric for disagreement, we don't directly use the overlap ratio, but instead compare the difference between the measured overlap ratio and the *expected* overlap ratio given the size of the ranges being compared. We note that for any range $[l, u]$, the expected overlap ratio of it compared to another uniformly randomly placed range $[l', u']$ is equal to the size of the other range $(u' - l')$. Given the observations above, we define (dis)-agreement as:

$$\begin{aligned}
\text{Overlap}(l, u, l', u') &= \max(\min(u, u') - \max(l, l'), 0)/(u - l) \\
\text{Agreement}(x, i) &= \sum_{j \neq i \in N} \text{Overlap}(l_i, u_i, l_j, u_j) - (u_j - l_j) \\
M_{\text{d}}(x) &= -\frac{1}{|N|} \sum_{i \in N} \text{Agreement}(x, i)
\end{aligned}$$

Intuitively, higher agreement scores for an annotator on an instance would indicate more agreement between that annotator and their peers. Positive scores imply that the agreement on this instance was higher than random—that annotators leaned towards *agreement*, while negative scores indicate lower than random agreement—that the annotators leaned towards *disagreement*. We note that such a definition of agreement for each annotator is generally not commutative (i.e., the agreement between a pair of annotators A, B measured from A is not necessarily equivalent to that measured from B). This reflects the natural asymmetry present in agreement as exposed through range overlap—for a hypothetical pair of annotators, the one with a "narrower" (subset) range may agree with their "wider" (superset) partner as both accept the ratings in the "narrow" range, while from the partner's perspective some ratings that they indicated as acceptable were not accepted by their "narrower"-ranged partner. Finally, to make the metric intuitive, we can take the negation of the "agreement" metric to define *dis*agreement. We can arrive at a per-instance disagreement score by taking the average disagreement across all annotators.

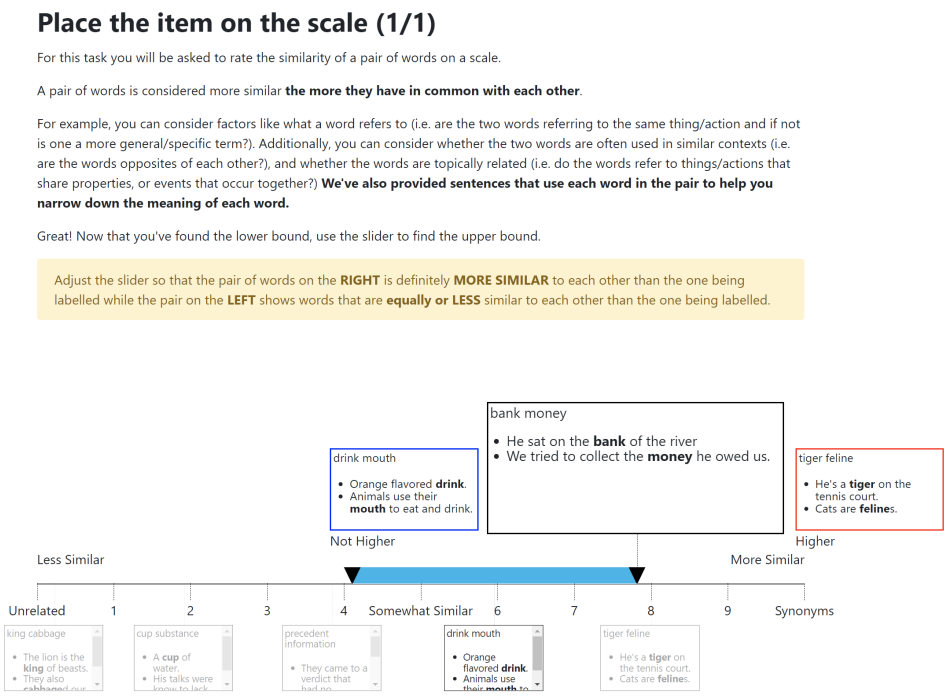


Figure 6.3: A screen capture of the interface used in the annotation process. This annotation tool allows us to collect measurements of individual judgments by annotators of their observed ambiguity of each item and allows us to measure disagreement through comparing the ranges across different annotators.

6.3.2 Gathering Additional Context

In cases where ambiguity is high among individual judgments, *context* has been shown to be an effective way to reduce this uncertainty in both traditional human judgment settings [236] as well as for group judgments facilitated in the form of crowdsourced tasks [189]. However, depending on the judgment task involved, *context* itself can encompass a wide variety of types of information, all of which come with varying amounts of cost involved to capture while not necessarily proving effective for reducing ambiguity. Even when capturing context is cheap, presenting too much context can risk exhausting the limited attention capacity of human adjudicators and bog down the judgment process [224] and misleading context

could result in bad decisions [267]. As a result, if we want to have human judgments that are scalable, it is likely that attempting to *comprehensively* capture context will be an intractable goal. Thus we need to build a process for gathering additional context that can be informed by measurements of what cases are actually ambiguous and may benefit from context. In this section, we will give some examples of how we envision additional context to be gathered for various task settings, and present some cases on how measuring sources of uncertainty can inform the collection of additional context.

In the first example, we look at how context may be added in a group judgment setting involving community or platform moderators making **content moderation** judgments. Content moderation in many online communities often takes the form of a group of *moderators* who need to collectively decide on a moderation action (such as demoting or removing content, placing a ban on the user, or doing nothing) [86, 188]. While many cases may have clear evidence supporting a certain action, historically there have been high-profile cases where limited context contributed to journalistic content being classified as pornographic [101, 134, 194]. In other cases, context, like historic behavior or metadata (e.g., IP addresses, time, device fingerprints), can be used to establish background information that affects the severity of the moderation action, where repeated offenses or attempts to evade enforcement can result in different (often more severe) actions being taken [251]. Our workflow facilitates this flexibility by surfacing judgments that may lack sufficient context to adjudicators. For example, a group of moderators using our workflow might identify a list of cases where observed ambiguity was high. These moderators can then examine each case, noting any types of context that may be helpful to clarify the judgment. Separate investigation processes might then be invoked to gather this information which can then be incorporated as part of a "casebook" surrounding the content [86].

In our second example, we look at what it might mean to gather context in a setting involving **data annotation** by a group of crowd workers. Ambiguity in annotation settings often happen because of the variations in the quality of data. For example, in large image annotation tasks, there can often be images that are unclear [67] or categories that are difficult to decide on with the available information such as species identification for animals [122]. Traditionally, because datasets are fixed ahead of time, context cannot be

gathered during the annotation process. Instead annotators may be asked to flag instances that are unclear, with some processes also allowing a description of why the case was unclear. Requesters may then go through the unclear instances, replacing them with clearer instances if possible or discarding them when they cannot acquire new data. For example, some natural language datasets may be constructed by extracting spans from a corpus. In this case, should some extracted span be ambiguous, the requester can adjust the extraction process to include additional context from the surrounding text. However, we do note that, more often the scale of the data involved in crowdsourced annotation often means that requesters forgo context entirely in favor of dropping data points as individual data points are often not worth the effort to disambiguate. Plus, requesters need to rely on self-reported flagged instances, which crowd workers may not be eager to indicate. Our workflow adapts to this by providing uncertainty measurements baked into the annotations themselves, meaning that requesters can more reliably identify cases they may want to collect context for. This can be useful when constructing challenge datasets where the goal is to create tasks that are difficult but not ambiguous.

As a consequence of the varied types of context, there isn't a single approach to gather context common across all domains. Instead, we envision that the process will vary based on what context is relevant to making better judgments on the task itself. Our workflow discusses the general process of acquiring context abstractly by modeling the process as a whole in the form of a CONTEXT intervention applied to an individual instance of a case. As it is not our goal to develop novel ways to acquire context, in our later experiments, we will use datasets that already contain pre-acquired contextual information and compare the case of withholding context against providing it as a proxy for a separate context acquisition process.

6.3.3 Using Deliberation to Resolve Disagreement

The design of our deliberation process is inspired by prior work on resolving disagreement using synchronous deliberation [237, 50]. In our workflow, the disagreement metric M_{d} is used to automatically *find* candidate instances that may benefit the most from deliberation—

cases where disagreement is the primary source of uncertainty. Then a group deliberates on each example by first independently performing a judgment on the item, and then collectively discussing synchronously. Judgments from each group member is visualized during the deliberation process and the group is collectively prompted to use this to compare their own judgment to those of their peers. Deliberation participants are prompted to consider and elaborate to peers the criteria they used to make their judgment. However, unlike in traditional deliberation systems where the outcome of the deliberation is a judgment on the instance, the goal of our deliberation process is to produce a generalizable guideline for resolving similar disagreements. After engaging in the discussion-based deliberation process, participants are prompted to consider the perspectives they observed during deliberation as well as the deliberation outcome to collaboratively propose a guideline for future examples that *resolve*s the difference in perspective for this instance.

After all the deliberations have concluded, proposed guidelines can be collected and, if needed, de-duplicated. This produces a final set of guidelines that can be incorporated back into the task, so that future disagreements of a similar type are accounted for in the task itself. We note that this overall workflow is also reminiscent of prior methods proposed to utilize worker-provided feedback to improve the quality of instructions in crowdsourcing tasks [181, 212]. However, our workflow makes use of the deliberation process to focus the participants on proposing more effective resolutions that account for the disagreements observed rather than inadequate instructions.

6.3.4 System Prototype Implementation

In prior sections, we've presented our workflow from a more conceptual perspective. Now we will describe the technical details around the prototype[1] that we used to conduct our experiments. To build out the system prototype for our workflow, we created 2 main components: (1) an **annotation** application to collect range-based scalar ratings enabling the measurement of ambiguity and disagreement (Figure 6.3); and (2) a **deliberation** application that collects range-based ratings and then matches participants into synchronous

[1]Code available: `https://github.com/Social-Futures-Lab/targeted-interventions-code`

deliberation sessions (Figure 6.4).

Our **annotation** application follows the general design of Goldilocks [49] and is implemented as a static web application with the input annotation data and output annotator responses stored directly through Amazon Mechanical Turk (AMT). We use a custom JavaScript toolkit[2] to interface with AMT and coordinate the experiment conditions and data storage. Our **deliberation** application is inspired by the design of prior synchronous deliberation systems [237, 50]. We use a front-end to interface with AMT and acquire data about instances to be judged in a way similar to our annotation application, with an additional Python-based back-end server that coordinates the real-time synchronous discussions. As our focus is on evaluating the workflow, unlike prior synchronous deliberation systems which incorporate complex dynamic matching algorithms to address issues like unresponsive partners, our back-end server uses a human-guided matching mechanism (the details of which we describe in Section 6.4.3) facilitated by an internal-facing dashboard.

6.4 Experiments

To evaluate the effects of interventions on group judgment uncertainty, we conducted annotation experiments to collect measurements on the uncertainty of group judgments both before any interventions were conducted and after each intervention was applied. For each instance annotated, we collected ambiguity and disagreement measurements using the our range based annotation application under the following conditions: BASELINE—no intervention applied, CONTEXT—context was included as a part of each instance, and DELIBERATION—additional guidelines from the deliberation intervention were provided as part of the task.

6.4.1 Tasks

For our experiments, we selected two annotation-based tasks that commonly produce uncertainty in group judgments: word similarity (**wordsim**) and toxicity rating (**toxicity**). Both task domains have seen use in prior work and are examples of tasks that contain multiple

[2]Code available: `https://github.com/jmchn1994/amt-shim-template`

sources of uncertainty during judgment.

The **wordsim** domain consists of examples based on an the WordSimilarity-535 Test Collection [88] and is structured as a task to judge the relatedness of pairs of words on a 0-10 scale. This domain was selected because it features varied sources that contribute to uncertainty of both the group and individuals. For one, the "relatedness" of words as a concept is only vaguely defined in the **wordsim** task itself, which can lead to different notions of the relatedness between different people reflected as different schools of thought such as comparing the relatedness of words through various facets such as their meaning, usage, generality and occurrence patterns. Additionally, many of the words involved in this task have multiple word senses. Because no context is provided to disambiguate which word sense is implied, individual annotators must also decide how to reconcile the ambiguity resulting from possible word senses. To seed the range-based annotation process, we used the existing similarity annotations to select 5 seed word pair examples that were evenly spaced along the range with the lowest variance. We then assembled our annotation dataset by selecting a random subset of 50 word pairs divided into 5 groups of 10 from the remaining items.

The **toxicity** domain consists of comments collected from a Wikipedia Talk Pages [208] and is structured as a task to judge the toxicity of each individual comment on a continuous rating scale with 7 point semantic differential scale labels. Judging toxicity itself is a task that comes with considerable uncertainty and disagreement. We note that prior work has shown that the background of each annotator and the circumstances in which comments are posted can greatly affect whether the annotator will see the same post as more toxic or not [234]. This gives rise to natural disagreement and ambiguity in annotations. As some comments can be many paragraphs long greatly increasing annotation effort, we first filtered the dataset to select only instances where neither the comment or parent comment exceeded a length of 280 characters. Then, to seed the range-based annotation process, we used the existing toxicity annotations from the dataset source to selected 5 seed comment examples that were evenly spaced along the range with the lowest variance. We then created our annotation dataset by selecting a random subset of 50 comments divided into 5 groups of 10 from the remaining items.

6.4.2 Acquiring Context

The process of acquiring context in general is usually dependent on the specific goals of the group and the task. As this process is separate from the workflow itself and we did not seek to evaluate the quality of context acquired, we instead simulated the process of acquiring context by using task datasets that were already augmented with context. During annotations in the BASELINE, context of each item was withheld from the annotators, while it was made available during the CONTEXT condition.

For the **wordsim** task, we took inspiration from prior work [124], which used example sentences that contained the word as a way to provide context. For each word in our dataset, we constructed its context by drawing an example sentence that made use of the word in the same form as it appears in the **wordsim** pair. These example sentences were drawn from WordNet [191] when available and when examples were not available, an online dictionary service[3] was used. When multiple word senses existed, a random one was selected to draw the example sentence from. Shorter example sentences were prioritized with long sentences manually simplified. Context for each word pair was then constructed by appending the example sentence for each word involved in the pair.

In the case of the **toxicity** task, our dataset source [208] already contains context information provided in the form of the parent comment of each comment. Context was provided to the annotators by appending the parent comment associated with the item along with a label indicating that it was the parent post.

6.4.3 Conducting Deliberation

We used a crowd task to conduct deliberation to produce guidelines for the DELIBERATION intervention. At the start of the task, each participant first goes through a training session that teaches them to use the annotation interface. After completing this session, participants are placed in a waiting room where they may be assigned either an assessment session or a deliberation session. In an assessment session, the participant uses the range-based annotation interface to provide their judgment for the instance annotated. In a deliberation

[3] `https://www.merriam-webster.com/`

session, a participant is matched with 1-2 partners and asked to use a real-time synchronous discussion interface (Figure 6.4) to discuss the disagreement observed in their range annotations and to collaboratively produce a guideline for future annotators. Guidelines can be proposed or updated by any participant and participants may only leave the discussion after a guideline has been proposed. The allocation of assessment and deliberation sessions was done semi-automatically: While a participant is in the waiting room, the deliberation system makes available a set of sessions available to that participant. A deliberation facilitator can then pick among these options to assign to the participant.

Once the deliberation was complete, the final guideline proposals were collected for each item. We then manually de-duplicated proposals by removing those that were similar. Minor modifications were also made to proposals so that they were phrased in a uniform way for each task domain. The proposals collected were then incorporated into the task instructions for the DELIBERATION condition annotation experiments, with 5 new guidelines added to the **toxicity** task and 6 added to the **wordsim** task.

6.4.4 Recruitment

We recruited crowd workers from Amazon Mechanical Turk (AMT) to conduct the annotations using an annotation interface based on Goldilocks [49] for each of the conditions: BASELINE, CONTEXT and DELIBERATION. For each condition in each domain, we recruited 25 workers (150 in total). Each participant was given 10 items to annotate for each task deployed. Within each domain, we made sure that a worker could not participate in more than 1 annotation task (displaying a notice and preventing further progression if any tasks beyond the first were attempted), ensuring unique worker pools between conditions in the same task domain. A base payment of $1.0 was given to participants for completing a training task with another $1.0 at the end if they completed all annotations. For each annotation completed, participants were paid $0.3 in the **wordsim** domain ($3.0 total) and $0.5 in the **toxicity** domain ($5.0 total). The median hourly pay was measured to be $13.5 and $15.9 for the two domains respectively.

Additionally, we also recruited separate AMT workers to participate in deliberation ses-

sions on instances in each domain in order to create the guidelines used in the DELIBERATION condition. For each domain and task group, we recruited 4 discussion participants (a total of 40). We used qualifications to ensure that the workers participating in the deliberation sessions did not participate in the annotations. Workers were paid $20 for participating in an hour-long discussion task involving 10 discussion and 10 annotation sessions. A bonus of $4 was given for workers who actively participated in discussions beyond the required 10.

6.4.5 Simulation Experiment

With the annotation experiment data for each of the 3 conditions collected, we are able to simulate the outcome of selecting a targeted intervention for each instance. For our simulation experiment, we used the ambiguity M_{a} and disagreement M_{d} scores collected during the BASELINE condition to decide the intervention to use for that instance.

For our experiments, we selected a threshold value of 0.1, which targets the instances that ranked in the top 10% in terms of either ambiguity score or disagreement score. To conduct the simulation, instances were sorted by their M_{a} and M_{d} scores collected from the BASELINE condition. We use this to determine a cutoff threshold for the ambiguity and disagreement scores $(\bar{M}_{\mathrm{a}}, \bar{M}_{\mathrm{d}})$. Then, for each instance in the dataset, we first check its ambiguity score. If $M_{\mathrm{a}}(x) \geq \bar{M}_{\mathrm{a}}$, we assign the context intervention by drawing annotation values from the CONTEXT condition for this instance and moving on to the next instance. Otherwise, we check the disagreement score, and if $M_{\mathrm{d}}(x) \geq \bar{M}_{\mathrm{d}}$, we will draw annotation values from the DELIBERATION condition for this instance. If neither uncertainty metric was above the threshold, we leave annotation values from the BASELINE condition unchanged.

6.4.6 Results

To evaluate our workflow, we focused on 2 main aspects: evaluating the effect of each intervention on the type of uncertainty it targets, and evaluating whether dynamically selecting a targeted intervention based on uncertainty measurements for each example can more efficiently reduce uncertainty compared to a uniform application of intervention.

Specifically, we evaluate the following hypotheses:

- **H1-a (Interventions are Effective)**: An intervention is effective at reducing the source of uncertainty it targets: CONTEXT will be most effective at reducing ambiguity, while DELIBERATION will be most effective at reducing disagreement.

- **H1-b (Interventions are Targeted)**: An intervention is not effective at reducing the type of uncertainty it does not target.

- **H2 (Efficient Uncertainty Reduction)**: A decision process based only on uncertainty measurements collected without any intervention can select a more optimal intervention for each instance that reduces uncertainty more efficiently than a uniform application of an intervention over all instances.

Effectiveness of Targeted Interventions

In this section, we will examine whether our hypotheses for the effectiveness and targeted nature of interventions is supported in our two task domains. To test our hypotheses, we extract 2 subsets of instances (slices) from each task based on the primary source of uncertainty measured during the BASELINE annotation. For each domain, we selected the top 10% instances that had the highest measured ambiguity as a "Most Ambiguous" slice and the top 10% instances that had the highest measured disagreement as a "Most Disagreement" slice. Then for each set of instances, we tracked their uncertainty after re-annotation following each intervention (CONTEXT and DELIBERATION). We visualize these measurements in Figure 6.6.

Looking at the slice of "Most Ambiguous" instances in each domain, we found that only the CONTEXT intervention condition was observed to be statistically significant in reducing the ambiguity across both the **wordsim** and **toxicity** task domains ($p < 0.001$, observed only between the BASELINE and CONTEXT conditions using Tukey's HSD). We found similar results for the slice of "Most Disagreement" instances when it came to disagreement, observing only statistically significant reduction in disagreement between DELIBERATION and BASELINE pairings ($p < 0.001$). This supports **H1-a** indicating that interventions are effective in reducing the type of uncertainty it targets.

We also examined how interventions affected the other (non-targeted) source of uncertainty. In both the **wordsim** and **toxicity** domains we did not observe statistically significant interactions of between the non-targeted condition and BASELINE. While lack of observing significance does not indicate that the non-targeted conditions had no effect on the source of uncertainty, it does indicate that they are not as effective as the targeted intervention, thus this provides some partial support for **H1-b**. Curiously, we did find that on the "Most Disagreement" slice in the **wordsim** domain, while DELIBERATION was significant in reducing that disagreement, it also had a significant effect on *increasing* ambiguity. Due to the nature of the task, we hypothesize that the guidelines produced from DELIBERATION resulted in participants considering more factors (word senses, indirect relationships) when determining the relatedness of words and as a consequence of the lack of any other context, they found the instances to be more ambiguous.

Efficiency of Decision Process

Popping up a level and looking at the case of uniformly applying each intervention across the all instances in the entire dataset (Figure 6.7), we found that CONTEXT was able to reduce ambiguity in both domains ($p = 0.0027 < 0.01$ and $p < 0.001$ for the **wordsim** and **toxicity** domains respectively). However, this seems to also come at a slight cost, also raising the mean disagreement in both cases ($p = 0.026 > 0.01$, not signif.[4], for **toxicity**, $p > 0.01$, not signif., for **wordsim**). This indicates that applying the same intervention across-the-board to all instances can come with tradeoffs, potentially causing increases in sources of uncertainty it was not meant to address. When looking at the DELIBERATION condition, we found no statistically significant effects on either uncertainty source when applied across the entire dataset, with slight increases in the mean value on both measurements. This suggests that while deliberation can be useful for instances with the most disagreement, applying it broadly may be harmful. This result is broadly in line with prior work on deliberation that suggests deliberation is likely only effective when items are already low in ambiguity [237] and should be used primarily on the challenging high disagreement cases.

[4]We set an a priori significance level at $p < 0.01$ throughout our statistical tests.

Next, we compare our results from the simulated decision process where instances are assigned different interventions based on whether their uncertainty is primarily caused by ambiguity or disagreement. When comparing against the BASELINE, we found that our simulated process (SIMULATION-0.1) resulted in lower mean values from both ambiguity and disagreement measures in both domains. However, this decrease was not measured to be statistically significant. The lack of significant results is not unexpected, though, as our simulated selection approach only applies an intervention to the top 10% of instances with highest ambiguity and disagreement as measured during the BASELINE annotations (only affecting at most 20% of instances) while all the remaining instances retained their original annotations. We also note that increasing the decision threshold biases results toward the CONTEXT condition—more significant decreases in ambiguity at the cost of higher disagreement. Interestingly, we observed that our two task domains responded differently to our simulated decision process, with **wordsim** achieving the most reduction of uncertainty through reducing disagreement (-8.7%), while **toxicity** achieved more reduction of ambiguity (-6.4%). We hypothesize that this may be due to disagreements being more challenging to resolve in **toxicity** judgments. In the end, while we don't show **H2** to be true in a statistically significant way with one round of targeted intervention, we do see a differences that may allow us to avoid trade-offs of balancing uniformly adding context or deliberating on all instances.

6.5 Discussion

In this section we will first examine the effect of varying the thresholds for selecting interventions and discuss how thresholds (which affect uncertainty reduction on a per-round basis) work in conjunction with iterative improvement style application of our workflow. Then we will discuss some qualitative observations on the guidelines produced through deliberation and how it may relate to the differences we observe across our two task domains. Following that, we will discuss how our workflow coordinates situations that involve both ambiguity and disagreement and discuss how our workflow can generalize across different tasks and modalities beyond the crowdsourced scalar rating annotation we used in our experiment. Finally, we will discuss some of the limitations of the two interventions we

explored—context and deliberation—as well as avenues for future work that may resolve some of these limitations.

6.5.1 Intervention Selection Thresholds and Iterative Improvement

In section 6.4.6, we found that we are able to observe reductions in both types of uncertainty by simulating a decision process that applied interventions to the top 10% of instances with highest ambiguity and disagreement, respectively, though not at a statistically significant level. As at most 20% of the instances would affected, one question that arises is what happens if we change this threshold to allow interventions to be applied to more (or fewer) instances. To explore this question, we adjusted the simulation parameters to simulate the decision process under additional thresholds as shown in Figure 6.8).

Through these simulations, we can observe that the two task domains tested respond differently in terms of their sensitivity to the targeted intervention selected. For the *word-sim* domain, we find that applying targeted interventions reduces *overall* disagreement but achieves relatively little benefit to *overall* ambiguity. From our results in Section 6.4.6, we know that the CONTEXT intervention is effective at reducing ambiguity for those most ambiguous instances, which indicates that the DELIBERATION intervention likely caused increases in ambiguity on the high-disagreement cases that canceled out the reduction of ambiguity provided by CONTEXT. We hypothesize that in this domain, the additional guidelines led to more comprehensive views on "word similarity" with annotators realizing that cases they would have been certain about (and thus disagreed with each other on) were actually ambiguous (and that they wouldn't have considered those alternative interpretations had it not been for the guidelines). On the other hand, for the *toxicity* domain, we find almost the opposite scenario where targeted interventions resulted decreased *overall* ambiguity but had minimal change to (or even increases to) *overall* disagreement. This suggests that for this domain, more context may have reduced the ambiguity around the setting of the online comments, but may have surfaced new disagreements on what toxicity means for the different annotators [234].

While this simulation result itself is interesting, we note that in practice, one would not

be able to find an "optimal" threshold using this approach as each intervention would need to be applied to all instances, resulting in a very inefficient process. Instead, we posit that improving the threshold to be more optimal would likely not be the most effective way to achieve gains in uncertainty reduction in practice, rather, a better approach lies in the application of our workflow in an **iterative improvement** [107] formulation where our workflow is run in additional iterations that operate on the data and task after application of the interventions from a previous round. Prior work has already shown that some uncertainty interventions, like deliberation, used in our workflow may only be effective on instances that have low ambiguity and may be counterproductive otherwise [237, 50]. Indeed, we even observe this in Figure 6.7, where we found that uniformly applying deliberation across all instances can slightly increase overall disagreement in both domains. However, targeted application of deliberation can reduce disagreement even if indiscriminate application does not (Figure 6.8d). This suggests that, a more effective approach lies in iterating on the workflow rather than optimizing thresholds: after each iteration of the workflow, instances that were ambiguous (and thus not suitable for deliberation) may now be less ambiguous, potentially opening them up to deliberation as an effective intervention in the next round. By focusing on tuning the threshold, we are unable to utilize this benefit as thresholds only affect how interventions are selected within a single iteration. In an iterative construction of the workflow, selection thresholds can instead be seen as a way to control the rate of uncertainty reduction per-round (almost akin to a "learning rate")—lower values are more conservative, affecting overall uncertainty less but more likely to avoid interventions cancelling out each others' benefits, whereas higher values reflect a more optimistic view on interventions, increasing the likelihood of failing to reduce uncertainty in a round, but having a larger impact at each step when it works.

Of course, there are more aspects to consider in the potential design of an iterative workflow. One aspect we briefly touched on in Section 6.2.5 is the idea of a conceptual drift in how groups make judgments. As tasks, group members, and social norms potentially shift over time [303], the judgments that are made and the uncertainty around them can also shift. Here we can envision a potential way where iterative workflows may allow us to adapt to these conditions depending on how we structure such iterations. For shorter term

decisions, groups such as communities, may want to make use of iteration that recruits or utilizes the same adjudicators (such as the group of moderators). This allows us to reap the benefits of uncertainty reduction as our interventions address the uncertainty from the same group of adjudicators. However, across longer time spans, a group may wish to switch to new adjudicators to re-calibrate uncertainty under new conditions. This opens up the potential of creating new guidelines or augmenting with new context that is more applicable new sources of uncertainty.

6.5.2 Utility of Guidelines Produced

In our results, we saw that applying deliberation across the entire dataset can result in increases in disagreement even though we also observe that it reduces disagreement for those cases with the highest disagreement. To explore this, we qualitatively examined several of the guidelines produced through the deliberation process to examine how they may not have been effective at scaling to more instances.

For the *wordsim* domain, we found that deliberation resulted in guidelines that outlined additional criteria for what would be considered as "similar", such as: "Antonyms (light/dark, good/evil) are similar.", "Causal [sic] and effect between words make them more similar.", and "Words part of a natural progression are more similar.". However, while these guidelines would have likely provided more consistent criteria around the word pairs that were deliberated on to produce them, they still leave opportunities for disagreements around applying them—e.g., would a certain word pair be considered a cause-effect pairing or natural progression? For the *toxicity* domain, we found that deliberation resulted in new guidelines such as the following: "Statements about policies not people are not considered toxic.", "Demeaning or condescending statements are likely to be toxic.". Like in *wordsim*, these guidelines are also overall rather narrow ("statements about policies") or could be vague when context was limited ("condescending statements").

While this is not a comprehensive exploration of the effectiveness of producing guidelines, we note that it does provide some insight into why guidelines produced by our particular deliberation formulation may not have generalized well. However, our goal in the evaluation

of Judgment Sieve is less focused on showing the effectiveness of our particular deliberation approach, which utilizes a flat hierarchy and only involves non-expert crowdworkers, and rather attempts to provide deliberation as a proof-of-concept. Indeed, we envision that in practical application of Judgment Sieve, groups are likely to decide to use alternatives to resolving disagreements that are not just a reproduction of our prototype. For example, in content moderation applications, an expert-led deliberation process may resolve some of the issues around the guidelines produced being too specific. Alternatively, for tasks like grading, groups of graders may forgo deliberation and instead defer the resolution of disagreement to others higher up in the hierarchy, like instructors.

6.5.3 Ambiguity and Disagreement All at Once

As we have observed in our experiments, while ambiguity and disagreement are largely distinct types of uncertainty, it is also not uncommon for an instance to have both high ambiguity and high disagreement. What should one choose to focus on when this occurs? In our simulated version of targeted intervention, we opted to prioritize resolving ambiguity before disagreement. This decision was informed by prior work indicating that the deliberation intervention we used (in the form of self-contained synchronous online discussions) may be ineffective when dealing with cases with high ambiguity [237], leading to lack of final resolution. However, in other more general applications of our workflow, it may be more productive to take a hybrid approach that actually starts off with a discussion. For example, in the case of content moderation, some moderation decisions may need to involve both creating new consensus on guidelines to address a novel type of content (as has been seen in platforms' adaptations to misinformation campaigns related to COVID-19), as well as collecting evidence (current scientific consensus, evaluating whether content is connected to larger misinformation campaigns, etc.) that backs a final decision. In these situations, starting off with an open deliberation setting can allow a group to first understand the space of context that will become necessary, directing a more effective context collection process later on in the process. A promising avenue of future work may be to develop approaches to hybridize the collection of context and the deliberation process, allowing groups to switch

back-and-forth between the two as needs arise.

6.5.4 Generalizing our Approach Across Different Tasks and Modalities

In our experiments, we mainly evaluated our workflow through a set of crowdsourced annotation tasks focusing on short text-based tasks under a continuous rating scale. However, more broadly speaking, there are many more scenarios (e.g., expert involved group judgments), tasks domains (e.g., longer form text, instances with multimedia), and judgment modalities (e.g., single or multi-label categorical classification) involving group human judgments where it can be beneficial to reduce uncertainty in a targeted way. We expect that the workflow proposed in this paper and the idea of separating sources of uncertainty should be able to generalize to these types of tasks and processes involving group judgments by adapting new mechanisms for collecting judgments and incorporating corresponding methods to distinguish the sources of uncertainty. In this section we will focus on discussing how we envision our process may be generalized to other input modalities, as well as different group judgment scenarios beyond crowdsourced annotation.

In this work, we evaluated Judgment Sieve on the specific judgment modality of continuous scalar ratings. However, other modalities for human judgments, such as categorical classification, are also commonly used, even in some of the task domains we explore like data annotation and content moderation. More generally though, the workflow we introduce in Judgment Sieve can theoretically be adapted to other decision modalities as long as there are annotation methods that allow us to separate sources of uncertainty and construct interventions that can effectively target those sources of uncertainty. Taking the example of categorical classification, prior work has tackled the problem of uncertainty in categorical classification through tools like soft labels [57] where annotators provide self-evaluated confidence weights over each class as opposed to a binary decision. Thus, to apply Judgment Sieve, we could make use of annotation tools that produce soft labels over categorical classification as our human judgment mechanism. We can then construct measures of ambiguity based on aspects like the *dispersion* (e.g., variance) of the distribution produced by each human annotator, and create measures of disagreement using distribution *divergence*

metrics (e.g., KL divergence, Wasserstein distance) between annotators. This would allow us to then apply domain specific interventions that target these sources of uncertanty.

Another dimension for generalizing our approach lies in its applicability beyond crowdsourcing settings, where non-expert human adjudicators are used to make judgments on lower-stake tasks. For example, take the case of an education setting, where instructors and teaching assistants may want to obtain a clearer picture of how well each students is learning the material by grading their assignments. In this case, the task of grading assignments is a group judgment scenario where graders are trying to reduce uncertainty over the assessment of the student's score. Just like with our annotations, this uncertainty can arise from different sources: sometimes a grading rubric may have criteria that are too coarse, where very different types of mistakes may evaluate to the same score; on the other hand, sometimes a rubric may have inconsistencies, leading to different graders disagreeing about how to score the same problem. Using the Judgment Sieve workflow, a group of teaching staff can systematically diagnose these issues by taking measurements on their judgments. Questions may be assigned to multiple graders to measure whether there is disagreement, while graders can also be prompted self report any answers that they found the rubric to not adequately address. Based on the sources identified, the uncertainty might be addressed by a staff discussion that resolves conflicting rubric items, or by assigning partial credit in cases where the rubric was too coarse. By coordinating measurements and targeted interventions following our general workflow, it becomes easy to make progress to systematically reducing uncertainty surrounding the judgment.

Of course, there are still limitations to our workflow and scenarios where it may fail to provide benefits. If the task skews heavily towards one particular type of uncertainty, the extra effort to measuring and distinguishing the sources of uncertainty may not be worth it if there are limited interventions available. For example, when a task is focused on capturing subjective personalized results such as preference elicitation, disagreement may not be a useful aspect of uncertainty to address. Similarly, when a task is purely perceptual such as evaluating sensory inputs, one may not be able to apply interventions to ambiguity as they arise from a limitation in one's ability to perceive. In these cases, while distinguishing the sources of uncertainty can be interesting, they likely don't provide a more effective way to

reduce uncertainty compared to just applying the intervention that is available as informed by overall uncertainty.

6.5.5 Caveats of Context

While in general context can be effective in reducing ambiguity, in our experiments, we also observed cases where additional context contributed to an increase in uncertainty, especially when the context is unexpected. For example, on the **wordsim** domain, while examining the cases where ambiguity increased after the introduction of context, we observed cases like "bank, money" showing increased ambiguity. Looking into these cases, we found that because our context is based on example sentences for randomly selected word senses, the example of "He sat on the **bank** of the river" was provided as a part of the context for this pair. In this case, the context was likely unexpected for the annotators, who, after realizing the presence of this alternative word sense, accounted for the increased ambiguity in their annotation of a word pair that would have otherwise been clear. Indeed in this case, the change is likely desirable and reflects a real increase in the uncertainty about instances, so focusing on reduction of a single uncertainty aspect doesn't paint the whole picture. We envision that these are situations where a more iterative approach can provide additional benefit—the additional context might then give rise to new disagreements on whether the sentence example is meant to ground the example or just supplement possible word senses, allowing focus to move to the new uncertainty about the under-specified task.

In a broader sense, though, there are limits to how far additional context can go, and we will eventually run into diminishing returns of more context, so an iterative workflow should account for this. Information about an open scientific problem or ongoing investigation can be interesting pieces of context, but also provide little to resolve ambiguity. Depending on the application, it may be desirable to keep track of changes during the re-annotation processes: if context is not improving uncertainty, the process may need to decide to hold uncertainty at the current level, only conducting more annotations if new information arrives or guidelines change.

6.5.6 Limits to Scaling Task Specification with Deliberation

Finally, we also note that there are limits to scaling the current design of our deliberation process. In the current process, deliberation produces additional guidelines which are incorporated into the instructions. While processes like de-duplication and reorganization can be done by task requesters, as the task specification becomes increasingly precise, the instructions grounding the task itself can eventually become too large for those making judgments to keep track of [293]. This issue can be seen in the case of content moderation, where paid contract moderators typically must go through extensive training and review many pages of instructions and examples in order to improve their consistency with other moderators. If the guidelines become too complex, their ability to resolve disagreement can be greatly reduced as people struggle to understand or even find a relevant guideline. A potential solution to dealing with complex task specifications may arise from looking at solutions in the realm of legal case building, another example of a space where 'guidelines' are almost impossibly complex. Taking inspiration from how lawyers build judgments from case law, instead of continuing to add additional specifications to a task's instructions, the instructions are kept at a manageable level of detail, and instead greater resolution around the specification comes primarily from reviewing prior judgments (in the form of precedent cases). More work is needed to consider how to design systems to allow easier exploration of prior judgments when determining how to judge a new item consistently.

6.6 Conclusion

In this paper, we present a new workflow for more efficiently reducing uncertainty in group judgments by applying a targeted intervention on each instance based measurements relating to ambiguity and disagreement. Through our experiments, we find that the interventions of adding context and conducting deliberation do most effectively reduce the type of uncertainty it targets. We also observe that dynamic selection of interventions on a per-item bases has the potential to avoid the trade-offs in uniformly applying interventions to all items.

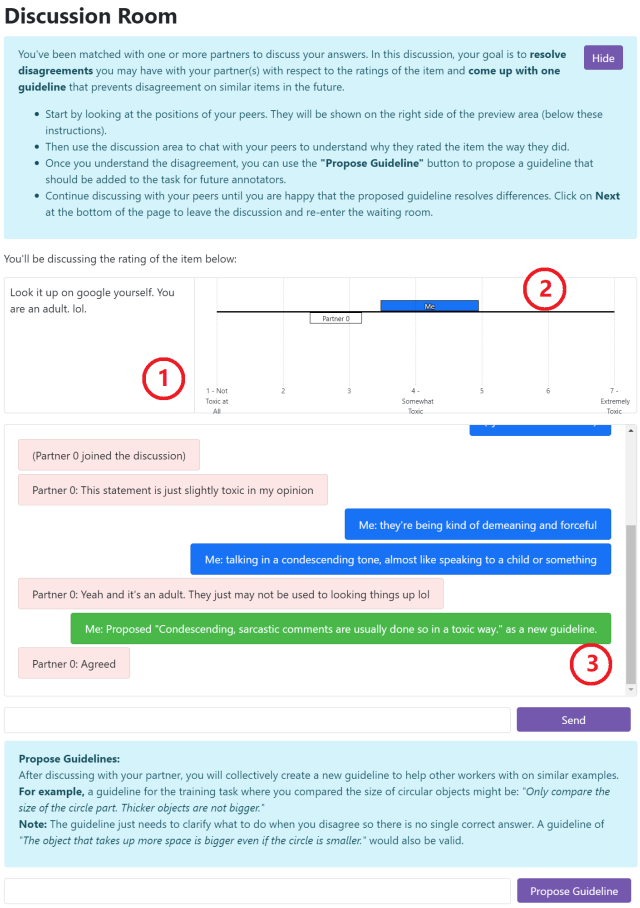
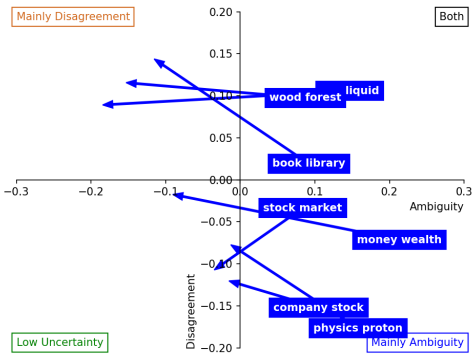
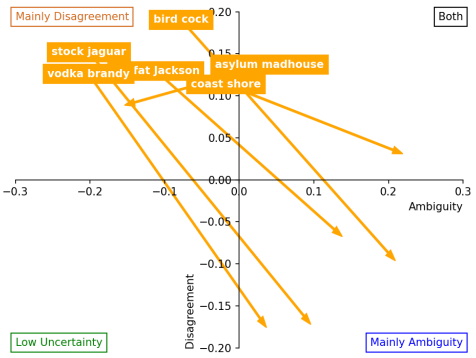


Figure 6.4: A screen capture of the deliberation interface used in our experiments. There are 3 main components to the interface: (1) A preview of the instance that was rated, (2) A visualization of the range answers of each participant shown on the same scale, and (3) The synchronous discussion area.

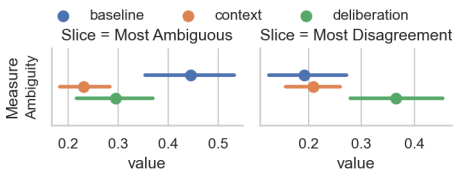


(a) A sample of items primarily exhibiting ambiguity (blue) and their new uncertainty after applying the CONTEXT intervention.

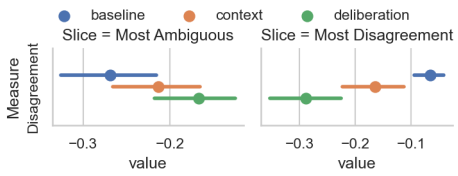


(b) A sample of items primarily exhibiting disagreement (orange) and their new uncertainty after applying the DELIBERATION intervention.

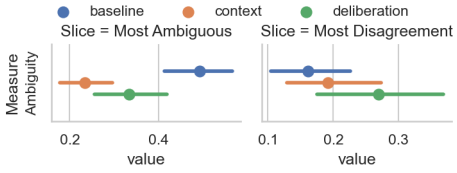
Figure 6.5: An illustrated figure showing how the uncertainty of a small sample of items moved within the uncertainty space. Items indicated in orange exhibited primarily disagreement. Items indicated in blue exhibited primarily ambiguity. Arrows point to the new location in the uncertainty space after applying the targeted intervention. Scores are rescaled such that the origin (0, 0) represents the average ambiguity and average disagreement across all items. Positive values indicate above average uncertainty score measurements.



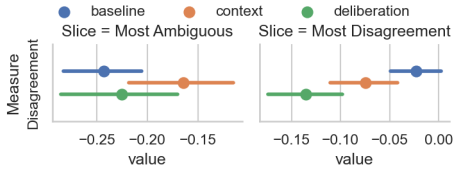
(a) Ambiguity for each slice on the *wordsim* task



(b) Disagreement for each slice on the *wordsim* task

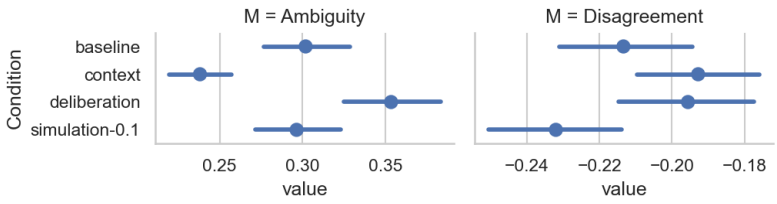


(c) Ambiguity for each slice on the *toxicity* task

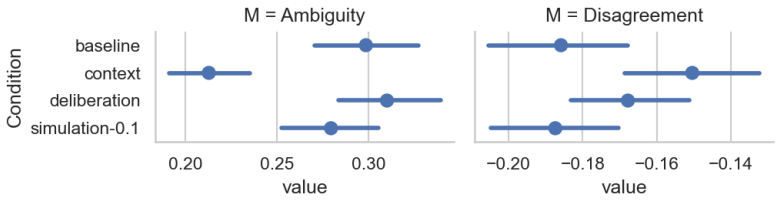


(d) Disagreement for each slice on the *toxicity* task

Figure 6.6: Point plots for each task domain that shows the ambiguity and disagreement measures under the BASELINE, CONTEXT and DELIBERATION intervention conditions. For each measure, we look at two slices of the dataset: The instances in the top 10% by ambiguity M_a ("Most Ambiguous") and those in the top 10% by disagreement M_d ("Most Disagreement"). Error bars indicate 95% confidence intervals.

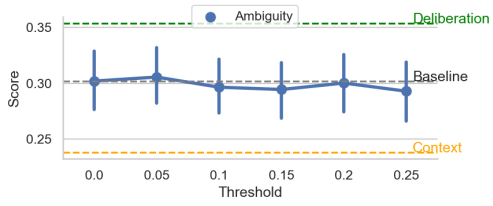


(a) Comparison for the *wordsim* task domain

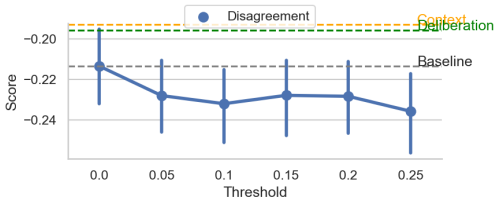


(b) Comparison for the *toxicity* task domain.

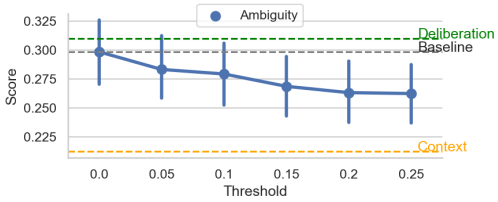
Figure 6.7: Point plots for each domain that show the ambiguity and disagreement measured after applying a uniform intervention (CONTEXT or DELIBERATION) across all instances and from simulating the selection of different interventions targeted to each instance SIMULATION-0.1. Error bars indicate 95% confidence intervals.



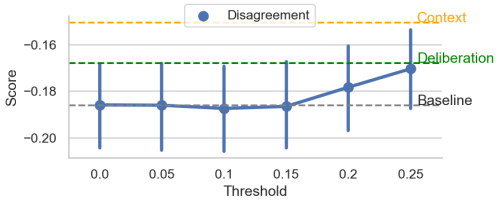
(a) Overall ambiguity on the *wordsim* task



(b) Overall disagreement on the *wordsim* task



(c) Overall ambiguity on the *toxicity* task



(d) Overall disagreement on the *toxicity* task

Figure 6.8: Plots showing the simulated interventions applied at different thresholds of 0% (no interventions applied), 5%, 10%, 15%, 20%, and 25%. For all plots, lower values reflect less uncertainty from the corresponding source. Three reference lines are provided on each graph to indicate the average uncertainty measurements of: BASELINE (grey), CONTEXT (orange), and DELIBERATION (green). Error bars indicate 95% confidence intervals around simulations, confidence intervals for the reference lines are not shown (see Figure 6.7 instead).

Chapter 7

DISCUSSION

In the previous chapters we have presented, evaluated, and discussed several novel tools and processes for understanding and addressing uncertainty. With this chapter, I will discuss some of the implications of this body of work.

First, I will discuss why uncertainty-aware tools are necessary for the current wave of scale-based human judgment tasks. Then, I will discuss how the tools and processes presented in this work can be applied to downstream usage scenarios. I will discuss some of the implications of the precedent-based approach to defining human judgment problems and what it means for the construction of datasets that can evolve along with society in the future. Following that, I will touch on some of the lessons we learned while building uncertainty-aware tools and the design implications they bring for incorporating these ideas in your own tools or tasks. At the end, I will overview some ongoing and envisioned future work that extend on and develop the ideas presented in this thesis.

7.1 The Importance of Uncertainty-Aware Tools for Human Judgment

From chapter 3 to chapter 6, I have demonstrated that by utilizing new tools and workflows that are built to account for uncertainty present in scale-based human judgments, we can achieve higher consistency across annotators on nuanced and subjective scales while also providing insights necessary to select and apply interventions to reduce uncertainty. However, some may note that a common aspect of these uncertainty-aware human judgment tools and workflows is that they require some extra effort from both crowd adjudicators and those conducting the judgment tasks to apply in practice. This extra effort can range from additional time to explore precedent or anchor cases, to full synchronous tasks where one needs to coordinate with another peer adjudicator. With the increasing power and expressivity of machine learning models, it is natural to wonder whether we can forgo this

extra expense by applying approaches that train on dis-aggregated data or utilize uncertainty measurements on data produced by existing human judgment tools and processes, as proposed by recent literature in the machine learning space [63, 89, 57]. Here I will make the case that, for complex and nuanced tasks, the additional cost involved in using uncertainty-aware human judgment tools is well justified, and that, while new models that can better take advantage uncertainty information will benefit the quality of automated judgments, they are complementary to collecting better human judgments in the first place.

There is no substitute for capturing uncertainty directly in initial judgments. As we have discussed in our overview of prior work, there have been various approaches that utilize collected judgments to provide post-hoc measurements of uncertainty [288, 262]. However, these approaches all depend on post-hoc interpretation of what may have been the source of uncertainty. When adjudicators are faced with uncertainty that they cannot express through existing tools, they end up being forced to work around it rather than recognize it. Individual adjudicators may resolve their uncertainty randomly to an agreeable judgment, or they may err on the side of trying to be conservative with their answer to stay in-line with their expectation of the majority [215]. By not providing adjudicators with options to express uncertainty, we lose the opportunity to gauge the adjudicators' natural evaluation of uncertainty and run the risk of reinforcing the confidence of adjudicators.

Uncertainty-aware meta-processes allow for targeted interventions for reducing uncertainty informed by uncertainty measurements. While incorporating uncertainty as a component of downstream systems can contribute to models that more faithfully reflect properties of the existing data, in some cases it can be desirable to diagnose and reduce uncertainty by refining the problem. In chapter 6, we found that while uncertainty reduction interventions can be generally effective, when applied indiscriminately across cases, they can create additional confusion if the targeted type of uncertainty is not present or is not a main source of uncertainty for individual cases. This can even result in an increase in uncertainty under some circumstances. Without methods that can capture and distinguish the sources of uncertainty present on an individual case level at the outset of data collection, downstream applications that attempt to reduce uncertainty may find that they now have to spend extra effort to investigate the uncertainty characteristics of

their task, counteracting any cost savings.

Uncertainty with sources distinguished can provide additional value in the form of transparency. In recent years, many have called out the need to introduce transparency into the datasets created through human judgements [97, 18, 211]. While dis-aggregated data and general uncertainty metrics provide some level of transparency into the data, the lack of insight into the factors contributing to the observed uncertainty means it can be difficult to diagnose problems that crop up down the line. With uncertainty-aware judgment tools, judgments themselves are encoded in a way that preserves information about the sources of uncertainty. This means that it is much easier to examine post-hoc whether issues arise from the selection of annotators (disagreement) or the cases that were judged (ambiguity).

7.2 Building on Understanding Uncertainty

So far we have described the ideas of uncertainty-aware tools, workflows, and meta-processes. However, one may be curious how these tools can be applied in practice. In this section I'll discuss two envisioned scenarios where the ideas from this work can be applied: using uncertainty-aware processes to build automated judgment tools based on machine learning, and using uncertainty-aware processes to coordinate community decisions.

For machine learning practitioners building machine learning models, uncertainty-aware processes mainly support two aspects of their task: assembling training data and of model evaluation. For example, say a machine learning practitioner is tasked with creating training data to automate a system to identify spam-like emails. They can make use of tools like Goldilocks, recruiting crowd participants to annotate a set of emails on a scale of how spam-like it is. The practitioner starts by recruiting experts to conduct annotations over a small set of messages, and then use those messages as the anchors for Goldilocks. The practitioner can then use the uncertainty information in this process to diagnose any issues with the data or problem specification, discovering any instances of messages that may be hard to classify or situations where annotators disagree. Using this information they can apply interventions of either adding more metadata to consider for ambiguous messages, or having annotators deliberate on definitions of spam. Once the level of uncertainty is

appropriate, the practitioner can use the resulting dataset to train a model. During evaluation, a practitioner can have the model make predictions about messages presented in the form of range-based scores. These can then be compared against ranges captured by human evaluators and used to assess how often the system agreed with humans as well as evaluate whether the system was over-confident (with self-assessed ambiguity that was lower than that of humans) when it should not be.

For communities coordinating decisions, uncertainty-aware processes and tools can guide larger procedures as a way to provide transparency and legitimacy of the process. For example, consider a community making a decision on how to moderate a set of posts. Such a community may decide to utilize a system like case law crowdsourcing (chapter 4) to conduct the moderation process. Moderators in the community may recruit a group of "jurors" by sampling a set of community members. These jurors can then be given access to a case law crowdsourcing system, where they are each assigned some cases to adjudicate and can utilize precedents to quickly assemble judgments without having to learn complex sets of moderation rules. Community moderators can then aggregate the judgments made by the juror panel to assess the uncertainty present in the judgments—for cases where jurors overwhelmingly agreed on the relevant precedents to apply, the judgments can be directly adopted, whereas cases can be elevated if there was disagreement or it was observed that precedents did not sufficiently inform the jurors' judgments. Given these cases, moderators can also examine what type of uncertainty was present. Were some precedents mostly agreed on while others had significant disagreement? Did some groups of jurors have much more significant disagreement with other groups? By using the information around uncertainty, the community moderators may get suggestions for whether effort was needed to look into the case or if they should solicit the wider community's opinion on how to settle a disagreement. Once cases arrive at a final judgment with a sufficiently low uncertainty, they can be incorporated into the set of precedents that inform the next round of adjudication.

7.3 Overturning Precedents and Living Datasets

One of the applications we envision with the datasets produced by uncertainty-aware tools is the ability to provide a source of ground truth for social judgments and norms [13, 161]. Because judgments from systems like Goldilocks and case law crowdsourcing encode uncertainty information on a per-annotator level, datasets of these judgments more accurately reflect the characteristics of real individuals rather than aggregated simulations of human adjudicators.

However, along with this also comes with a potential limitation that needs to be considered. Systems like Goldilocks and case law crowdsourcing achieve their consistency by utilizing past judgments to calibrate scales across adjudicators either through anchors or in the form of precedents. When used to construct socially informed datasets, such as moderation decisions, social norms and collective identities can also become embedded into these past judgments [69] which can serve to perpetuate past norms into the future. Of course, this is not a problem unique to the tools we have introduced. Similar issues also pertain to the precedents that ground and define the case law legal systems that inspired case law crowdsourcing, and like with the concept of precedents, the legal system also presents a solution that can inspire how we address the issue of facilitating this concept drift over time. Within case law systems, there is the idea of overturning a precedent decision as a way to indicate that norms and reasoning applied then should no longer apply now. Taking inspiration from this, we can create meta-processes for maintaining datasets into the future where old judgments may be re-visited and, should current adjudicators change their decision, overturned. In fact the construction of case law crowdsourcing even presents a simple way to apply the idea of overturning precedents in practice. Because each judgment is associated with a set of positive and negative precedents, these judgments actually form networks of citations where judgments in the future establish dependencies on judgments from the past. When a past judgment is revisited and overturned, we can relatively easily evaluate the effect this has on other judgments that may have depended on it by traversing the dependency graph. This can allow us to additionally flag candidate sets of judgments that we may also want to revisit due to their dependence on the no longer valid precedent.

The idea of revisiting past judgments can even be applied to the judgments made under Goldilocks. Like with sets of precedents, the bounds in each Goldilocks judgment also form an implicit dependency against the neighboring cases that ground it. Should cases around the bounds of a case be adjusted, then we would get a useful signal that that case may also need to be revisited.

So when should we revisit past precedents? In both systems, as more judgments are made, it becomes more likely that the judgments of the past no longer reflect the norms of the current adjudicators. Thus one way of coordinating and maintaining these datasets for social ground truth can involve tracking expiration dates for past judgments. Once a certain amount of time has elapsed after a judgment has been made, we might decide that it should no longer be used as a precedent or anchor, and instead mark it as a candidate for revisiting. One might also make the broader connection that this kind of gradual fading of past judgments is in some ways similar to how learning rates in reinforcement learning allow us to bias new decisions towards older or more recent observations.

7.4 Design Implications for Uncertainty-Aware Tools and Processes

Finally, we will discuss some design implications for creating new tools and processes to capture, distinguish, and address uncertainty.

Utilize familiar aspects of existing tools and focus on re-framing how they are used: In our work on Goldilocks, we found that while annotators often recognized ambiguous instances during annotation, they can struggle to quantify how much ambiguity is present. In our pilot tests, we experimented with designs where annotators could create ranges similar to confidence intervals centered around a value and then adjust the size or mean value of these ranges. However, this proved to be difficult to use as annotators had to estimate the ambiguity of the instance and correct it after the fact. This prompted us to approach the problem from the annotators' side by looking at what judgments were easier to make for them—in this case, comparative judgments. With this in mind, we ended up with our final design for Goldilocks where a two-step process, each based on familiar interactions similar to traditional slider scales but framed as using comparative judgments to find a specific value for one of the bounds rather than an average placement of the item.

We applied similar ideas when designing the case law crowdsourcing workflow. We expect that for other domains and input modalities, it may also be important to consider utilizing smaller judgments that are easier to conduct by the human adjudicators.

Through our experiments in Goldilocks, we also found that there was a significant variety in the type of task that could affect the utility of different types of anchors. We found that on tasks like age estimation, human annotators can often start off with high agreement on the understanding of the scale, compared to tasks like toxicity and satiety estimation where scales need to be interpreted or learned. This means that while tasks like age estimation also contain instance level ambiguity, there is relatively little uncertainty resulting from the interpretation of the scale itself, leading to example-based anchors not producing benefits for consistency. For designs that work with some aspect of abstractly defined scales or criteria, it may be important to include both example contexts and the original abstract scale.

Training adjudicators on how to think critically can be important: In our work on Cicero, we found that while the workers were given training on the task itself through a gated instructions [178] process, the challenging nature of the task still meant that we observed a considerable amount of disagreement in initial judgments. However, by training the laypeople crowd to identify high quality justifications, we were able to improve the ability for them to resolve disagreement and arrive at the more accurate answers through the use of deliberation.

Uncertainty reducing interventions are often targeted and applying the wrong intervention can have detrimental effects: In the work on targeted interventions, while we did expect uncertainty interventions to target the type of uncertainty they were suitable for, we also surprisingly found that applying them uniformly regardless of the main source of uncertainty can reduce one type of uncertainty while raising the other. Indeed prior work on deliberation [237] has shown that this intervention is not suitable for all situations. However, we found that if it was applied without discerning irresolvable cases whether context was insufficient, uncertainty can increase leading to a negative effect.

7.5 Ongoing and Future Work

In the sections above, I discussed some of the implications of the work presented in this thesis. In this section, I will outline and discuss some ongoing projects that build on the ideas presented in this thesis as well as potential avenues to develop the ideas further, addressing more aspects of the human judgment ecosystem.

7.5.1 Ongoing: Extending Uncertainty Tools to New Modalities and Tasks

In this thesis, I present two tools and workflows, Goldilocks and case law crowdsourcing, aim to improve how human judgments are collected for two common types of tasks—scalar rating and categorical classification. With these tools we simultaneously tackle the issue of improving consistency for human judgments in complex and nuanced tasks, while also providing a way to capture uncertainty during initial judgments. However, beyond these there are many other modalities where human judgments can be involved. For example, a significant amount of annotation work is conducted in the visual domain for producing image datasets that range from medical imaging diagnosis [8] to identifying pedestrian behavior to train self driving vehicles [51]. Many of these domains involve uncertain human judgments and while also being high stakes and sensitive to failures. An important future direction of work would be to extend the set of uncertainty tools into other domains like image annotation.

In addition to this, while annotation tools like Goldilocks can produce ground truth data to construct datasets, the ability to distinguish sources of uncertainty also allows us to gain understanding into the uncertainty surrounding existing annotation tasks. In line with this idea, one additional direction for future work is to utilize uncertainty aware annotation tools to produce measurements that allow us to understand the uncertainty present in existing datasets and diagnose biases that may form as a result of limitations in the data and annotators selected ahead of time [38].

7.5.2 Building Uncertainty Tools to Support Online Communities

While the work in this dissertation focuses mainly on tools for understanding and addressing uncertainty in a crowd annotation setting, many of the tools presented are not limited to being used by layperson crowds. As we have seen in related work, communities are increasingly using data and technology mediated decision processes to conduct tasks such as content moderation [86]. However, many of these tasks also depend on humans making judgments under uncertainty. Considering this, yet another area of research that could be interesting to explore is how communities (with different social dynamics compared to crowd work) can also utilize the tools and processes we built to address uncertainty in a way that may be more transparent or legitimate to stakeholders in these communities.

7.5.3 Uncertainty-Aware AI Systems

Finally, as the main consumer of human judgment data, the field of AI and machine learning has also been tackling challenges related to uncertainty. While there is an increasing amount of models that make use of dis-aggregated data [63] or uncertainty distributions in the form of soft labels [57] to improve training performance, few attempt to model uncertainty as a result of distinct sources. This has led to models that are trained to reflect the judgments of an 'average annotator'. With new tools that allow us to separate the sources of uncertainty in human annotation, there is the potential for the creation of new uncertainty-aware AI systems that make judgments in a way that more faithfully reflects the characteristics of the datasets [106] use to train them—in the form of a collection of judgments from distinct individuals who each experience ambiguity and collectively may disagree.

Chapter 8

CONCLUSION

As we have seen through the research presented in this dissertation, we introduced and studied several novel tools that are aimed at understanding and addressing uncertainty throughout each aspect of human judgments with the crowd. We introduced novel interfaces for conducting annotation in multiple input modalities (scalar rating and categorical classification) and provided a way to distinguish sources of uncertainty through the lenses of ambiguity and disagreement using these new annotation tools. On the other end, we also showed how context and multi-turn deliberation can result in better resolution of disagreement, one of the sources of uncertainty. Finally, we explored how we could take advantage of disentangled uncertainty measurements to build a workflow that can dynamically apply targeted interventions to achieve more efficient and effective reduction of overall uncertainty.

In this chapter, we will summarize our main contributions, outline how our these findings support the thesis statement, and conclude by providing several directions for future research on exploring the question of understanding and addressing uncertainty of human judgments from the crowd.

8.1 Contributions

This dissertation makes contributions towards our understanding in designing new tools and processes that allow us to better understand and address uncertainty in human judgments. chapter 2 examines the existing space of defining, interpreting, measuring, and reducing uncertainty in human judgments. We explore the tools and processes related to uncertainty have been proposed and the issues surrounding lack of adequately addressing uncertainty in the fields of human computer interaction, crowdsourcing and artificial intelligence.

Then with *Goldilocks* (chapter 3), we explore a design that improves upon existing tools for scalar rating annotation. Using a two-step range based annotation along with example-

based anchors, Goldilocks provided better inter-worker consistency while allowing us to separately capture and distinguish *ambiguity* and *disagreement* measurements as part of the annotation process itself. Through experiments conducted on several subjective and/or ambiguous task domains, I find evidence to support that the separate measurement of *ambiguity* and *disagreement* better characterizes the uncertainty by allowing better recovery of pairwise relationships. I also find evidence to support that example-based anchors improved inter-worker consistency when the scale was subjective.

Following that, in chapter 4 we explore *case law crowdsourcing* as a design for an annotation tool that enables us to scale up categorical adjudication on tasks where decision bounds can be based on complex subjective reasoning and can't be efficiently specified through rubrics. By asking adjudicators to construct sets of positive and negative precedents, we can understand the uncertainty around judgments of individuals and groups of adjudicators. I demonstrate through experiments that the precedent-based judgments in case law crowdsourcing produced more consistent judgments compared to traditional rubrics. I also examined how uncertainty measured through precedent sets allowed identification of cases when the precedents were not sufficient.

In chapter 5 we switch gears and examine how the disagreement component of uncertainty can be addressed. Building off of prior work in crowdsourced argumentation tools, we created a novel workflow, Cicero, that uses contextual and multi-turn deliberation to resolve disagreement. We show that this approach results in higher accuracy of the final consensus judgments.

In chapter 6 we tie together the tools for understanding and addressing uncertainty through a dynamic workflow that uses the measurements of *ambiguity* and *disagreement* produced by Goldilocks annotations to inform the choice of interventions targeted to each instance. We found that the interventions of adding context and creating new guidelines through deliberation did correspondingly target the *ambiguity* and *disagreement* sources of uncertainty. We also found that uniform application of an intervention can be ineffective, causing one source of uncertainty to increase while the other lowers. Using simulations we found that applying a targeted intervention can result in a better balance where we don't need to trade off one type of uncertainty reduction for the other.

Finally, in chapter 7 , I discuss some limitations of the systems presented and potential solutions. I also discuss some design implications based on our experience creating these uncertainty tools and workflows.

Overall, in this thesis I make the following contributions:

- **Improving Consistency through Precedents and Anchors**: I present the idea of using past cases to ground complex or subjective scales in the form of precedents and anchors, reducing uncertainty and improving the quality of data annotation work with the crowd.

- **Novel Tools and Interactions for Collecting Human Judgments and Resolving Disagreement**: I present novel interfaces and interactions for collecting judgments in a way that also allows us to distinguish different sources of uncertainty as well as novel workflows to resolve disagreement through multi-turn and contextual deliberation.

- **Empirical Insights about Uncertainty in Human Judgment Tasks**: I provide empirical insights into characteristics and behaviors of uncertainty for a set of human judgment tasks, showing that consistency in interpretation of scales depends on subjectivity of tasks as well as showing that applying the wrong intervention can produce overall harmful results that increase uncertainty.

BIBLIOGRAPHY

[1] Azad Abad, Moin Nabi, and Alessandro Moschitti. Self-crowdsourcing training for relation extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 518–523, 2017.

[2] Abhaya Agarwal and A. Lavie. Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. In *WMT@ACL*, 2008.

[3] Jon D. Agley, Yunyu Xiao, Dr. Rachael D. Nolan, and Lilian Golzarri-Arroyo. Quality control questions on amazon's mechanical turk (mturk): A randomized trial of impact on the usaudit, phq-9, and gad-7. *Behavior Research Methods*, 54:885–897, 2021.

[4] JR Alan M. Jones. Victims of groupthink: A psychological study of foreign policy decisions and fiascoes. pp. iii, 276. boston, mass.: Houghton mifflin, 1972. $4.50. *The ANNALS of the American Academy of Political and Social Science*, 407(1):179–180, 1973.

[5] Jennifer Allen, Cameron Martel, and David G Rand. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in twitter's birdwatch crowdsourced fact-checking program. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[6] Christopher Anderson. Wisdom of the crowds. *Nature*, 2006.

[7] Gabor Angeli, Julie Tibshirani, Jean Wu, and Christopher D. Manning. Combining distant and partial supervision for relation extraction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1556–1567, 2014.

[8] Samuel G. Armato, Geoffrey McLennan, Luc M. Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao, Denise R. Aberle, Claudia I. Henschke, Eric A. Hoffman, Ella Annabelle Kazerooni, Heber MacMahon, Edwin J R Van Beeke, David F. Yankelevitz, Alberto M. Biancardi, Peyton H. Bland, Matthew S. Brown, Roger M. Engelmann, G. E. Laderach, Daniel Max, Richard C. Pais, D. P. Qing, Rachael Y. Roberts, Amanda R. Smith, Adam Starkey, Poonam Batrah, Philip Caligiuri, Ali O. Farooqi, Gregory W Gladish, Cecilia Matilda Jude,

Reginald Munden, Iva Petkovska, Leslie E. Quint, Lawrence H. Schwartz, Baskaran Sundaram, Lori E. Dodd, Charles Fenimore, David Gur, Nicholas A. Petrick, John B. Freymann, Justin S. Kirby, Brian Hughes, Alessi Vande Casteele, Sangeeta Gupte, Maha Sallamm, Michael Heath, M. Kuhn, Ekta Dharaiya, Richard Burns, David Fryd, Marcos Salganicoff, V. Anand, Uri Shreter, Stephen Vastagh, and Barbara Y. Croft. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38 2:915–31, 2011.

[9] Ishaan Arora, Julia Guo, Sarah Ita Levitan, Susan McGregor, and Julia Hirschberg. A novel methodology for developing automatic harassment classifiers for twitter. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 7–15, 2020.

[10] Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1100–1105, 2019.

[11] Lora Aroyo and Chris Welty. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013(2013), 2013.

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

[13] Shubham Atreja, Libby Hemphill, and Paul Resnick. What is the will of the people? moderation preferences for misinformation. *ArXiv*, abs/2202.00799, 2022.

[14] Tal August, Nigini Oliveira, Chenhao Tan, Noah Smith, and Katharina Reinecke. Framing effects: Choice of slogans used to advertise online experiments can boost recruitment and lead to sample biases. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), nov 2018.

[15] Stephanie Alice Baker, Matthew Wade, and Michael James Walsh. The challenges of responding to misinformation during a pandemic: content moderation and the limitations of the concept of harm. *Media International Australia*, 177:103–107, 2020.

[16] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *AAAI Conference on Human Computation & Crowdsourcing*, 2019.

[17] Gagan Bansal, Tongshuang Sherry Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2020.

[18] Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

[19] Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

[20] Birgitta Berglund, Giovanni Battista Rossi, James T. Townsend, and Leslie R. Pendrill. *Measurement with persons : Theory, methods and implementation areas.* Psychology Press, 2011.

[21] Michael S. Bernstein, Joel Brandt, Robert C. Miller, and David R. Karger. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, page 33–42, New York, NY, USA, 2011. Association for Computing Machinery.

[22] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. Soylent: a word processor with a crowd inside. In *UIST '10 Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 313–322. ACM Press, 2010.

[23] Lucas Beyer, Olivier J. H'enaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *ArXiv*, abs/2006.07159, 2020.

[24] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. *Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty*, page 401–413. Association for Computing Machinery, New York, NY, USA, 2021.

[25] Md Momen Bhuiyan, Amy X Zhang, Connie Moon Sehat, and Tanushree Mitra. Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26, 2020.

[26] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and Tom Yeh. Vizwiz: Nearly real-time answers to visual questions. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*, W4A '10, pages 24:1–24:2, New York, NY, USA, 2010. ACM.

[27] Roger Bilisoly. Generalizing the mean and variance to categorical data using metrics. *arXiv: Applications*, 2014.

[28] Abeba Birhane. The impossibility of automating ambiguity. *Artificial Life*, 27:44–61, 2021.

[29] Flora Blangis, Slimane Allali, Jérémie F Cohen, Nathalie Vabres, Catherine Adamsbaum, Caroline Rey-Salmon, Andreas Werner, Yacine Refes, Pauline Adnot, Christèle Gras-Le Guen, et al. Variations in guidelines for diagnosis of child physical abuse in high-income countries: a systematic review. *JAMA network open*, 4(11):e2129068–e2129068, 2021.

[30] Ria Mae Borromeo and Motomichi Toyama. Automatic vs. crowdsourced sentiment analysis. In *Proceedings of the 19th International Database Engineering & Applications Symposium*, IDEAS '15, page 90–95, New York, NY, USA, 2015. Association for Computing Machinery.

[31] Lukas Bossard, M. Guillaumin, and L. Gool. Food-101 - mining discriminative components with random forests. In *ECCV*, 2014.

[32] C. Bouveyron and S. Girard. Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognit.*, 42:2649–2658, 2009.

[33] Vaughn Malcolm Bradley. Learning management system (lms) use with online instruction. *International Journal of Technology in Education*, 2020.

[34] Jonathan Bragg, WASHINGTON EDU, and Daniel S Weld. Learning on the job: Optimal instruction for crowdsourcing. In *ICML Workshop on Crowdsourcing and Machine Learning*, 2015.

[35] Jonathan Bragg, Mausam, and Daniel S. Weld. Sprout: Crowd-powered task design for crowdsourcing. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, UIST '18, page 165–176, New York, NY, USA, 2018. Association for Computing Machinery.

[36] G. Brown, I. Neath, and N. Chater. A temporal ratio model of memory. *Psychological review*, 114 3:539–76, 2007.

[37] Jeffrey M Brunstrom, Nicholas G Shakeshaft, and Nicholas E Scott-Samuel. Measuring 'expected satiety' in a range of common foods using a method of constant stimuli. *Appetite*, 51(3):604–614, 2008.

[38] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT*, 2018.

[39] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1101–1110, 2008.

[40] Chris Callison-Burch, M. Osborne, and Philipp Koehn. Re-evaluation the role of bleu in machine translation research. In *EACL*, 2006.

[41] Hancheng Cao, Vivian Yang, Victor Chen, Yu Jin Lee, Lydia Stone, N'godjigui Junior Diarrassouba, Mark E. Whiting, and Michael S. Bernstein. My team will go on: Differentiating high and low viability teams through team interaction. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW3), jan 2021.

[42] J. Mck. Cattell. V.—mental tests and measurements. *Mind*, pages 373–381, 1890.

[43] James Chalmers, Fiona Leverick, and Vanessa E. Munro. Handle with care: Jury deliberation and demeanour-based assessments of witness credibility. *The International Journal of Evidence & Proof*, 26:381–406, 2022.

[44] Eshwar Chandrasekharan and Eric Gilbert. Hybrid approaches to detect comments violating macro norms on reddit. *arXiv preprint arXiv:1904.03596*, 2019.

[45] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25, 2018.

[46] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 2334–2346, New York, NY, USA, 2017. Association for Computing Machinery.

[47] Joseph Chee Chang, Nathan Hahn, Adam Perer, and Aniket Kittur. Searchlens: Composing and capturing complex user interests for exploratory search. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 498–509, New York, NY, USA, 2019. Association for Computing Machinery.

[48] Joseph Chee Chang, Aniket Kittur, and Nathan Hahn. Alloy: Clustering with crowds and computation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3180–3191, 2016.

[49] Quan Ze Chen, Daniel S. Weld, and Amy X. Zhang. Goldilocks: Consistent crowdsourced scalar annotations with relative uncertainty. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021.

[50] Quanze Chen, Jonathan Bragg, Lydia B Chilton, and Dan S Weld. Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[51] Tina Chen, Renran Tian, Yaobin Chen, Joshua E. Domeyer, Heishiro Toyoda, Rini Sherony, Taotao Jing, and Zhengming Ding. Psi: A pedestrian behavior dataset for socially intelligent autonomous car. *ArXiv*, abs/2112.02604, 2021.

[52] Justin Cheng, Jaime Teevan, and Michael S. Bernstein. Measuring crowdsourcing effort with error-time curves. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1365–1374, New York, NY, USA, 2015. ACM.

[53] Lydia B. Chilton, Juho Kim, Paul André, Felicia Cordeiro, James A. Landay, Daniel S. Weld, Steven W Dow, Rob Miller, and Haoqi Zhang. Frenzy: collaborative data organization for creating conference sessions. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014.

[54] Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1999–2008, New York, NY, USA, 2013. ACM.

[55] John Joon Young Chung, Jean Y Song, Sindhu Kutty, Sungsoo Hong, Juho Kim, and Walter S Lasecki. Efficient elicitation approaches to estimate collective crowd answers. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25, 2019.

[56] Andrew P Clark, Kate L. Howard, A. Woods, I. Penton-Voak, and Christof Neumann. Why rate when you could compare? using the "elochoice" package to assess pairwise comparisons of perceived physical strength. *PLoS ONE*, 13, 2018.

[57] Katherine M. Collins, Umang Bhatt, and Adrian Weller. Eliciting and learning with soft labels from every annotator. In *HCOMP*, 2022.

[58] Corinna Cortes and Neil D. Lawrence. Inconsistency in conference peer review: Revisiting the 2014 neurips experiment. *ArXiv*, abs/2109.09774, 2021.

[59] Kate Crawford and Trevor Paglen. Correction to: Excavating ai: the politics of images in machine learning training sets. *AI & SOCIETY*, 36:1399–1399, 2021.

[60] Stephen Crowder, Collin Delker, Eric Forrest, and Nevin Martin. *Introduction to Statistics in Metrology*. Springer, 2020.

[61] Peng Dai, Christopher H. Lin, Mausam, and Daniel S. Weld. Pomdp-based control of workflows for crowdsourcing. *Artificial Intelligence*, 202:52–85, 2013.

[62] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Comput. Surv.*, 51(1), jan 2018.

[63] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 01 2022.

[64] Todd Davies and Reid Chandler. Online deliberation design: Choices, criteria, and evidence. *arXiv preprint arXiv:1302.5177*, 2013.

[65] A. Philip Dawid and Allan Skene. Maximum likelihood estimation of observer errorrates using the em algorithm. *Journal of The Royal Statistical Society Series C-applied Statistics*, 28:20–28, 1979.

[66] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. Zencrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 469–478, New York, NY, USA, 2012. ACM.

[67] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[68] Michael Denkowski and Alon Lavie. Choosing the right evaluation for machine translation: an examination of annotator and automatic metric performance on human judgment tasks. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA, October 31-November 4 2010. Association for Machine Translation in the Americas.

[69] Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554*, 2021.

[70] Terrance Devries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *CVPR Workshops*, 2019.

[71] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 135–143, 2018.

[72] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In *CrowdSearch*, 2012.

[73] Jeff Donahue and Kristen Grauman. Annotator rationales for visual recognition. *2011 International Conference on Computer Vision*, pages 1395–1402, 2011.

[74] Shayan Doroudi, Ece Kamar, Emma Brunskill, and E. Horvitz. Toward a learning science for complex crowdsourcing tasks. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.

[75] Stephen L Dorton, Samantha B Harper, Glory A Creed, and H George Banta. Up for debate: Effects of formal structure on argumentation quality in a crowdsourcing platform. In *International Conference on Human-Computer Interaction*, pages 36–53. Springer, 2021.

[76] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1013–1022, New York, NY, USA, 2012. ACM.

[77] Ryan Drapeau, Lydia Chilton, Jonathan Bragg, and Daniel Weld. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 4, pages 32–41, 2016.

[78] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. A checklist to combat cognitive biases in crowdsourcing. In *AAAI Conference on Human Computation & Crowdsourcing*, 2021.

[79] A. Dumitrache. Crowdsourcing disagreement for collecting semantic annotation. In *ESWC*, 2015.

[80] A. Dumitrache, Lora Aroyo, and Chris Welty. Capturing ambiguity in crowdsourcing frame disambiguation. In *HCOMP*, 2018.

[81] A. Dumitrache, Lora Aroyo, and Chris Welty. Crowdsourcing ground truth for medical relation extraction. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8:1–20, 2018.

[82] Anca Dumitrache, Lora Aroyo, and Chris Welty. Crowdsourcing ground truth for medical relation extraction. *ACM Trans. Interact. Intell. Syst.*, 8(2), jul 2018.

[83] Carsten Eickhoff. Cognitive biases in crowdsourcing. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018.

[84] Joel S. Elson, Douglas C. Derrick, and Gina Scott Ligon. Examining trust and reliance in collaborations between humans and automated agents. In *HICSS*, 2018.

[85] A. R. Fabbri, Wojciech Kryscinski, Bryan McCann, R. Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. *ArXiv*, abs/2007.12626, 2020.

[86] Jenny Fan and Amy X Zhang. Digital juries: A civics-oriented approach to platform governance. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

[87] Daniel Fenner, Benjamin Bechtel, Matthias Demuzere, Jonas Kittner, and Fred Meier. Crowdqc+—a quality-control for crowdsourced air-temperature observations enabling world-wide urban climate applications. In *Frontiers in Environmental Science*, 2021.

[88] Lev Finkelstein, Evgeniy Gabrilovich, Y. Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. *ACM Trans. Inf. Syst.*, 20(1):116–131, jan 2002.

[89] Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online, June 2021. Association for Computational Linguistics.

[90] Paula Fortuna, Juan Soler, and Leo Wanner. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th language resources and evaluation conference*, pages 6786–6794, 2020.

[91] Craig R Fox and Gülden Ülkümen. Distinguishing two dimensions of uncertainty. *Fox, Craig R. and Gülden Ülkümen (2011),"Distinguishing Two Dimensions of Uncertainty," in Essays in Judgment and Decision Making, Brun, W., Kirkebøen, G. and Montgomery, H., eds. Oslo: Universitetsforlaget*, 2011.

[92] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, Yuan Yao, and Shaogang Gong. Interestingness prediction by robust learning to rank. In *ECCV*, 2014.

[93] Ujwal Gadiraju, Besnik Fetahu, and Ricardo Kawase. Training workers for improving performance in crowdsourcing microtasks. In *European Conference on Technology Enhanced Learning*, pages 100–114. Springer, 2015.

[94] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, page 5–14, New York, NY, USA, 2017. Association for Computing Machinery.

[95] Ruijiang Gao and M. Saar-Tsechansky. Cost-accuracy aware adaptive labeling for active learning. In *AAAI*, 2020.

[96] Henry J. Gardner and M. Martin. Analyzing ordinal scales in studies of virtual environments: Likert or lump it! *PRESENCE: Teleoperators and Virtual Environments*, 16:439–446, 2007.

[97] Timnit Gebru, Jamie H. Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64:86–92, 2021.

[98] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *EMNLP*, 2020.

[99] R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. Garbage in, garbage out?: do machine learning application papers in social computing report where human-labeled training data comes from? *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

[100] Mor Geva, Y. Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *ArXiv*, abs/1908.07898, 2019.

[101] Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[102] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

[103] Benjamin Alan Goldstein, Ann Marie Navar, Michael J. Pencina, and John P. A. Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 24:198–208, 2017.

[104] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2015.

[105] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[106] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human*

Factors in Computing Systems, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[107] Shinsuke Goto, Toru Ishida, and Donghui Lin. Understanding crowdsourcing workflow: Modeling and optimizing iterative and parallel processes. In *AAAI Conference on Human Computation & Crowdsourcing*, 2016.

[108] Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[109] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3:1–24, 2019.

[110] Miriam Greis, Thorsten Ohler, Niels Henze, and Albrecht Schmidt. Investigating representation alternatives for communicating uncertainty to non-experts. In *INTERACT*, 2015.

[111] Ralph Grishman. Information extraction: Techniques and challenges. In *Information extraction a multidisciplinary approach to an emerging information technology*, pages 10–27. Springer, 1997.

[112] The Guardian. The facebook files. `https://www.theguardian.com/news/series/facebook-files`, 2017.

[113] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *ArXiv*, abs/1706.04599, 2017.

[114] Danna Gurari and Kristen Grauman. Crowdverge: Predicting if people will agree on the answer to a visual question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3511–3522, 2017.

[115] Kevin A. Hallgren. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in quantitative methods for psychology*, 8 1:23–34, 2012.

[116] S. G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183, 1988.

[117] Peter Hase and Mohit Bansal. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? In *Annual Meeting of the Association for Computational Linguistics*, 2020.

[118] Danula Hettiachchi, Mike Schaekermann, Tristan J. McKinney, and Matthew Lease. The challenge of variable effort crowdsourcing and how visible gold can help. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021.

[119] Martin Hilbert. Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. *Psychological bulletin*, 138 2:211–37, 2012.

[120] Chien-Ju Ho and Ming Yin. Working in pairs: Understanding the effects of worker interactions in crowdwork. *Computing Research Repository, CoRR*, abs/1810.09634, 2018.

[121] Stephen C. Hora. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54:217–223, 1996.

[122] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist species classification and detection dataset. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2017.

[123] Xinlan Emily Hu, Mark E Whiting, and Michael S Bernstein. Can online juries make consistent, repeatable decisions? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[124] Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[125] Ting-Hao Kenneth Huang and Jeffrey P Bigham. A 10-month-long deployment study of on-demand recruiting for low-latency crowdsourcing. In *Proceedings of The fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2017)*, 2017.

[126] Ting-Hao Kenneth Huang, Walter S Lasecki, Amos Azaria, and Jeffrey P Bigham. " is there anything else i can help you with?" challenges in deploying an on-demand crowd-powered conversational agent. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*, 2016.

[127] Zihan Huang, Charles Low, Mengqiu Teng, Hongyi Zhang, Daniel E. Ho, Mark S. Krass, and Matthias Grabmair. Context-aware legal citation recommendation using deep learning. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 2021.

[128] Jordan S. Huffaker, Jonathan K. Kummerfeld, Walter S. Lasecki, and M. Ackerman. Crowdsourced detection of emotionally manipulative language. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[129] Scott Huffman. Search evaluation at google. `https://googleblog.blogspot.com/2008/09/search-evaluation-at-google.html`, 2008.

[130] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110:457–506, 2019.

[131] Jane Im, Amy X Zhang, Christopher J Schilling, and David Karger. Deliberation and resolution on wikipedia: A case study of requests for comments. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–24, 2018.

[132] Abhaya Indrayan and Rajeev Kumar Malhotra. *Medical biostatistics*. Chapman and Hall/CRC, 2017.

[133] Oana Inel and Lora Aroyo. Harnessing diversity in crowds and machines for better ner performance. In *ESWC*, 2017.

[134] Matthew Ingram. Here's why facebook removing that vietnam war photo is so important. *Fortune*, 2016.

[135] Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, page 64–67, New York, NY, USA, 2010. Association for Computing Machinery.

[136] R. L. Rogers J. P. Kincaid, R. P. Fishburne Jr and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. In *Technical report, DTIC Document*, 1975.

[137] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. "did you suspect the post would be removed?" understanding user reactions to content removals on reddit. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–33, 2019.

[138] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27, 2019.

[139] Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R Brubaker. Understanding international perceptions of the severity of harmful content online. *PloS one*, 16(8):e0256762, 2021.

[140] Liwei Jiang, Jena D. Hwang, Chandrasekhar Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. Delphi: Towards machine ethics and norms. *ArXiv*, abs/2110.07574, 2021.

[141] Nan-Jiang Jiang and Marie-Catherine de Marneffe. Investigating reasons for disagreement in natural language inference. *arXiv preprint arXiv:2209.03392*, 2022.

[142] Tao Jin, Pan Xu, Quanquan Gu, and F. Farnoud. Rank aggregation via heterogeneous thurstone preference models. In *AAAI*, 2020.

[143] Audun Josang, Jin-Hee Cho, and Feng Chen. Uncertainty characteristics of subjective opinions. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 1998–2005, 2018.

[144] David Jurgens. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *HLT-NAACL*, 2013.

[145] Sanjay Kairam and Jeffrey Heer. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1637–1648, 2016.

[146] Ece Kamar, Severin Hacker, and Eric Horvitz. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 467–474. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

[147] David R. Karger, Sewoong Oh, and Devavrat Shah. Budget-optimal crowdsourcing using low-rank matrix approximations. In *Conference on Communication, Control, and Computing*, 2011.

[148] Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A Smith, and Daniel S Weld. Genie: A leaderboard for human-in-the-loop evaluation of text generation. *arXiv preprint arXiv:2101.06561*, 2021.

[149] Svetlana Kiritchenko and Saif M. Mohammad. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–817, San Diego, California, June 2016. Association for Computational Linguistics.

[150] Les Kirkup and Robert B. Frenkel. *An Introduction to Uncertainty in Measurement: Using the GUM (Guide to the Expression of Uncertainty in Measurement)*. Cambridge University Press, 2006.

[151] Armen Der Kiureghian and O. Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, 31:105–112, 2009.

[152] Kate Klonick. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.*, 131:1598, 2017.

[153] Kate Klonick. The facebook oversight board: Creating an independent institution to adjudicate online free expression. *Yale Law Journal*, 129(2418), 2020.

[154] Masaki Kobayashi, Hiromi Morita, Masaki Matsubara, Nobuyuki Shimizu, and Atsuyuki Morishima. An empirical study on short- and long-term effects of self-correction in crowdsourced microtasks. In *AAAI Conference on Human Computation and Crowdsourcing, HCOMP*, 2018.

[155] Stefan H Krieger and Katrina Fischer Kuh. Accessing law: An empirical study exploring the influence of legal research medium. *Vand. J. Ent. & Tech. L.*, 16:757, 2013.

[156] Travis Kriplean, Jonathan T. Morgan, Deen Freelon, Alan Borning, and Lance Bennett. Considerit: Improving structured public deliberation. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, pages 1831–1836, New York, NY, USA, 2011. ACM.

[157] Jacques Kruger and David Dunning. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77 6:1121–34, 1999.

[158] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. Structured labeling for facilitating concept evolution in machine learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 3075–3084, New York, NY, USA, 2014. Association for Computing Machinery.

[159] Anand Kulkarni, Matthew Can, and Björn Hartmann. Collaboratively crowdsourcing workflows with turkomatic. In *CSCW*, New York, New York, USA, 2012. ACM Press.

[160] Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R Klemmer. Peer and self assessment in massive online classes. In *Design thinking research*, pages 131–168. Springer, 2015.

[161] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 299–318, 2021.

[162] Mucahid Kutlu, Tyler McDonnell, Tamer Elsayed, and Matthew Lease. Annotator rationales for labeling tasks in crowdsourcing. *Journal of Artificial Intelligence Research*, 69:143–189, 2020.

[163] S. Kwon, Woo Kim, Chan Hyo Bae, Min gyun Cho, Seunghoon Lee, and Neal Dreamson. The identity changes in online learning and teaching: instructors, learners, and learning management systems. *International Journal of Educational Technology in Higher Education*, 18, 2021.

[164] Andrew Stuart Ian Donald Lang and Joshua Rio-Ross. Using amazon mechanical turk to transcribe historical handwritten documents. *Code4Lib Journal*, 1, 2011.

[165] Samuel Läubli, Rico Sennrich, and M. Volk. Has machine translation achieved human parity? a case for document-level evaluation. *ArXiv*, abs/1808.07048, 2018.

[166] Ji-Ung Lee, Jan-Christoph Klie, and Iryna Gurevych. Annotation curricula to implicitly train non-expert annotators. *Computational Linguistics*, 48(2):343–373, 2022.

[167] Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[168] Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE transactions on image processing*, 28(4):1575–1590, 2018.

[169] Tianyi Li, Kurt Luther, and Chris North. Crowdia: Solving mysteries with crowdsourced sensemaking. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):105:1–105:29, November 2018.

[170] Weixin Liang, J. Zou, and Zhou Yu. Beyond user self-reported likert scale ratings: A comparison model for automatic dialog evaluation. In *ACL*, 2020.

[171] Daniel Licht, Cynthia Gao, Janice Lam, Francisco Guzmán, Mona T. Diab, and Philipp Koehn. Consistent human evaluation of machine translation across language pairs. In *Conference of the Association for Machine Translation in the Americas*, 2022.

[172] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[173] Christopher H. Lin, Mausam, and Daniel S. Weld. To re(label), or not to re(label). In *HCOMP*, 2014.

[174] Christopher H. Lin, Mausam, and Daniel S. Weld. Re-active learning: Active learning with relabeling. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 1845–1852. AAAI Press, 2016.

[175] E Allan Lind, John Thibaut, and Laurens Walker. Discovery and presentation of evidence in adversary and nonadversary proceedings. *Michigan Law Review*, 71(6):1129–1144, 1973.

[176] Greg Little, Lydia B. Chilton, Max Goldman, and Robert C. Miller. Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, page 68–76, New York, NY, USA, 2010. Association for Computing Machinery.

[177] Greg Little, Lydia B. Chilton, Max Goldman, and Robert C. Miller. Turkit: Human computation algorithms on mechanical turk. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, page 57–66, New York, NY, USA, 2010. Association for Computing Machinery.

[178] Angli Liu, Stephen Soderland, Jonathan Bragg, Christopher H Lin, Xiao Ling, and Daniel S Weld. Effective crowd annotation for relation extraction. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 897–906, 2016.

[179] Bulou Liu, Yueyue Wu, Yiqun Liu, Fan Zhang, Yunqiu Shao, Chenliang Li, Min Zhang, and Shaoping Ma. Conversational vs traditional: Comparing search behavior and outcome in legal case retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1622–1626, 2021.

[180] Jian Lu, Wei Li, Qingren Wang, and Yiwen Zhang. Research on data quality control of crowdsourcing annotation: A survey. *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 201–208, 2020.

[181] V. K. Chaithanya Manam, Dwarakanath Jampani, Mariam Zaim, Meng-Han Wu, and Alexander J. Quinn. Taskmate: A mechanism to improve the quality of instructions in crowdsourcing. *Companion Proceedings of The 2019 World Wide Web Conference*, 2019.

[182] VK Chaithanya Manam and Alexander J Quinn. Wingit: Efficient refinement of unclear task instructions. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*, 2018.

[183] Andrew Mao, Yiling Chen, Krzysztof Z Gajos, David Parkes, Ariel D Procaccia, and Haoqi Zhang. Turkserver: Enabling synchronous and longitudinal online experiments. *Proceedings of HCOMP*, 12, 2012.

[184] Emily Megan Marshman, Ryan Thomas Sayer, Charles Henderson, Edit Yerushalmi, and Chandralekha Singh. The challenges of changing teaching assistants' grading practices: Requiring students to show evidence of understanding. *Canadian Journal of Physics*, 96:420–437, 2018.

[185] J Nathan Matias. The civic labor of volunteer moderators online. *Social Media+ Society*, 5(2):2056305119836778, 2019.

[186] Toshiko Matsui, Yukino Baba, Toshihiro Kamishima, and Hisashi Kashima. Crowdsourcing quality control for item ordering tasks. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2013.

[187] Tyler McDonnell, Matthew Lease, Tamer Elsayad, and Mucahid Kutlu. Why is that relevant? collecting annotator rationales for relevance judgments. In *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, page 10, 2016.

[188] Aiden R. McGillicuddy, Jean-Grégoire Bernard, and Jocelyn Cranefield. Controlling bad behavior in online communities: An examination of moderation work. In *International Conference on Interaction Sciences*, 2020.

[189] Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. Abuse is contextual, what about nlp? the role of context in abusive language annotation and detection. *ArXiv*, abs/2103.14916, 2021.

[190] Sarah Michaels, Catherine O'Connor, and Lauren B Resnick. Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in philosophy and education*, 27(4):283–297, 2008.

[191] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, nov 1995.

[192] Tanushree Mitra and Eric Gilbert. Credbank: A large-scale social media corpus with associated credibility annotations. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):258–267, Aug. 2015.

[193] Vikram Mohanty, David Thames, Sneha Mehta, and Kurt Luther. Photo sleuth: Combining human expertise and face recognition to identify historical portraits. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 547–557, New York, NY, USA, 2019. Association for Computing Machinery.

[194] Jethro Mullen and Charles Riley. After outcry, facebook will reinstate iconic vietnam war photo. *CNN Business*, 2016.

[195] Michael J. Muller, Christine T. Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, and Casey Dugan. Designing ground truth and the social life of labels. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[196] Raymond S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2:175–220, 1998.

[197] Jon Noronha, Eric Hysen, Haoqi Zhang, and Krzysztof Z. Gajos. Platemate: Crowdsourcing nutritional analysis from food photographs. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 1–12, New York, NY, USA, 2011. ACM.

[198] Jeppe Nørregaard, Benjamin D Horne, and Sibel Adalı. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 630–638, 2019.

[199] Elle O'Brien. iterative/aita_dataset: Praw rescrape of entire dataset, February 2020.

[200] David Oleson, Alexander Sorokin, Greg P Laughlin, Vaughn Hester, John Le, and Lukas Biewald. Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. In *Human Computation Workshop*, page 11, 2011.

[201] Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. *The measurement of meaning*. University of Illinois press, 1957.

[202] Christina A. Pan, Sahil Yakhmi, Tara P. Iyer, Evan Strasnick, Amy X. Zhang, and Michael S. Bernstein. Comparing the perceived legitimacy of content moderation processes: Contractors, algorithms, expert panels, and digital juries. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1), apr 2022.

[203] Alexandra Papoutsaki, Hua Guo, Danaë Metaxa-Kakavouli, Connor Gramazio, Jeff Rasley, Wenting Xie, Guan Wang, and Jeff Huang. Crowdsourcing from scratch: A pragmatic experiment in data collection by novice requesters. In *HCOMP*, 2015.

[204] Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. Don't blame the annotator: Bias already starts in the annotation instructions. *ArXiv*, abs/2205.00415, 2022.

[205] R. Passonneau and Bob Carpenter. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326, 2013.

[206] Desmond U Patton, William R Frey, Kyle A McGregor, Fei-Tzin Lee, Kathleen McKeown, and Emanuel Moss. Contextual analysis of social media: The promise and challenge of eliciting context in social media posts with natural language processing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 337–342, 2020.

[207] Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, 2019.

[208] John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online, July 2020. Association for Computational Linguistics.

[209] Maria Pershina, Bonan Min, Wei Xu, and Ralph Grishman. Infusion of labeled data into distant supervision for relation extraction. In *Proceedings of ACL*, 2014.

[210] Maja Popović. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 234–243, Online, November 2021. Association for Computational Linguistics.

[211] Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. On releasing annotator-level labels and information in datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[212] Vivek Pradhan, Mike Schaekermann, and Matthew Lease. In search of ambiguity: A three-stage workflow design to clarify annotation guidelines for crowd workers. *ArXiv*, abs/2112.02255, 2021.

[213] Vivek Krishna Pradhan, Mike Schaekermann, and Matthew Lease. In search of ambiguity: A three-stage workflow design to clarify annotation guidelines for crowd workers. *Frontiers in Artificial Intelligence*, 5, 2022.

[214] Drazen Prelec. A bayesian truth serum for subjective data. *science*, 306(5695):462–466, 2004.

[215] Drazen Prelec and H. Sebastian Seung. An algorithm that finds truth even if most people are wrong, 2007. Working Paper.

[216] Y. Qin, Xuezhi Wang, Alex Beutel, and Ed Huai hsin Chi. Improving uncertainty estimates through the relationship with adversarial robustness. *ArXiv*, abs/2006.16375, 2020.

[217] Stefan Räbiger, Gizem Gezici, Yücel Saygın, and Myra Spiliopoulou. Predicting worker disagreement for more effective crowd labeling. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 179–188. IEEE, 2018.

[218] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *ArXiv*, abs/1806.03822, 2018.

[219] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[220] Daniela Retelny, Sébastien Robaszkiewicz, Alexandra To, Walter S Lasecki, Jay Patel, Negar Rahmati, Tulsee Doshi, Melissa Valentine, and Michael S Bernstein. Expert crowdsourcing with flash teams. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 75–85. ACM, 2014.

[221] Daniela Retelny, Sébastien Robaszkiewicz, Alexandra To, Walter S. Lasecki, Jay Patel, Negar Rahmati, Tulsee Doshi, Melissa A. Valentine, and Michael S. Bernstein. Expert crowdsourcing with flash teams. *Proceedings of the 27th annual ACM symposium on User interface software and technology*, 2014.

[222] Lionel Peter Robert and Daniel M. Romero. Crowd size, diversity and performance. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015.

[223] Sarah T Roberts. *Behind the screen: The hidden digital labor of commercial content moderation*. University of Illinois at Urbana-Champaign, 2014.

[224] David L. Rosenhan, Sara L. Eisner, and Robert J. Robinson. Notetaking can aid juror recall. *Law and Human Behavior*, 18:53–61, 1994.

[225] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*, 2017.

[226] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2014.

[227] Keisuke Sakaguchi and Benjamin Van Durme. Efficient online scalar annotation with bounded support. *ArXiv*, abs/1806.01170, 2018.

[228] Niloufar Salehi and Michael S. Bernstein. Hive: Collective design through network rotation. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), nov 2018.

[229] Joni Salminen, Fabio Veronesi, Hind Almerekhi, Soon-Gvo Jung, and Bernard J Jansen. Online hate interpretation varies by country, but more by individual: A statistical analysis using crowdsourced ratings. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 88–94. IEEE, 2018.

[230] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen K. Paritosh, and Lora Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[231] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[232] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *ACL*, 2019.

[233] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics.

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

[235] Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. A framework of high-stakes algorithmic decision-making for the public sector developed through a case study of child-welfare. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–41, 2021.

[236] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. Understanding expert disagreement in medical data analysis through structured adjudication. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

[237] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), nov 2018.

[238] Morgan Klaus Scheuerman, Emily L. Denton, and A. Hanna. Do datasets have politics? disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5:1–37, 2021.

[239] João Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, L. Ungar, and Chris Callison-Burch. Chateval: A tool for chatbot evaluation. In *NAACL-HLT*, 2019.

[240] Glenn Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.

[241] Nihar Shah and Dengyong Zhou. No oops, you won't do it again: Mechanisms for self-correction in crowdsourcing. In *International Conference on Machine Learning*, pages 1–10, 2016.

[242] Hawal Shamon and Carl C. Berning. Attention check items and instructions in online surveys with incentivized and non-incentivized samples: Boon or bane for data quality? *Econometrics: Econometric & Statistical Methods - Special Topics eJournal*, 2019.

[243] Christian Siderius, Robel Geressu, Martin C Todd, Seshagiri Rao Kolusu, Julien J Harou, Japhet J Kashaigili, and Declan Conway. High stakes decisions under uncertainty: dams, development and climate change in the rufiji river basin. In *Climate Risk in Africa*, pages 93–113. Palgrave Macmillan, Cham, 2021.

[244] Rachel N Simons, Danna Gurari, and Kenneth R Fleischmann. " i hope this is helpful" understanding crowdworkers' challenges and motivations for an image description task. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26, 2020.

[245] Edwin Simpson and Iryna Gurevych. Finding convincing arguments using scalable bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371, 2018.

[246] Arjun Singh, Sergey Karayev, Kevin Gutowski, and Pieter Abbeel. Gradescope: A fast, flexible, and fair system for scalable assessment of handwritten work. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*, L@S '17, page 81–88, New York, NY, USA, 2017. Association for Computing Machinery.

[247] Adish Singla, Ilija Bogunovic, Gábor Bartók, Amin Karbasi, and Andreas Krause. Near-optimally teaching the crowd to classify. In *International Conference on Machine Learning*, pages 154–162. PMLR, 2014.

[248] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

[249] Eloise C. Snyder. Uncertainty and the supreme court's decisions. *American Journal of Sociology*, 65(3):241–245, 1959.

[250] Robert Soden, Laura Devendorf, Richmond Y. Wong, Yoko Akama, and Ann Light. Modes of uncertainty in hci. *Found. Trends Hum. Comput. Interact.*, 15:317–426, 2022.

[251] Thamar Solorio, Ragib Hasan, and Mainul Mizan. Sockpuppet detection in Wikipedia: A corpus of real-world deceptive writing for linking identities. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1355–1358, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[252] Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. Findings of the WMT 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91, Online, November 2020. Association for Computational Linguistics.

[253] James A Sprowl. Computer-assisted legal research—an analysis of full-text document retrieval systems, particularly the lexis system. *American Bar Foundation Research Journal*, 1(1):175–226, 1976.

[254] Michael St Pierre, Gesine Hofinger, and Robert Simon. *Crisis management in acute care settings: human factors and team psychology in a high-stakes environment.* Springer International Publishing AG, 2016.

[255] N. Stewart, G. Brown, and N. Chater. Absolute identification by relative judgment. *Psychological review*, 112 4:881–911, 2005.

[256] Neil Stewart, Nick Chater, and Gordon D.A. Brown. Decision by sampling. *Cognitive Psychology*, 53(1):1–26, 2006.

[257] George Stoica, Emmanouil Antonios Platanios, and Barnab'as P'oczos. Re-tacred: Addressing shortcomings of the tacred dataset. In *AAAI Conference on Artificial Intelligence*, 2021.

[258] David Q Sun, Hadas Kotek, Christopher Klein, Mayank Gupta, William Li, and Jason D Williams. Improving human-labeled data through dynamic automatic conflict resolution. *arXiv preprint arXiv:2012.04169*, 2020.

[259] Yu-An Sun, Christopher R Dance, Shourya Roy, and Greg Little. How to assure the quality of human computation tasks when majority voting fails. In *Workshop on Computational Social Science and the Wisdom of Crowds, NIPS*, 2011.

[260] Mihai Surdeanu. Overview of the TAC2013 knowledge base population evaluation: English slot filling and temporal slot filling. In *TAC 2013*, 2013.

[261] Nicolas Suzor and Darryl Woodford. Evaluating consent and legitimacy amongst shifting community norms: an eve online case study. *Suzor, Nicolas P. & Woodford, Darryl (2013) Evaluating consent and legitimacy amongst shifting community norms: an EVE Online case study. Journal of Virtual Worlds Research*, 6(3):1–14, 2013.

[262] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online, November 2020. Association for Computational Linguistics.

[263] David Tait and Meredith Rossner. Making sense of the evidence: Jury deliberation and common sense. In David Tait and Jane Goodman-Delahunty, editors, *Juries, Science and Popular Culture in the Age of Terror: The Case of the Sydney Bomber*, pages 249–271, London, 2017. Palgrave Macmillan UK.

[264] Shengpu Tang, Aditya Modi, Michael W. Sjoding, and Jenna Wiens. Clinician-in-the-loop decision making: Reinforcement learning with near-optimal set-valued policies. *ArXiv*, abs/2007.12678, 2020.

[265] R Peter Terrebonne. A strictly evolutionary model of common law. *The Journal of Legal Studies*, 10(2):397–407, 1981.

[266] Merine Thomas, Thomas Vacek, Xin Shuai, Wenhui Liao, George Sanchez, Paras Sethia, Don Teo, Kanika Madan, and Tonya Custis. Quick check: a legal research recommendation system. In *NLLP@ KDD*, 2020.

[267] Craig Thorley, Lara Beaton, Phillip Deguara, Brittany Jerome, Dua Khan, and Kaela Schopp. Misinformation encountered during a simulated jury deliberation can distort jurors' memory of a trial and bias their verdicts. *Legal and Criminological Psychology*, 25:150–164, 2020.

[268] Louis L Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.

[269] TIAN TIAN and Jun Zhu. Max-margin majority voting for learning from crowds. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[270] Yuandong Tian and Jun Zhu. Learning from crowds in the presence of schools of thought. In *KDD*, 2012.

[271] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185:1124–1131, 1974.

[272] Petsiuk V., Siemenn A., Surbehera S., Chin Z., Tyser K., Hunter G., Raghavan A., Hicke Y., Plummer B., Kerret O., Buonassisi T., Saenko K., Solar-Lezama A., and Drori I. Human evaluation of text-to-image models on a multi-task benchmark. *ArXiv*, abs/2211.12112, 2022.

[273] Betty Van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. Challenges for toxic comment classification: An in-depth error analysis. *arXiv preprint arXiv:1809.07572*, 2018.

[274] Anne Marthe van der Bles, Sander van der Linden, Alexandra L. J. Freeman, James Mitchell, Ana B. Galvao, Lisa Zaval, and David J. Spiegelhalter. Communicating uncertainty about facts, numbers and science. *Royal Society Open Science*, 6, 2019.

[275] Helena Vasconcelos, Matthew Jorke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on ai systems during decision-making. *ArXiv*, 2022.

[276] Carl Vogel, Maria Koutsombogera, and Rachel Costello. Analyzing likert scale inter-annotator disagreement. In *Neural Approaches to Dynamics of Signal Exchanges*, 2020.

[277] C. Völker, T. Bisitz, R. Huber, B. Kollmeier, and Stephan M. A. Ernst. Modifications of the multi stimulus test with hidden reference and anchor (mushra) for use in audiology. *International Journal of Audiology*, 57:S104 – S92, 2018.

[278] Jeroen Vuurens, Arjen P de Vries, and Carsten Eickhoff. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11)*, pages 21–26, 2011.

[279] C. Wah, Grant Van Horn, Steve Branson, Subhransu Maji, P. Perona, and Serge J. Belongie. Similarity comparisons for interactive fine-grained categorization. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 859–866, 2014.

[280] Warren E. Walker, Poul Harremoës, Jan Rotmans, Jeroen P. van der Sluijs, M. B. A. Asselt, Paul Janssen, and Martin Krayer von Krauss. Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support. *Integrated Assessment*, 4:5–17, 2003.

[281] Chenguang Wang, A. Akbik, Laura Chiticariu, Yunyao Li, F. Xia, and Anbang Xu. Crowd-in-the-loop: A hybrid approach for annotating semantic roles. In *EMNLP*, 2017.

[282] Dongsheng Wang, Prayag Tiwari, Mohammad Shorfuzzaman, and Ingo Schmitt. Deep neural learning on weighted datasets utilizing label disagreement from crowdsourcing. *Computer Networks*, 196:108227, 2021.

[283] Claire Wardle and Hossein Derakhshan. Information disorder: Toward an interdisciplinary framework for research and policymaking, 2017.

[284] Zeerak Waseem. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas, November 2016. Association for Computational Linguistics.

[285] Bert Weijters, Hans Baumgartner, and Maggie Geuens. The calibrated sigma method: An efficient remedy for between-group differences in response category use on likert scales. *International Journal of Research in Marketing*, 33(4):944–960, 2016.

[286] Galen Weld, Amy X Zhang, and Tim Althoff. Making online communities' better': A taxonomy of community values on reddit. *arXiv preprint arXiv:2109.05152*, 2021.

[287] Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie. The multidimensional wisdom of crowds. In *Advances in neural information processing systems (NIPS)*, pages 2424–2432, 2010.

[288] Chris Welty, Lora Mois Aroyo, and Praveen Kumar Paritosh. A metrological framework for evaluating crowd-powered instruments. In *HCOMP-2019: AAAI Conference on Human Computation*, 2019.

[289] Keenon Werling, Arun Tejasvi Chaganty, Percy S Liang, and Christopher D Manning. On-the-job learning with bayesian decision theory. *Advances in Neural Information Processing Systems*, 28, 2015.

[290] Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. Whose vote should count more: Optimal integration of labels from laberlers of unknown expertise. In *In Proc. of NIPS*, pages 2035–2043, 2009.

[291] Mark E. Whiting, Allie Blaising, Chloe Barreau, Laura Fiuza, Nik Marda, Melissa Valentine, and Michael S. Bernstein. Did it have to end this way? understanding the consistency of team fracture. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

[292] Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 246–253, College Park, Maryland, USA, June 1999. Association for Computational Linguistics.

[293] Meng-Han Wu and Alexander J. Quinn. Confusing the crowd: Task instruction quality on amazon mechanical turk. In *HCOMP*, 2017.

[294] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399, 2017.

[295] S. Yan, H. Wang, X. Tang, Jianzhuang Liu, and T. Huang. Regression from uncertain labels and its applications to soft biometrics. *IEEE Transactions on Information Forensics and Security*, 3:698–708, 2008.

[296] Shuicheng Yan, Huan Wang, Thomas S. Huang, Qiong Yang, and Xiaoou Tang. Ranking with uncertain labels. *2007 IEEE International Conference on Multimedia and Expo*, pages 96–99, 2007.

[297] Hao-Yu Yang, Junling Yang, Yue Pan, Kunlin Cao, Qi Song, Feng Gao, and Youbing Yin. Learn to be uncertain: Leveraging uncertain labels in chest x-rays with bayesian neural networks. In *CVPR Workshops*, 2019.

[298] Alvin Yuan, Kurt Luther, Markus Krause, Sophie Isabel Vennix, Steven P Dow, and Bjorn Hartmann. Almost an expert: The effects of rubrics and expertise on perceived value of crowdsourced design critiques. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, page 1005–1017, New York, NY, USA, 2016. Association for Computing Machinery.

[299] Lotfi A. Zadeh. Fuzzy sets. *Inf. Control.*, 8:338–353, 1996.

[300] Omar F Zaidan, Jason Eisner, and Christine D Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *Proceedings of NAACL and HLT 2007*, 2007.

[301] Amy X Zhang, Jilin Chen, Wei Chai, Jinjun Xu, Lichan Hong, and Ed Chi. Evaluation and refinement of clustered search results with the crowd. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–28, 2018.

[302] Amy X. Zhang and Justin Cranshaw. Making sense of group chat through collaborative tagging and summarization. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), nov 2018.

[303] Amy X Zhang, Bryan Culbertson, and Praveen Paritosh. Characterizing online discussion using coarse discourse sequences. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[304] Ce Zhang, Feng Niu, Christopher Ré, and Jude Shavlik. Big data versus the crowd: Looking for relationships in all the right places. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 825–834. Association for Computational Linguistics, 2012.

[305] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[306] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[307] Yunfeng Zhang, Qingzi Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

[308] Yunxuan Zhang, Li Liu, Cheng Li, and Chen Change Loy. Quantifying facial age by posterior of age comparisons. *ArXiv*, abs/1708.09687, 2017.

[309] Sharon Zhou, Melissa Valentine, and Michael S. Bernstein. In search of the dream team: Temporally constrained multi-armed bandits for identifying effective team structures. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 108:1–108:13, New York, NY, USA, 2018. ACM.

[310] Haiyi Zhu, Steven P. Dow, Robert E. Kraut, and Aniket Kittur. Reviewing versus doing: Learning and performance in crowd assessment. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 1445–1455, New York, NY, USA, 2014. ACM.

[311] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren G. Terveen. Value-sensitive algorithm design. *Proceedings of the ACM on Human-Computer Interaction*, 2:1–23, 2018.

[312] A. Çelikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey. *ArXiv*, abs/2006.14799, 2020.