Wisdom of the Crowd, Without the Crowd: A Socratic LLM for Asynchronous Deliberation on Perspectivist Data

MALIK KHADAR*, University of Minnesota, USA
DANIEL RUNNINGEN*, University of Minnesota, USA
JULIA TANG, University of Minnesota, USA
STEVIE CHANCELLOR†, University of Minnesota, USA
HARMANPREET KAUR†, University of Minnesota, USA

Data annotation underpins the success of modern AI, but the aggregation of crowd-collected datasets can harm the preservation of diverse perspectives in data. Difficult and ambiguous tasks cannot easily be collapsed into unitary labels. Prior work has shown that deliberation and discussion improve data quality and preserve diverse perspectives—however, synchronous deliberation through crowdsourcing platforms is time-intensive and costly. In this work, we create a Socratic dialog system using Large Language Models (LLMs) to act as a deliberation partner in place of other crowdworkers. Against a benchmark of synchronous deliberation on two tasks (Sarcasm and Relation detection), our Socratic LLM encouraged participants to consider alternate annotation perspectives, update their labels as needed (with higher confidence), and resulted in higher annotation accuracy (for the Relation task where ground truth is available). Qualitative findings show that our agent's Socratic approach was effective at encouraging reasoned arguments from our participants, and that the intervention was well-received. Our methodology lays the groundwork for building scalable systems that preserve individual perspectives in generating more representative datasets.

CCS Concepts: • **Human-centered computing** → *Computer supported cooperative work*; *Empirical studies in HCI*; • **Information systems** → *Decision support systems.*

Additional Key Words and Phrases: Annotation, Crowdsourcing, Deliberation, Large Language Models, Socratic Dialogue

ACM Reference Format:
Malik Khadar, Daniel Runningen, Julia Tang, Stevie Chancellor, and Harmanpreet Kaur. 2025. Wisdom of the Crowd, Without the Crowd: A Socratic LLM for Asynchronous Deliberation on Perspectivist Data. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW526 (November 2025), 35 pages. https://doi.org/10.1145/3757707

*Both authors contributed equally to this paper.
†Both senior authors contributed equally to this paper.

Authors' Contact Information: Malik Khadar, khada005@umn.edu, University of Minnesota, Department of Computer Science & Engineering, Minneapolis, USA; Daniel Runningen, runni028@umn.edu, University of Minnesota, Department of Computer Science & Engineering, Minneapolis, USA; Julia Tang, jyt@umn.edu, University of Minnesota, Department of Computer Science & Engineering, Minneapolis, USA; Stevie Chancellor, steviec@umn.edu, University of Minnesota, Department of Computer Science & Engineering, Minneapolis, USA ; Harmanpreet Kaur, harmank@umn.edu, University of Minnesota, Department of Computer Science & Engineering, Minneapolis, USA.

ACM 2573-0142/2025/11-ARTCSCW526
https://doi.org/10.1145/3757707

1 Introduction

Data annotation by human crowds is the backbone for the success of many AI systems. Annotators have worked on tasks ranging from image labeling [96] and straightforward content classification [80] to more subjective domains, such as content moderation [55, 82] and toxicity detection [28]. In many cases, large teams of annotators provide individual labels, and their collective opinions are condensed to a single label [28].

When annotating data for AI, there is a tension between producing datasets that result in accurate, generalizable models, and producing datasets that maintain diverse perspectives. Common data curation methods often remove nuance or disagreement in annotations, such as by creating "ground truth" datasets using majority voting [6, 28, 103], domain- or expertise-weighted voting [39, 107], crowd consensus via post-hoc clustering [7, 114], and "noise removal" strategies [66]. Annotations are standardized in a way that does not represent different perspectives. Moreover, those standardized labels risk perpetuating existing systemic harms both directly [16, 100, 101] and indirectly [11, 18, 99].

Treating all differences as noise erases the uncertainty inherent in many kinds of annotation tasks—uncertainty that is relevant if we want models to work for everyone, not just the majority. In response, perspectivist approaches encourage annotators to reflect on their opinions, consider their own and others' perspectives, and discuss *how* they came to their annotation decisions [16, 41]. The term *perspectivism* stems from the belief that data can have multiple "correct" interpretations, and that removing voices from a dataset discards nuance and propagates bias [16, 18, 61]. Prior work suggests that perspectivist approaches to data annotation and model building capture the inherent ambiguity of classification tasks more effectively than their traditional counterparts [27, 43, 73]. Therefore, to enable AI models to take more nuanced approaches to problems, we also need datasets that respect nuanced opinions.

Synchronous deliberation between annotators has successfully supported perspectivist approaches for data annotation, but synchronous work has significant resource costs. Design and social science theories argue that deliberation with others richly supports engagement between differing perspectives [33, 88]. In practice, research on crowdsourcing has confirmed the value of deliberation in resolving disagreements [102] and promoting higher-quality annotations through systems like MicroTalk [34], Revolt [19], and Cicero [22]. However, synchronous deliberation has drawbacks—it demands significantly more time from participants, is challenging to facilitate due to dropout, and is costly. For example, participants in Schaekermann et al. [102] spent as much as four hours annotating and synchronously deliberating 10 datapoints. Other studies show similar time and cost considerations [19, 24].

In this paper, we introduce a Socratic LLM dialogue system for asynchronous deliberation in data annotation, reaping the perspectivist benefits of synchronous deliberation without the prohibitive costs. Our system facilitates a Socratic dialogue with annotators, prompting them to reflect on and defend their annotation choices [1, 2, 31]. Socratic dialogues have been shown to be effective in facilitating meta-cognitive tasks like reflection in educational domains [12, 49, 64]. We build on the argumentative benefits identified in prior crowdsourcing and deliberation work [22, 34, 102] and embody these in an asynchronous workflow using an LLM. As an added benefit, we see this work as a positive use case of LLMs, highlighting a responsible application in a landscape where LLMs are here to stay, yet best practices are still evolving. By limiting the LLM to a Socratic persona, our approach has built-in guardrails to center annotator perspectives and avoid known problems of using LLMs (*e.g.*, hallucinations, misrepresentation, and other social biases [9, 75]).

To demonstrate the potential of the Socratic LLM for improving perspectivist data annotation, we benchmark the performance of our Socratic LLM against the synchronous deliberation approach of Schaekermann et al. [102]. We mirror the experimental approach and dataset selection against this work because they provide ample implementation details and publicly released their results for follow-up work [102]. They test their approach on two datasets well-suited to study the consequences of deliberation on annotation—an inherently ambiguous Sarcasm detection task [36] and a comparatively objective semantic Relation task [34]. We compare the two approaches using quantitative metrics such as the number of annotations that change due to deliberation, annotator confidence pre- *vs.* post-deliberation, ground truth accuracy on the Relation task, number and length of dialogues exchanged, and task experience measured with task load index [48]. We also report results from qualitatively coding the conversation logs generated from interactions with our system and the responses to our post-study experiential questions.

Our results show that our Socratic LLM performs better than the synchronous deliberation approach of our benchmark on several key metrics. With our approach, participants were more likely to change their labels on the Relation task compared to the benchmark, with these changes resulting in more accurate annotations (64.79% post-deliberation accuracy for us *vs.* 48.86% for benchmark); no difference in the Sarcasm task. Not only were participants more accurate, but they were also more confident about their labels after discussion with our Socratic LLM: 28% of participants switched to higher confidence in our case, compared to no change in confidence after deliberation in the benchmark. Finally, our participants were more engaged in their conversations as the amount and lengths of their messages exceeded those in the benchmark. Our qualitative analysis of conversation logs showed the Socratic LLM played various important roles in deliberation: as an argument evaluator that questioned annotators' claims, evidence, and warrant; as a negotiator of different class boundaries relevant to the task; as a cognitive support tool that re-articulated prior reasoning; and as a validator that signaled when sufficient levels of deliberation had occurred.

We highlight the benefits of using a Socratic LLM to assist annotators during the labeling process, proposing the technique as a scalable method for obtaining perspectivist data annotations. Thus, our key contributions are:

(1) The development of a theoretically grounded Socratic LLM for asynchronous deliberation and insights on its use for a data annotation task.
(2) Quantitative evidence demonstrating the value of our Socratic LLM: improvements in task performance and costs compared to a synchronous deliberation benchmark.
(3) Qualitative evidence synthesizing the potential roles for Socratic LLMs in annotation deliberation workflows.

Finally, we also discuss connections with HCI theories like distributed cognition and boundary objects, articulate the costs and benefits of asymmetric deliberation, consider ethical implications of systems like ours, and present design implications for data annotation tasks and positive LLM roles for future work.

2 Related Work

In this section, we review prior work in three areas: 1) collective annotation, disagreements, and perspectivism; 2) deliberation in annotation; and 3) LLM applications and Socratic dialogue systems. While deliberation and discussion between annotators yield rich datasets that preserve diverse perspectives, coordinating interactions among disjoint annotators introduces immense costs. We posit that deliberation with Socratic LLMs can alleviate the logistical difficulty of taking a perspectivist approach to data annotation and minimizing coordination issues.

2.1 Collective Annotation

Using crowdsourced data annotation to label datasets used by AI systems is a widespread practice [8, 29, 67]. However, researchers recognized that there were challenges with the quality crowdsourced labels, due in part to bad data and a belief that crowdworkers lacked expertise [26, 46, 59, 90, 119]. As a result, crowdsourcing tasks began to gather multiple annotations for each datapoint in a dataset to improve quality control [44, 74, 79, 95].

To handle multiple annotations per datapoint, data curators made a necessary assumption: when decisions were not unanimous, annotations were viewed as erroneously "noisy" [45, 54, 122]. Recently, this assumption has been challenged [76, 101] as science investigates how to distinguish between bad-faith errors, accidents, uncertainty, and explicit disagreement [40, 100]. Thus, this section reviews the research on resolving annotation *errors* versus *disagreements.*

2.1.1 Addressing Errors in Annotations. Prior work has attempted to identify and resolve noisy data caused by errors [45, 54, 122]. Frénay and Verleysen [40] propose a taxonomy of noise and methods to reduce it. For example, researchers have investigated ways to improve the quality of codebooks to mitigate the mistakes caused by task confusion and uncertainty [13, 60, 70]. Some studies allow the annotators to provide more detailed labels than what practitioners predicted would be necessary (*e.g.* by adding sub-classes for annotation) [19, 91]. In other cases, the innovation is in communicating uncertainty itself. When dealing with binary choices or sliding scales, Chen et al. [23] devised a way for annotators to embed their level of certainty in the label. Finally, researchers have developed algorithms using noisy data to train models to reflect the crowd's collective confidence [78]. Most of these methods, however, only address the problem of "error" when it is defined as something to be removed or dealt with.

2.1.2 Addressing Annotation Disagreement via Perspectivism. We argue that not all noise is erroneous. Indeed, there are longstanding critiques of attempts to resolve differences in annotation as error [16, 56]. Our approach stems from these critiques, and in this section, we overview the "perspectivist turn" in data annotation.

A growing body of work questions the assumptions of a single ground truth always existing in annotation [8, 16, 37, 94], and proposes that, instead, multiple perspectives are valid when annotating data. The term *perspectivism* is grounded in the belief that data can have various "correct" interpretations and that removing voices from a dataset discards nuance that can damage the overall annotation task [16, 18]. For example, existing content moderation models have been shown to disproportionately affect minority groups because of a lack of diversity and missing social context in the training data [11, 98]. The impact of perspectivism reaches far beyond social media platforms [20, 35]. Perspectivist approaches to annotation and model building enable system designs to be sensitive to social biases [35]. There is evidence to suggest that *perspectivist* approaches to model building *perform better* than their traditional counterparts [27, 42, 43, 73].

A perspectivist approach can be applied at the annotation level (*e.g.* by soliciting more diverse annotations [57, 123]), or at the algorithm level. At the annotation level, perspectivism acknowledges that data could be interpreted differently and be simultaneously correct depending on the viewpoint. It also means keeping the disagreeing annotations in the resulting datasets instead of resorting to aggregation methods [43]. At the algorithm level, perspectivism means developing methods for equitably representing these annotation differences in models. Approaches like Wallace et al. [112] suggest different weighting for models that emphasize nuanced perspectives. Meanwhile, Gordon et al. [42] take advantage of diverse annotations to bootstrap a sample population of mock "jurors" for deriving a final decision. The one thing these model-building and training approaches have in common is the need for non-aggregated datasets to be made available for research, such as the datasets provided by Schaekermann et al. [102]. Without them, progress on perspectivist approaches will be slow to develop [16, 37, 94].

We build on this work by proposing a new method for improving annotation-level consideration of perspectives: one that preserves the voices of all annotators of a dataset. Our method extends prior work on capturing diversity using deliberation, discussed below.

2.2 Deliberation and its Impact on Annotation

Deliberation has been proposed to solve sociotechnical problems, where multiple people work together to decide on the outcome of a task or action. Deliberation has been heavily studied in social computing systems because of its relevance to human communication, reasoning, and decision-making (*e.g.* [70, 121]). Of relevance to our work on perspectivism is deliberation that does not resolve cleanly and that embodies dissensus or disagreement as a reasonable outcome [33, 88]. Thus, in this section, we explore deliberation and dialogue, when it can lead to disagreement, and what the research says regarding data annotation tasks.

2.2.1 Deliberation in Annotation Work. Crowdwork uses discussion between annotators and within teams to resolve disagreements in annotations in a variety of contexts [19, 47, 69]. This is because deliberation improves the quality and explainability of the resulting data [25], and helps derive a correct answer when the individuals did not know one beforehand [63]. This type of intervention can be beneficial in simple cases of misunderstanding, but vexing when annotators hold fundamentally different beliefs. Examining this specifically, Schaekermann et al. [102] studied the impact discussion had on determining if a consensus could be reached on any given annotation item, calling it resolvable *vs.* irresolvable disagreement. Their results support that reaching consensus is influenced by the nature of the task, but the act of deliberation improves task quality regardless.

Closest to our work are systems that facilitate dialogue and deliberation for annotation. For example, Drapeau et al. [34]'s system Microtalk helped annotators supply additional reasoning for their labels on a relation task. Though MicroTalk did not involve direct discussion, it presented annotators with counter-reasoning in situations where others had already labeled differently. Chang et al. [19] expanded on the design of MicroTalk [34] to allow annotators to better describe the items where they disagreed. In doing so, they helped develop a workflow, Revolt, that generated richer guidelines to describe a dataset. Schaekermann et al. [102] built on MicroTalk [34] and Revolt [19] to investigate if disagreements between annotators could be reliably resolved through synchronous discussion. Their solution involved time-boxed stages to coordinate workers based on their disagreements for the items they labeled. Finally, Chen et al. [22] avoid time-boxed stages by dynamically pairing individuals working on the same task from the start.

These examples of prior work show the benefits of explicit reasoning and deliberation between annotators, but face challenges regarding worker coordination. We build on this work by eliminating the need for coordinating annotator effort, which allows all annotators to complete their tasks independently. We hope to address the coordination and time constraints by using an LLM as a stand-in discussion partner, benchmarking against the publicly available datasets and results of Schaekermann et al. [102].

2.3 LLMs as Discussion Partners and Socratic Dialogues

The recent success of LLMs in research, business, and education applications is partly due to their conversational abilities. Chatbots are now supporting decision-making as mediators for multi-stakeholder applications [104]. Other chatbot research encourages debate with users on video platforms about content to promote media literacy [106]. LLMs are even being used to help guide students through their learning [3, 64] or perform assessments on their understanding [52]. Prior work has also begun to explore using conversational LLMs to independently perform and assist with the labeling of datasets [81, 117], but these approaches struggle with the same bias propagation issues as label aggregation methods [16]. We believe in a critical optimist approach to using LLMs: with appropriate guardrails and constrained task responsibilities, we see value in using LLMs for their language generation capabilities. Our approach considers how we can use an LLM to provide a means of reflection with a structured Socratic dialogue.

Many applications of LLMs describe the use of the Socratic method as part of their prompt design considerations — in debugging code [3], teaching math [32], conducting oral assessment in classrooms [52], and evaluating programming assignments [68]. These classroom-oriented applications come as no surprise, given the studies in educational settings which have long benefited from the use of the Socratic method to promote critical thinking and reflection [5, 38, 111]. Because this approach to dialogue is well-established in its use for LLM prompting and meta-cognitive benefits, we find it an appropriate framework for exploring deliberation in dataset annotation.

We build on prior work by designing a Socratic LLM for an unstructured reasoning task (*i.e.*, annotating subjective and objective data), giving individuals a mechanism to think more critically about their annotations. The following section provides details on the Socratic method and its incorporation into our approach.

3 Socratic LLM: Design and System

In this section, we describe our Socratic LLM system design and implementation. This includes design goals and principles, prompt engineering strategies, and finally, an architecture overview of our system implementation.

3.1 Design Goals

A major bottleneck of deliberation is the need for two or more people to actively and synchronously engage in discussion [19, 22]. Additionally, designers need to ensure the mode of deliberation is appropriate for the task at hand, such as debates [113] or panel [116] formats.

We developed a **Socratic LLM** to act as an inquisitor from the Socratic Method to address these challenges. Our goal was to provide users with a conversational partner that provokes moments of reflection while limiting auxiliary input. First, our Socratic LLM helps lower logistical constraints by supporting asynchronous deliberation with an individual annotator. Second, recent work in programming education (*e.g.*, [3, 64, 68]) and mathematics [32] has demonstrated the efficacy of LLMs to act as a Socratic *instructor and inquisitor* to support meta-cognitive skills. We conceptualize this process similarly to Hung et al. [52]'s work using a Socratic agent in educational settings. Our application is similar to education domains by engaging in higher-order cognitive processes to facilitate an outcome instead of serving as a question-answering system [2, 5, 12, 31, 38].

Our Socratic LLM facilitated the latter half of the Socratic Method [1]. The crowdworker independently completed the two initial stages—*wonder* and *hypothesis*—before engaging with the Socratic LLM in *elenchus*: "a form of cooperative argumentative dialogue between individuals, based on asking and answering questions to stimulate critical thinking and to draw out ideas and underlying presumptions" [1]. Then, we asked participants to engage in *approving or revising of hypothesis:* if the hypothesis fails to stand up to scrutiny, the discussion returns to elenchus. Once the hypothesis is approved, participants can move to the final stage of *application*, where the new hypothesis can be used to revise one's understanding or behavior. Figure 1 shows how the Socratic Method maps onto our annotation process.

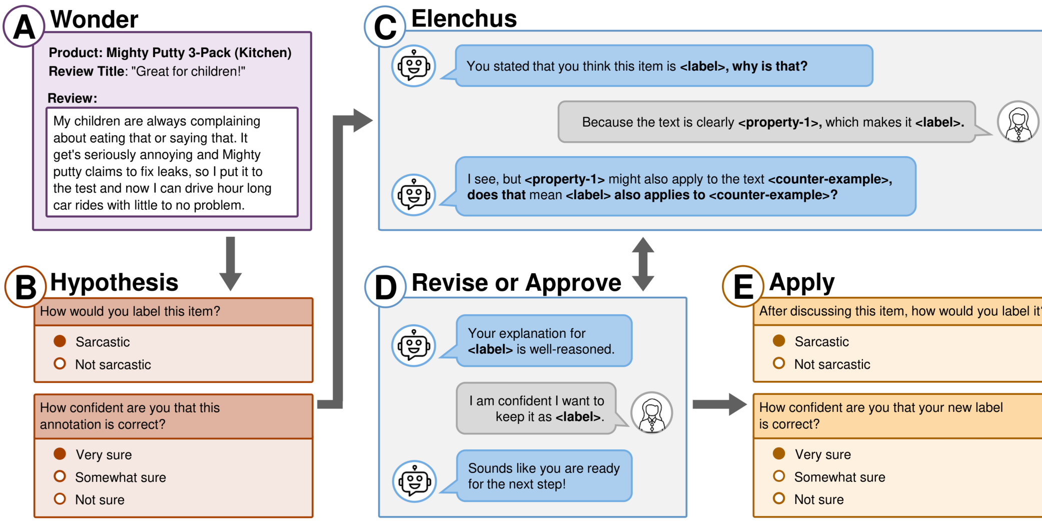

Fig. 1. **Socratic Method in practice.** Participants progress through the phases of the Socratic Method as they complete annotation tasks. (A) Wonder, the participant considers the datapoint provided. (B) Hypothesis, participants generate a hypothesis based on the data and respond to questions to record their annotation. (C) Elenchus, engagement with the Socratic LLM encourages participants to reflect on their hypothesis and explain their position. (D) Revise or Approve, if their hypothesis does not withstand scrutiny, then participants revise it and return to the discussion style of Elenchus. Otherwise, their hypothesis is collectively approved. (E) Apply, participants record their final annotation based on their approved hypothesis.

3.2 Prompt Engineering

To ensure a successful dialogue, we designed three components in our prompt: instructions for performing the elenchus, guidelines on a "Socratic Temperament" [1], and general guardrails to keep responses manageable. The full, human-readable prompt is provided in Appendix A.

Socratic Processes. First, prompt engineering and development focused on the process of elenchus, the cooperative dialogue that interrogates the participants' hypothesis. The goal of the Socratic dialogue system is to help revise and approve the hypothesis before the user applies the result of their hypothesis during re-annotation. To do this, five steps were provided in the system prompt to the LLM, mapping to prior work on Socratic systems (*e.g.* [1, 52]). The first step expects the participant to assert a claim based on their label. This step is followed by the dialogue system interrogating the claim, either by asking clarifying questions or prompting further reflection to illustrate potential gaps in logic. The participant is then expected to engage with the dialogue system's line of inquiry. If inconsistencies are identified, the Socratic LLM should ask the participant to update their claim and restart with the new claim. If the reasoning is sound, the participant should be encouraged to affirm their hypothesis using the annotation interface and move forward.

Temperament. Second, we wanted the Socratic LLM to maintain a reasonable temperament that supported inquisitive engagement. To that end, our prompt outlined four primary traits that emerged through pilot testing: humility, respect, joy in the dialogue, and prioritizing mutual understanding. Temperament is key for dialogue systems like ours [21], and we were mindful that the process of elenchus can come across as overly and unintentionally forceful or manipulative [1]. The Socratic LLM needed to remain impartial to ensure that the dialogue was centered on the participant's developing hypothesis rather than biases expressed by the Socratic LLM.

Guardrails. Finally, we added guardrails and rules to the prompt to ensure the responses were manageable for participants in a crowdworking task. We instructed the LLM not to use outside knowledge; we did not want to bias annotators with ground truth data available through the LLM's internal datasets or outside resources it had scraped (*e.g.*, Wikipedia data may confirm a Relation annotation). Through extensive pilot testing, we also identified structural rules necessary to keep the Socratic LLM's responses dialogic rather than prescriptive, ensure good conversation practices, and avoid unwanted behaviors. For example, we included low-level instructions about the interaction format, like limiting the responses to two or three sentences, as LLM responses can otherwise be quite long [89]. Other rules helped constrain the conversation to promote the turn-taking intentions of the Socratic process and prevent the Socratic LLM from getting off-task.

3.3 Architecture

LLM. Our LLM was built with Anthropic's Claude 3 Haiku model[1]. The prompt underwent six major iterations before pilot testing showed behavior that sufficiently reflected the Socratic process.
System Architecture. Our system was built on the Express framework[2] for the back-end and VueJS framework[3] for the user interface, with styling implemented via Bootstrap libraries[4]. Data storage was handled by an instance of MySQL server[5] and retrieved through RESTful API calls with our back-end. Our LLM was integrated as a service in Express using Anthropic's API interface.[6]

4 Experimental Design

We based our experimental design on our intended benchmark, Schaekermann et al. [102], which examined how synchronous deliberation helps crowdworkers resolve disagreements on data annotation tasks. Though we discuss several studies from prior work that are related to our goals (see Section 2.2.1), Schaekermann et al. [102]'s work uniquely affords benchmarking given their publicly-available data and results. We designed a study to evaluate the impacts of our Socratic LLM on annotation outcomes, mirroring the experimental design and artifacts used by our benchmark. Our research was reviewed and approved by the University of Minnesota's Institutional Review Board (IRB).

4.1 Benchmark Datasets

We used the following two datasets from Schaekermann et al. [102], each containing 40 items which we refer to as "datapoints."

Sarcasm Dataset. The first dataset asked annotators to identify whether an Amazon product review was sarcastic. Sarcasm detection is considered a subjective task because there is no universal definition of sarcasm, making it well-suited for deliberation. For this task, Schaekermann et al. [102] built their dataset from the sarcasm detection of Filatova [36], filtering their dataset to identify the labels with the highest inter-rater disagreement. An example from this dataset is included in our system screenshots under Appendix C.

Relation Dataset. The second dataset asked annotators to identify whether a particular relation between two objects in a sentence was present. Depending on the datapoint, the relation would either be "lived in" or "died in." This is a common benchmark in deliberation tasks [22, 34] and is considered objective because there are official labeling guidelines on *LivedIn* and *DiedIn* relations

[1]https://www.anthropic.com/news/claude-3-haiku
[2]https://expressjs.com
[3]https://vuejs.org
[4]https://getbootstrap.com
[5]https://www.mysql.com
[6]https://www.anthropic.com/api

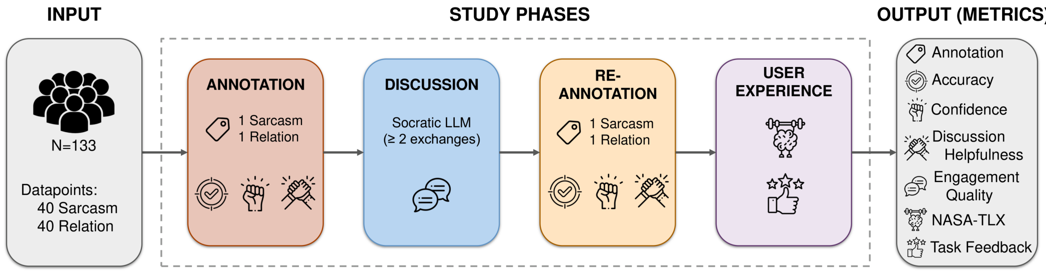

Fig. 2. Workflow diagram of the Socratic LLM-assisted annotation process. Crowdworkers (N=133) first performed initial labeling tasks for Sarcasm and Relation detection, followed by an AI-driven discussion phase using Socratic questioning. The crowdworkers then reconsidered their annotations and confidence levels before completing NASA-TLX and experience assessments.

published by the Linguistic Data Consortium.[7] However, individual datapoints can be ambiguous and thus lack a ground truth label. Our specific dataset was taken from MicroTalk [34] and modified by our benchmark [102] to clarify the task instructions. One example from this dataset asked if the relationship "Litvinenko 'lived in' London" was expressed in the sentence "Litvinenko died of radioactive poisoning, from Polonium 210, in his home in London."

4.2 Study Task

Our study was split into four phases: (1) an annotation phase where participants labeled datapoints from the datasets; (2) an asynchronous discussion phase with our Socratic LLM, where participants reasoned about their annotations; (3) a re-annotation phase where they had the opportunity to reconsider their labels; and (4) a user experience survey. Figure 2 includes an overview of these phases. The exact wording and setup of the questions can be found in Appendix B, and example screenshots can be found in Appendix C.

4.2.1 Annotation Phase. First, participants were asked to annotate two randomly-selected datapoints, one each from the **Sarcasm** and **Relation** datasets. Only one datapoint was displayed at a time. The task included the same context and instructions as our benchmark for consistency. We randomly sampled from each dataset until at least three participants had labeled each datapoint. After annotating each datapoint, participants rated: (1) their confidence that their annotation was correct (options: "Very Sure", "Somewhat Sure", and "Not Sure"); and (2) whether having a discussion would reduce their uncertainty about their label (options: "Yes" or "No").

We also placed attention check questions among those listed above, one for each annotation (see Appendix B.2 for the two attention checks). If participants failed both attention checks, they were removed from the study, and their data was excluded from analysis. After annotating, participants were asked to confirm that they wished to proceed and warned that they would not be able to return to this phase.

4.2.2 Discussion Phase with Socratic LLM. Participants then entered the Discussion Phase, engaging with the Socratic LLM. In addition to reviewing their datapoint and the annotation instructions, participants had a chat interface to converse with the dialogue system (see Figure 6 in Appendix C). They were instructed to discuss their reasoning for each annotation. Thus, participants engaged in two distinct chat sessions with the Socratic LLM, one for each datapoint.

[7] https://tac.nist.gov/2017/KBP/index.html

Participants could submit as many messages as desired to our Socratic LLM but were required to respond at least twice to each discussion before proceeding. This mirrored the minimum participation required from collaborative discussions in our benchmark.

4.2.3 Re-annotation Phase. After chatting with our Socratic LLM, participants were allowed to re-annotate each datapoint, with clear communication that changing their labels was not required. This phase was critical in capturing the impact of our asynchronous discussion and included questions about both the (re-)annotation and the influence of the conversation (see Appendix B.3 for exact questions). After the re-annotation phase of the first datapoint, we shared a short 12-second GIF[8] as a transition to give participants a short break before discussing the second datapoint.

4.2.4 User Experience. Finally, participants were directed to answer experiential questions in a Qualtrics survey.[9] These questions were focused on prior deliberation experience and task difficulty to help lay the foundations of comparison for future work.

We measured task difficulty using five of the NASA Task Load Index (NASA-TLX) questions [48]. We omitted the sixth question on the physical demand of the task, following guidance from prior work where NASA-TLX has been used to compute task load in computational settings [17, 58, 65]. Additionally, we asked several Likert-scale questions to measure current and prior annotation experience, and to determine the perceived helpfulness of the Socratic LLM; yes/no questions about opinions on and experiences with deliberation for annotation; and open-ended questions for feedback about the experience (see Appendix B.4 for exact questions).

4.3 Participants, Consent, and Procedure

We recruited 133 participants on Prolific to participate in our experiment.[10] Prolific is an online research platform where participants can be recruited for crowdsourcing tasks. The study was released in batches, with approximately 30 slots in each batch to allow for manual verification of work and for observation of system functionality.

We used Prolific's built-in screening mechanisms to filter participants by location (restricted to the United States) and English language fluency. Our participants ranged in age from 18 to 70 (mean = 37.23, sd = 10.87). 56.39% of participants identified as female through the Prolific platform, 42.86% as male, and 0.75% selected "Prefer not to say". 61.65% of the sample identified themselves as White, 21.05% as Black, 7.52% as Asian, 7.52% as Mixed, and 2.26% as Other.[11]

To receive compensation of $5 USD, participants had to meet our selection criteria, opt into participation on Prolific, confirm consent via a linked Qualtrics survey, and complete all study tasks. We disqualified participants for: 1) failing our attention checks; 2) when there were clear signs of misconduct, such as willfully ignoring task instructions; 3) providing responses generated from outside sources like ChatGPT; or 4) when their task completion speed for their provided responses was much too high to have been typed by hand. Participants could opt out of this study at any time until payment was submitted.

Based on pilot testing, we estimated this task would take approximately 20 minutes to complete. On average, participants spent 20.5 minutes (sd = 11.7) on the study, with 15.4 minutes (sd = 8.3) spent on the annotation and deliberation sections. We targeted our compensation at $15 USD/hr.

[8]https://media1.tenor.com/m/2EXMLYXzJAYAAAAd/dog-stand-up-dog-wake-up.gif
[9]https://www.qualtrics.com
[10]https://www.prolific.com
[11]Prolific's platform provides only the options listed in this section to describe a user's legal sex and ethnic group. We recognize that these categories may not fully represent the identities of our participants and note this as a limitation.

4.4 Analysis and Baseline Comparisons

Our analysis involved quantitative evaluations of the participants' annotations and qualitative assessments of their messages with the LLM and their experiences with the study.

The quantitative data we collected was directly compared to data and evaluations provided by Schaekermann et al. [102]. Using this prior work and their metrics as a benchmark, we analyzed:

(1) **Changes in Annotations.** The frequency of annotation changes after discussions. This includes changes in annotations generated by an individual before *vs.* after discussion, and across all annotations for a datapoint.
(2) **Accuracy of Relation Data Annotations.** The accuracy of final annotations for datapoints where ground truth was available for the Relation dataset.
(3) **User Confidence.** Participants' confidence in their annotations before *vs.* after discussion.
(4) **Engagement Quality.** The length and number of messages sent by individual participants.

The benchmark did not always have a perfectly matched population to ours to serve as a baseline due to the constraints associated with synchronous deliberation and their time-boxed staging approach. We note two important cases of this. *First*, for metrics 1–3, we identify the subset of annotations from Schaekermann et al. [102] where participants completed all task assignments and thus would be a fair comparison to our case. This included some annotations where there was not enough disagreement to require discussion. *Second*, for the Engagement Quality metric, note that the benchmark's and our analysis is contingent upon deliberation. Therefore, we only use data from items that were deliberated on in both our study and the benchmark to ensure a fair comparison. Although we could leverage other engagement quality metrics (*e.g.*, time) for comparison, these would be unfair to the benchmark's time-boxed staging approach and always show results in our favor (details in Section 5.4). Our metrics reflect what was collected and measured by Schaekermann et al. [102], excluding metrics yielding unfair or unreasonable comparisons.

For qualitative coding, we conducted inductive thematic analysis [14]. The first author open-coded the LLM chat logs, and the third author open-coded the open-text study experience responses. A review of all open codes and axial coding was jointly conducted between the first and second authors. Finally, all authors discussed the axial codes and collaboratively extracted themes.

5 Findings

Our results show the Socratic LLM promoted perspectivist thinking in the annotation process and outperformed the synchronous deliberation benchmark along several key metrics. Comparing to Schaekermann et al. [102]'s work as a benchmark, our main findings are:

(1) Participants' switched annotations significantly more often after discussion with the Socratic LLM for Relation tasks.
(2) Importantly, these changes to the Relation annotations from interacting with the Socratic LLM improved annotation accuracy post-deliberation.
(3) Overall accuracy of our final Relation dataset labels was higher compared to the benchmark.
(4) Confidence in annotations increased after engagement with the Socratic LLM.
(5) The quality of our participants' engagement was higher on average than the benchmark.
(6) Our qualitative analysis of conversations with the Socratic LLM identified different LLM roles that embodied the Socratic processes and facilitated better reasoning.

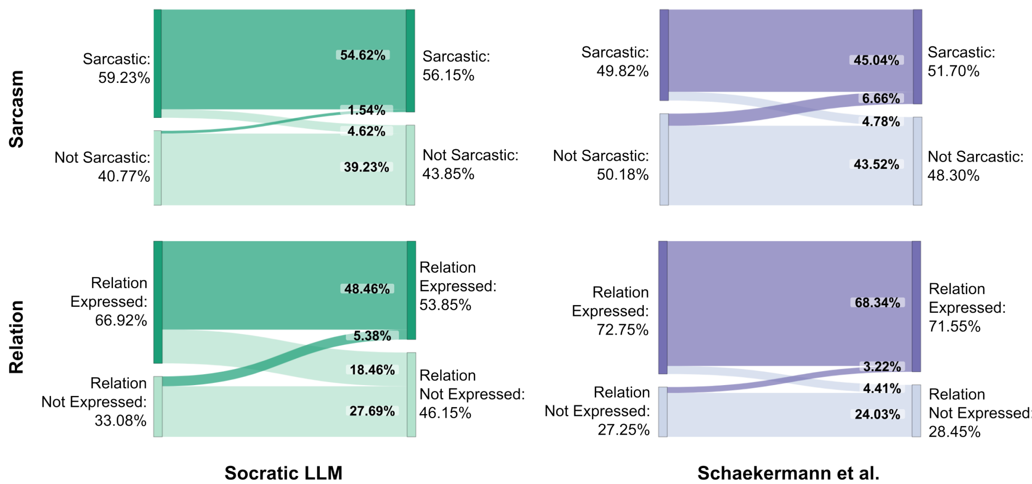

Fig. 3. Annotation-level flips from both our results (left) and Schaekermann et al. [102] (right). Results are split between the Sarcasm (top) and Relation (bottom) datasets.

5.1 Changes in Annotations

We define a "flip" as an annotation that changed pre- *vs.* post-deliberation. When analyzing flips, we focus on the changes to the binary labels (*e.g.*, expressed/not expressed and sarcasm/not sarcasm) and omit post-deliberation labels of "Not Sure" in both the benchmark and our data.[12]

5.1.1 Annotation-Level Flips. Table 1 shows the percentages of annotation-level flips between our Socratic LLM and the benchmark. An annotation-level flip is a flip on an individual's annotation for a given datapoint. For the Sarcasm dataset, our approach resulted in 6.15% of flips on annotations compared to 11.44% flips in the benchmark. However, using a two-population z-test on annotation changes, the rate of annotation flips between us and the benchmark for the Sarcasm dataset were not significant (z = -1.84, p = 0.06576). For the Relation dataset, 23.85% of annotations flipped compared to the benchmark's 7.63%, with this difference being significant (z = 6.34, p « 0.001).

We next consider flips between the pairwise pre- *vs.* post-deliberation annotations in more detail. Figure 3 contextualizes these annotation flips of all pre- *vs.* post-deliberation annotations using Sankey diagrams. These figures illustrate a similar proportion of flips between the binary labels of Sarcasm in our task and the benchmark, as the insignificant findings demonstrate.

We highlight a key difference in outcomes for the Relation task compared to the benchmark. In the Relation task for our study, 18.46% of annotations change from "Expressed" to "Not Expressed" post-deliberation. The benchmark did not see a comparable flip post-deliberation (only 4.41% flipped to "Not Expressed"). We consider whether these flips are consequential later in the Findings Section 5.2.

[12]In our study, there were 45 annotation-level flips. Of these, 6 were changed to "Not Sure" (or 2.25% of the total), evenly split between the two datasets with no overlap on a particular datapoint or participant. In the benchmark, there were 43 changes to "Irresolvable" for the Sarcasm dataset (3.02% of 1424 total annotations) and 61 for the Relation dataset (3.22% of 1896 total annotations).

BY ANNOTATION	Socratic LLM	Schaekermann et al.	Difference
Sarcasm	6.15%	11.44%	-5.29%
Relation	23.85%	7.63%	16.22%***
All Datapoints	15.00%	9.27%	5.73%
BY DATAPOINT			
Sarcasm	5.92%	8.69%	-2.77%
Relation	24.19%	6.80%	17.39% ***
All Datapoints	14.41%	7.81%	6.60%

Table 1. Percent of annotation flips from our results compared to Schaekermann et al. [102] separated by dataset. The top rows describe flips on an annotation level. Lower rows describe average flip rates per datapoint. We mark all significant results with *** to indicate ($p < 0.001$).

5.1.2 Datapoint-Level Flips. Next, we consider datapoint-level flips, *i.e.*, the proportion of annotation flips aggregated by datapoint. To benchmark against Schaekermann et al. [102], we must create a proportional rate of change because the benchmark had many more annotators per datapoint than us. The common denominator in both approaches is the datapoints, which we compare in this section. Recall that each datapoint in our condition was labeled by at least three annotators and in the benchmark by many annotators. Therefore, we calculate the average difference between the datapoints, or the average percentage difference of total annotations that were changed post-deliberation, divided by total annotations for every datapoint. We then average across all proportions for each datapoint using the following formula, where r_i is the flip rate of a particular datapoint in one of the populations and n is the population size:

$$\frac{\sum_{i=1}^{n}(r_{i_{\text{LLM}}} - r_{i_{\text{S}}})}{n}$$

Our datapoint-level findings present a similar narrative to the annotation-level flips—there were significantly more flips on the Relation task. Table 1 also shows the average proportion of annotations that flipped for each datapoint in the Sarcasm and Relation datasets. For the Sarcasm dataset, our approach resulted in 5.92% label flips per datapoint, compared to 8.69% in the benchmark. However, there was no significant difference in the datapoint-level label flips between us and the benchmark, as measured by a paired Mann-Whitney U test ($U = 350$, $p = 0.1095$). For the Relation dataset, datapoint-level label flips were significantly different: our approach resulted in 24.19% flips per datapoint compared to 6.8% for the benchmark ($U = 87$, $p \ll 0.001$).

5.2 Accuracy of Relation Data Annotations

Next, we demonstrate that interaction with our Socratic LLM helped people annotate the data more correctly compared to the baseline. Schaekermann et al. [102] noted that 21 of the Relation datapoints had ground truth associated with them.[13] Therefore, we scrutinize the annotation-level flips of all participants who annotated the 21 Relation datapoints for which ground truth is available (n=71 for us; n=1003 for the benchmark).

[13] 25 of the labels from the dataset have ground truth, but 4 of those were not deliberated upon in Schaekermann et al. [102]. For a fair comparison, we only include the 21 data points for which both we and the benchmark have annotations.

		Socratic LLM		Schaekermann et al.	
Initial Annotation	**Ground Truth**	**Expressed**	**Not Expressed**	**Expressed**	**Not Expressed**
	Expressed	25.35%	11.27%	23.33%	2.39%
	Not Expressed	35.21%	28.17%	50.45%	23.83%
Post-Deliberation	**Ground Truth**	**Expressed**	**Not Expressed**	**Expressed**	**Not Expressed**
	Expressed	23.94%	12.68%	24.13%	1.60%
	Not Expressed	22.54%	40.85%	49.55%	24.73%

Table 2. Confusion matrices for annotations from datapoints provided with a "ground truth" in the Relation dataset. Cells represent the percentage of annotations that fall under each category. Shaded cells denote where the annotation matched the ground truth.

Pre-deliberation, our accuracy ratings on these 21 datapoints is comparable to the benchmark. Table 2 shows the confusion matrices for each annotation stage. In each confusion matrix, the colored diagonal corresponds to the number of participants whose labels matched the ground truth values from the original dataset. Pre-deliberation, our participants had slightly higher accuracy in labeling than the benchmark (53.52% *vs.* 47.16%). We also note that in both the Socratic LLM and the benchmark, most errors were participants generating false positives, where the participant labeled a Relation as being Expressed when the ground truth said Not Expressed. In Schaekermann et al. [102], their participants also had notably fewer false negatives than our participants.

Post-deliberation, however, interacting with the Socratic LLM improved the overall labeling accuracy compared to the benchmark. Our overall accuracy increased substantially to 64.79% (from 52.52%), a considerable improvement over the post-deliberation accuracy change of the benchmark (48.86% post-deliberation from 47.16% initial accuracy). Crucially, engagement with the Socratic LLM substantially *lowered* our false positive rate to 22.54% (from 35.21%), meaning that engagement with the Socratic LLM helped our participants change erroneous "Expressed" relations to the correct "Not Expressed" label. In comparison, the benchmark saw a slight decrease in the false positive rate after deliberation to 49.55% (from 50.45%).

5.3 User Confidence

Confidence in annotations increased after engagement with the Socratic LLM. For this evaluation, we consider changes in confidence pre- *vs.* post-deliberation for the 266 individual annotations for both datasets (two annotations from all 133 participants). We benchmark against confidence evaluations pre- *vs.* post-synchronous deliberation using the extended dataset from Schaekermann et al. [102]. In both setups, participants indicated their confidence level as "very sure", "somewhat sure", or "not sure", representing high, medium, and low confidence, respectively. While high confidence does not indicate that an annotator made a "correct" choice in annotation (assuming one even exists for Sarcasm), we wanted to understand how our system impacted users' perceived confidence after the deliberation process.

Our results show that the Socratic LLM increased confidence in annotations compared to the synchronous deliberation approach of the benchmark. Pre-deliberation, the proportion of high confidence expressed in the benchmark was higher than ours (65.72% versus 57.52%, respectively). However, the Socratic LLM's process increased participants' confidence in their labels: post-deliberation, the percentage of our annotations marked with high confidence increased to 85.34%; in comparison, the benchmark had a minimal increase to 66.39%. The change in confidence (pre-intervention confidence minus post-intervention confidence) was significant with two-sample, unpaired t-tests for both the Sarcasm ($t(1555) = 3.46$, $p \ll 0.001$, Cohen's $d = 0.33$) and Relation ($t(2027) = 3.23$, $p \ll 0.001$, Cohen's $d = 0.36$) datasets with a small effect size.

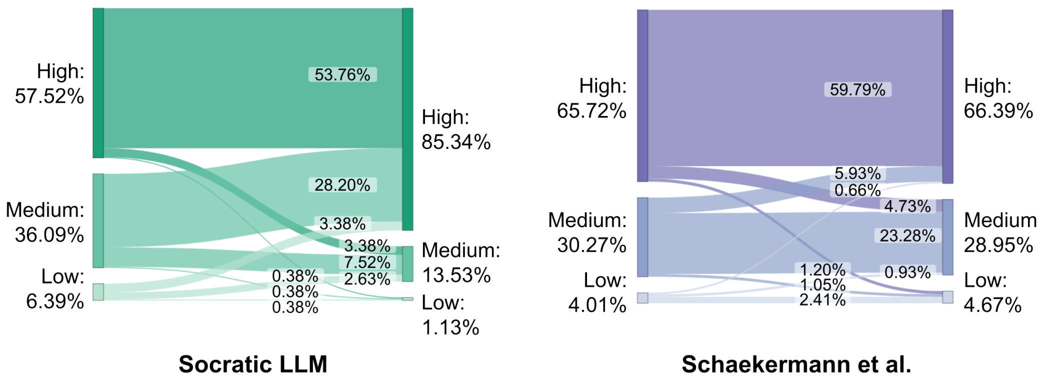

Fig. 4. Changes in confidence between initial annotations and post-deliberation re-annotations, both for us (left) and Schaekermann et al. [102] (right).

We also examined the pathways of confidence changes per annotation to ensure that the Socratic LLM was not overly damaging to individual confidence pre- and post-deliberation. This helps us contextualize how individual annotation confidence changed compared to the population-level averages. Figure 4 presents a Sankey diagram showing these individual pathways for our study and the benchmark. Our annotators' overall confidence increases after interacting with the Socratic LLM, as 28.2% of annotations go from medium to high confidence post-deliberation. For the benchmark, confidence numbers do not change substantially pre- *vs.* post-deliberation.

Qualitatively, participants reported that the Socratic LLM made them feel more confident. When asked to describe their feelings on the process, P33 stated, "Because the chatbot understood the distinction once explained, that seemed to indicate that others would also understand the logic behind my decision. That makes me feel more confident about the label." This is most clearly demonstrated by the annotations that shifted from a "medium" to "high" confidence level, suggesting that participants reasoned about their annotation regardless of whether they kept their original annotation.

5.4 Engagement Quality

Next, we evaluate the differences in engagement between our approach and the benchmark. In many crowdworking tasks [97, 118], time would be the traditional metric to capture objective engagement—faster is considered better because the cost of labeling is a factor of time. However, Schaekermann et al. [102] relied on a synchronous setting with time delays; our design specifically removes this bottleneck. Given the coordination costs of synchronous deliberation, it would be unfair to use time as a measure of engagement since our approach would win by design. Therefore, we rely on a fairer comparison of engagement based on the benchmark's metrics about deliberation quality: the length and number of messages sent by participants. These metrics have been used as a reasonable proxy for evaluating engagement in text-based discussions in prior work (*e.g.*, [77]).

Our Socratic LLM increases discussion engagement quality compared to the benchmark. We measured this by the number and lengths of messages sent to the LLM. The average number of messages sent in our dialogues (omitting the initial default message from the Socratic LLM) was 7.6 (sd = 2.6), compared to the average 5.4 (sd = 1.9) messages exchanged in the benchmark. When it comes to the lengths of these messages, the average number of characters in our participants' messages was 104.7 (sd = 106.8), exceeding the average of 75.3 characters (sd = 57.6) for the benchmark. Our annotators' discussion engagement was high even when the average number of characters contained in messages from our Socratic LLM was 449 (sd = 182.6). This indicates that our LLM contributed much to the dialogues but did not diminish the input of our annotators.

5.5 LLM Roles for Supporting Data Annotation

Through our qualitative analysis of the conversation logs between the participants and our Socratic LLM, we identified that the Socratic LLM's adherence to the prompt instructions (provided in Appendix A) cause distinct patterns of LLM behavior to emerge. We present these patterns as emergent *roles* that the LLM intuitively shifted between to facilitate Socratic dialogue. These roles are:

- *Argument Evaluator*, structuring the provision of claims, evidence, and warrants.
- *Classification Boundary Negotiator*, determining through hypothetical questions the point at which annotators change their annotations.
- *Cognitive Support Tool*, re-articulating information for the convenience of the annotator.
- *Validator*, approving the annotator's label once sufficient deliberation has occurred.

Through this lens, we describe how our Socratic LLM successfully encouraged most participants to justify their reasoning behind their labels.

5.5.1 Socratic LLM as an Argument Evaluator. An overwhelming majority of participants engaged in active reasoning with the Socratic LLM as it requested basic elements of a good argument: a claim (their annotation), evidence (excerpts supporting their claim), and a warrant (an explanation of how the evidence was being used to support their claim) [108]. This was due to the Socratic LLM's initial prompting of participants to provide "important parts of the text" relevant to their annotation, followed by requests for additional evidence. Accordingly, the most frequent type of message from participants contained evidence paired with a warrant. In subjective feedback about our system, participants reported they found this behavior "useful by helping to uncover the 'why'" (P22) and to "articulate very well" (P70) how the text was being used to support their annotation.

During this argument evaluation process, our Socratic LLM helped make participants aware of their perspectives and consider alternate points of view. When requesting a warrant, the Socratic LLM would use language like "I'm curious to hear your perspective" (P36) or, in the expression task, "Can you tell me more about how you're defining a 'relationship' in this context?" (P124). This behavior caused the participants to acknowledge the perspective they brought to the annotation, with one participant commenting in the post-survey how "[the bot] helps lay out points more clearly and concisely without bringing any personal feelings/biases to the table" (P84). In order to encourage perspective shifts, the Socratic LLM would ask for or about counter-evidence and alternative interpretations of evidence. This behavior was crucial for the quality of deliberation, with many participants appreciating how the Socratic LLM prompted them to "second guess" their opinion (P49) and "look at [the data] from different perspectives" (P94). We provide an example interaction of this behavior that led to a change in annotation:

> **P104:** The review appears to be more straight forward *vs.* sarcastic in tone, expressing her unfavorable opinion.
> **Socratic LLM:** I understand your reasoning. Let me ask a follow-up question - do you see any parts of the review that could potentially be interpreted as sarcastic or ironic?
> **P104:** This part seems sarcastic "[excerpt from data]"

Our Socratic LLM's adherence to a neutral perspective was key to this behavior, as it allowed the dialogue to center on the participant's reasoning. One participant pointed out that "the chatbot is great for non-biased analysis. For example, you may not as efficiently convince a friend of a contrasting opinion" (P133), while another stated "[the argumentation] helped me actually talk through it with myself while talking to a chatbot" (P70).

5.5.2 Socratic LLM as a Classification Boundary Negotiator. Another recurring behavior of the Socratic LLM was that it would encourage participants to explore the nuances of the classification boundaries between our binary labels. This behavior is critical for defining decision boundaries when annotating training data; such boundaries ultimately determine the performance of machine learning models and are often the subject of research [19]. When exhibiting this behavior, the Socratic LLM proposed or requested hypothetical evidence that may change the participant's label, specifically at the edge of the boundary. Consider this example from P114's dialogue, where they discuss the title of a review in the Sarcasm task:

> **P114**: "Love at First Sip? Not" I found as sarcastic. Like, playful but also decisive.
> **Socratic LLM:** Okay, I follow your logic there. The title with that phrasing does have a bit of a sarcastic tone to it, even if the review itself is more matter-of-fact.
> Let me ask you this - if the title was just "Didn't Like This At All", do you think that would change your assessment at all? Or is it specifically the "Love at First Sip? Not" phrasing that makes the title seem sarcastic to you?
> **P114:** Yes, specifically the play on words with "Love at First Sip?" If the title was "Didn't Like This At All" I wouldn't even think it was being sarcastic.

This negotiation also elicits counterfactual examples from annotators, providing valuable insights for dataset curators.

5.5.3 Socratic LLM as a Cognitive Support Tool. Our Socratic LLM affirmed a shared understanding throughout the Socratic process and, in doing so, allowed cognitive support and offloading. Almost every message sent by the Socratic LLM would reiterate what the participant had just said and summarize the participant's reasoning up to that point, creating an opportunity for the participant to interject if there was a misunderstanding. Only once did we find an instance where the participant indicated a misunderstanding, reflecting the LLM's capacity for summarization. The Socratic LLM reduced the cognitive load on the message-to-message scale by fully articulating the participants' points. Consider this example from P132:

> **P132:** "brought up" is almost equivalent to "lived in". Here, it's the same thing.
> **Socratic LLM:** Okay, I understand your reasoning now. You're interpreting the phrase "brought up by a star-struck mother in Los Angeles" as essentially meaning that Bachardy lived in Los Angeles during that time period.
> That's a reasonable interpretation based on the language used in the text. The phrase "brought up" does strongly imply a more long-term living situation, rather than just a temporary visit or stay.

This tendency to re-articulate created a beneficial asymmetry in the deliberation process. Participants were free to dictate the broad strokes of the dialogue and ensure that the Socratic LLM was on the same page by relying on its extensive re-articulations. This is supported by the difference in length between participant and LLM messages: as pointed out in Section 5.4, messages from our Socratic LLM were 4 times longer on average. This behavior resulted in natural-feeling dialogue for our participants, expressed in compliments directed toward our Socratic LLM and how it "listened pretty well" (P5), was "sharper than expected" (P106), and "actually felt like I was chatting with a human" (P95).

5.5.4 Socratic LLM as a Validator. Finally, understanding the participant's reasoning was a prerequisite for the Socratic LLM to approve their annotation—we associate this approval behavior with the LLM's role as a validator. Participants were surprisingly committed to seeking this validation, even though it was not required. While they could have moved on from the deliberation phase after sending only two messages to the Socratic LLM, our average dialogue lengths far exceeded this minimum (see Section 5.4 for details). In fact, the most common type of message sent by the Socratic LLM was one in which the participant's annotation was approved. The frequency of hypothesis approval messages indicated a corresponding abundance of active reasoning, where participants went beyond the bare minimum and instead engaged in a full Socratic dialogue with the LLM.

The hypothesis approval messages outlined the justification for the approvals, highlighting the reasoning performed by the participant. Take this example directed toward P123:

> **Socratic LLM:** Great, I think I have a good understanding of your reasoning now. The sarcastic title, the exaggerated language about throwing the product away despite being "really tight with the dollar", and the overall negative tone throughout the review led you to confidently label this as a sarcastic review.
> Since your logic and explanations make sense based on the information provided, I'd encourage you to go ahead and re-annotate the item with your "Sarcastic" label and high confidence.

Such summaries of the participants' reasoning, combined with an upbeat and supportive tone, left many participants feeling "validated" (P109) after the deliberation process.

5.6 Experiential Factors of Socratic LLM Usage

Finally, we consider participants' experiences with the Socratic LLM through their post-study questionnaire (see Section 4.2.4 for details on questions asked). In short, our participants were accepting of the Socratic LLM as a deliberation partner and did not find it burdensome to use. Their open-ended responses provide additional nuance on how future iterations of the system can fulfill the diverse needs of data annotators.

5.6.1 Perceived Helpfulness of the Socratic LLM. Overall, participants found the Socratic LLM to be helpful. Recall that we measured anticipated helpfulness of discussion before—and retrospective helpfulness after—the discussion phase for each annotation. Participants anticipated that discussion would be helpful in slightly less than half of annotations (48%). After the discussions, we saw a substantial increase where participants considered 68% of discussions to be helpful retroactively with only 32% (85 annotations) holding a negative opinion. The perceived value of discussions improved in almost one-third of annotations after a discussion with our Socratic LLM.

The sentiment around the LLM discussions was also overwhelmingly positive, regardless of prior deliberation experience. While only 22 of our participants (16.54%) had prior experience with human-human deliberation, all but one of them agreed that deliberation was at least somewhat important to the annotation process. Likewise, participants who were new to deliberations also viewed the experience positively—of the 111 (83.46%) people who had not done human-human deliberations before, 102 said that deliberation was either somewhat or very important to the annotation process, with 99 of them also expressing a willingness to use a similar system in the future.

In short, the sentiment around the asynchronous LLM-based discussions was overwhelmingly positive, regardless of prior deliberation experience.

5.6.2 Effort and Cognitive Load. Recorded NASA-TLX scores show that participants felt engaged by the process rather than overwhelmed. Recall that all TLX items were answered on a scale of 1–21. On average, the task was seen as somewhat mentally demanding (mean = 8.8, sd = 5.3) and effortful (mean = 9.7, sd = 5.2), but low in temporal demand (mean = 4.4, sd = 3.7), impedance to personal performance (mean = 3.4, std = 3.6), and frustration (mean = 3.4, sd = 3.4).

We also examined qualitative data for participants who reported high effort in their NASA-TLX evaluations. In these cases, participants reported that they were actively contemplating the decisions they had made. Despite having a "high" effort score (19 out of 21), P47 reflected "I liked getting direct feedback that makes me more active in what I'm participating in. It helps me think more critically." Similar sentiments about alternate perspectives and evaluating arguments were common from participants with high effort scores.

The relatively low NASA-TLX scores for frustration, time, and task effort likely reflect our small-scale task design. Having participants annotate just two datapoints allowed for each discussion to develop naturally. The Socratic dialogue encouraged continuous engagement with each annotation, resulting in higher mental effort but with limited frustration or time pressure typically associated with more expansive (and therefore time-consuming) annotation tasks.

5.7 Failures and Design Opportunities

5.7.1 Socratic LLM Conversation Logs. Our approach was not without failures: the Socratic LLM occasionally failed to actualize the Socratic method. Despite being instructed not to, there were a few instances where the LLM would incorporate outside information for the Relation task or infer a warrant when the participant only provided evidence. While these infractions were largely inconsequential, the Socratic LLM did commit more egregious mistakes, albeit rarely.

The Socratic LLM's most harmful behavior was its occasional misrepresentation of the task to the participant. This issue occurred most often for the Relation task, as the Socratic LLM enforced the misconception that the relationships must be explicitly stated in the data rather than implied. Unfortunately, this led to some annotation flips, and one participant even mistakenly complimented the Socratic LLM as it made them "understand that we are determining whether the relationship is directly expressed vs just implied" (P35). Fortunately, in most cases, the opposite occurred where the Socratic LLM would correct the participants' misunderstanding of the task.

The Socratic LLM would also occasionally push an idea with too much or too little intensity. In one extreme case, the LLM outright told the participant "I think the more accurate label would be 'Relationship is Not Expressed'" (P66). Instances of such behavior led one participant to state, "I thought the chatbot was very insistent and felt like it would never consider my inference as valid" (P32). On the other end of the spectrum, the Socratic LLM could be too mild, approving poor reasoning and causing some participants to state that it felt like "an echo chamber" (P19) where deliberation was a "waste of time" as one could "steam roll the AI" (P23).

Such sentiments about the Socratic LLM's failure as a deliberation partner were rarely expressed, with most participants sharing a positive opinion on its efficacy. Nonetheless, these issues demand attention in future iterations of Socratic LLMs, whether via refined system prompts, more robust guardrails, or more capable underlying LLMs.

5.7.2 User Experience Feedback. A small portion of participants did not feel that the Socratic LLM contributed much to the discussion. One participant wrote, "I felt like I was teaching it. I do not want to have to teach something to do a job of another person" (P35). A handful of others stated things like they "see no reason for it" (P16) or that they would "not find it personally helpful" (P122).

Interestingly, some of the more tepid responses came from people who were at least somewhat aware of the limitations regarding LLMs in practice. P25 wrote "I would consider them only if I could be sure that the chatbot was correct in what they were saying. I feel too unsure that a chatbot would help me reason out the correct answer without leading me astray," suggesting an understanding that LLMs have the potential for hallucinations. P10 added to this sentiment, saying "Since AI uses information that's already out there, it can be inherently flawed. They're using human data. We are flawed", implicating the underlying bias present in any model's training data. These concerns reflect back to the design of the initial prompt for the Socratic LLM.

Although there were some criticisms of the Socratic LLM, these concerns still represented a small minority of the feedback we received. It was rare for the LLM to misrepresent the task, and even rarer for it to hold fast to an opinion. In general, we saw many of our participants enjoy their experience and think deeply about the choices they were making. P58 captured our goals almost verbatim in their parting words to us:

> It's one of the only use cases of AI that I'm completely comfortable with. Having an entity that you can direct your thoughts around is almost like a slightly improved version of journaling. The AI isn't solving anything, but rather acting as a way to challenge concepts and explore other avenues.

6 Discussion

Our work highlights the benefits of using a Socratic LLM in supporting a perspectivist approach to data annotation. Our findings clearly signal improvement on key metrics: change in annotation pre- *vs.* post-deliberation, improved Relation task accuracy, improved confidence, and higher engagement. Moreover, our qualitative results highlight the varied roles that an LLM can take on, even with a simple Socratic design, including an argumentation device, cognitive support tool, and validator. Importantly, participants found value in the discussion process, which is reflected in their increased confidence and perceptions of usefulness. These findings collectively demonstrate the value of incorporating deliberation in data domains that can benefit from perspectivism. Moreover, our approach presents a successful alternative to work that relies on synchronous deliberation (*e.g.*, [19, 22, 102]), by resolving coordination and logistical costs.

6.1 Connection to HCI Theories

We are excited by opportunities to connect our work with well-regarded HCI theory in both describing the practices of our participants and providing new inferences for AI systems grounded in historical HCI knowledge.

Distributed Cognition. Our findings suggest that our Socratic LLM was helping our participants with distributed cognition processes [50, 53, 93]. Data annotation can be cognitively demanding, especially when the dataset being labeled is complex or requires consideration of multiple perspectives, such as toxicity, hate speech, or mental health labels. With our computationally-mediated approach, the Socratic LLM served as an artifact that "stored" a participant's thinking. Our analysis showed that, at each dialogic point, the Socratic LLM reminded people of their thought processes so far by summarizing then asking for follow-up. We believe this helped participants offload some of the cognitive tasks that people conduct when considering classification boundaries; instead, participants could engage more thoughtfully in reasoning about the questions at hand. Our Socratic LLM helps make these complex annotation tasks less cognitively demanding. We regard this as a classic case of "How a Cockpit Remembers its Speeds" [53].

Conversations as Boundary Objects. Similarly, we find that the conversations with the Socratic LLM can serve as a boundary object in annotation tasks [72, 105]. Boundary objects are "both plastic enough to adapt to local needs…yet robust enough to maintain a common identity across sites. They are weakly structured in common use, and become strongly structured in individual-site use" [105][p.393]. Our proposed boundary object—conversations generated via Socratic exchange—have a common identity across stakeholders but can be used in unique ways by different stakeholders. Annotators can use conversations to understand their own labels, negotiate the boundaries of various classes in an annotation task, and re-consider their labels if need be. Our participants used conversations in all of these aforementioned ways (see Section 5.5). Researchers and data curators can also use conversations to interrogate and further study important boundaries of tasks, including uncertainty and tensions in interpretation. Moreover, we hypothesize that the conversational artifacts could be useful for adding contextual insights to machine learning architectures, an area of emergent interest [42, 115, 120] for a different set of stakeholders.

6.2 Mechanisms for Supporting Perspectivist Data Annotation

Historically, data annotation was rooted in assumptions of accuracy and alignment to a single, objective "ground truth". Following this, prior work in the space conceptualized disagreement as "noise" and error to resolve [45, 54, 122]. This led to a cascade of negative outcomes such as non-representative datasets [16, 99], biased models [20, 35], diminished representation [42, 112], and missing critical "edge cases" [84].

On the other hand, a perspectivist approach to annotation recognizes these "negative" outcomes as unexplored opportunities to think about the task. Our approach in this paper embodies perspectivism at three levels: 1) the task itself; 2) the system we designed; and 3) the metrics we prioritize.

Task Design. The addition of a "Not Sure" label option during the re-annotation phase reflects a perspectivist task goal by giving space for uncertainty in labels. This additional choice was not provided in the first annotation phase to require the user to formulate a hypothesis for one binary label over the other. Only after the deliberation—during which this initial perspective and others are considered—were participants able to select "Not Sure". Such an option holds much value for perspectivist datasets, as it pinpoints annotations where ambiguity is high enough for dataset curators to address directly [76].

System Design. Of the many possibilities when setting up a discussion-based agent, we selected the Socratic method because it mirrors a perspectivist approach. A focus on other kinds of deliberation could have resulted in, for example, a dialectic or debate-centered discussion (*e.g.*, [113]) where the Socratic LLM presented an alternate view, or a panel-style discussion (*e.g.*, [116]) where the LLM asked specific questions about the participants' selected label. Our perspectivist focus led to design choices that required participants to be the ones initiating the cognitive engagement with a different perspective. That is, participants defined alternate hypotheses/perspectives rather than engaging with pre-selected alternatives from us. This is a notable strength of the Socratic method [2, 5, 12, 31, 38, 52].

Metrics. Taking this perspectivist approach to data annotation helps us re-focus on relevant metrics beyond accuracy. For us, a critical metric that reflects this is people's confidence in their labels. Our method's positive quantitative impact on confidence and our qualitative findings suggest that participants' annotations became more aligned with their underlying perspectives. Regardless of the source—whether it stemmed from a change in annotation, the verbalization of the participant's reasoning, or the development of the participant's understanding of the data being annotated—this increased confidence supports the goals of curating quality perspectivist datasets.

6.3 Benefits of Asymmetric Deliberation

Although synchronous deliberation has thus far been considered best practice for resolving disagreement on annotation (and other) tasks, in practice, it can be impeded by social dynamics [51, 87]. Individuals can sometimes commandeer conversations, or social pressures can cause some people to hesitate in expressing ideas [62, 86]. The field of groupwork has long studied the issues that arise when getting multiple people to collaborate on a shared objective [19, 69, 70, 102].

In contrast, the asymmetry of human-AI pairings results in more effective deliberation. People feel comfortable expressing a certain level of curtness with LLMs, which can be effective for evaluating multiple perspectives. This is evident in the short messages sent by participants, which were expanded in the Socratic LLM's verbose summative responses. Participants could then confirm or rethink elements of the Socratic LLM's summary. The social pressures of human-human interaction do not allow for similar bluntness, asymmetry in communication, or impatience.

That said, there are benefits to human-human deliberation that are lost with AI partners. AI, including LLMs, is biased toward perspectives that represent the majority because of training data, so a deliberator may not be exposed to underrepresented perspectives that they could find among human deliberation partners. Furthermore, the neutrality of our Socratic approach does not fully emulate contestation, which is characteristic of a healthy community engaged in problem-solving [71, 83, 88]. Respecting this, we argue that the use of AI-assisted deliberation should be balanced with synchronous deliberation in perspectivist annotation pipelines, accounting for the logistic context (*e.g.*, coordination costs) and the case-by-case value of contestation.

6.4 Design Implications for Positive LLM Use-Cases

In addition to the theoretical aspects of our findings, we examine practical design implications.

Generating Counterfactuals. From our results emerged a positive, unintended use-case for a Socratic LLM: to help people reason about class boundaries in a classification task (Section 5.5). We believe this could result in an exciting use-case of generating counterfactuals, *i.e.*, hypothetical examples testing how slight changes to a datapoint can affect annotations. Future work could update our Socratic LLM system prompt to engage in example-based dialogue as a means of generating these counterfactuals. The resulting pair of datapoints—the original and the counterfactual—could clarify the decision boundary more precisely. Finding ways to generate counterfactuals is central to fields like explainable AI, to improve the transparency of AI systems [85, 87].

Balancing Guardrails based on Perspectivist Needs. Future iterations of Socratic LLMs should test the guardrails we applied to our system, to balance the potential for LLM errors against the external knowledge needed for the LLM to be a helpful partner. Our system prompt heavily constrained the LLM to avoid using external knowledge and to use generic conversation prompts based on high-level Socratic principles. Future implementations can carefully adjust the information and instructions available to the Socratic LLM to reflect the needs of different annotation tasks, emphasizing behaviors that yield the richest perspectivist data.

Additional Personas to Support Deliberation. Our findings highlight the different roles that LLMs can take on (even when prompted to only be a Socratic conversation aid) and how these shape participant behaviors. The Socratic LLM fulfilled the roles of an argumentation device, boundary negotiator, cognitive support tool, and validator via its adoption of a Socratic persona. Looking ahead, there are other personas that may fulfill different, equally valuable roles depending on task goals for deliberation and annotation. For example, designers could apply De Bono's Six Thinking Hats [124] to provide LLM personas that facilitate different styles critical thinking, such as adversarial, creative, or emotional. There are also opportunities to model personas on those who will be impacted by the product of deliberation. Recent work in the writing domain has explored modeling personas on anticipated audience members [10]—applying this idea to the data annotation realm, we might create personas representing communities that will be impacted by models trained on the data being annotated.

6.5 Ethical and Pragmatic Concerns

While the Socratic LLM was largely successful in our study, there are notable risks of adoption in future applications. Here, we describe ethical and pragmatic concerns when introducing Socratic LLMs into data annotation contexts, and provide suggestions for addressing those concerns.

Bias Amplification and Over-reliance. Socratic LLMs must strike a careful balance in rigor and forcefulness: while some degree of contention is required to draw out the reasoning of annotators, an overly contentious LLM runs the risk of overriding annotator perspectives. We saw this happen when observing failures of the Socratic LLM (Section 5.7.1). The social affability and strategy chosen by the LLM could lead to different annotation outcomes, potentially amplifying the biases in data annotation tasks.

Moreover, even when the LLM strikes an appropriate balance in dialogue about annotations, there is a risk that LLMs afford too much cognitive support and thus precipitate over-reliance and rationalization of the LLM's responses. Ongoing work on AI-assisted tasks has highlighted the need for cognitive forcing functions when using AI assistance for tasks (*e.g.*, [15, 109]), though designing appropriate functions has proved challenging. More research is needed to unpack the impact of LLM assistance on people's cognition for the sake of avoiding over-reliance.

Equity Issues and Legitimate Crowdwork Crowdwork and large-scale annotation are the backbone of data labeling for many AI systems. Pragmatically, most of our participants felt like the Socratic LLM was helpful, and we imagine that large-scale deployment of LLMs like ours could support legitimate workers in moving quickly through tasks where they have high confidence. On the other hand, there are ongoing challenges in assessing the quality of crowdworker labor. The effort required to engage with the Socratic LLM in combination with financial pressures can exacerbate this issue when annotators attempt to "game the system". Indeed, we encountered participants who attempted to complete our study using an external LLM for the portions that required writing. Research on crowdsourced data annotation is aware of these behaviors and tries to work around them [4, 92]—guidelines we also followed (*e.g.*, by using attention and LLM checks).

Bias in Annotator Selection and Divergent Perspectives. LLMs like ours also have the potential to undermine unique and legitimate perspectives. Here, we consider a scenario where annotators are only routed to a Socratic LLM when there is a risk of divergence in an item's labels. Pragmatically, annotators with unique perspectives—those whom we want to center in perspectivist approaches to annotation—will be more likely to engage with a Socratic LLM as their annotations differ from dominant viewpoints, meaning that their rate of task completion could be slowed. This could be prevented by avoiding any conditional use of the LLM, with all annotators for a task being given the same probability of encountering the Socratic LLM. Alternatively, annotators could be appropriately compensated for their time of engagement with Socratic LLMs. These suggestions will ensure that annotators are not financially punished for having minority perspectives.

However, theoretical concerns remain for annotators with diverse viewpoints and the tie-in to bias amplification. The LLM's biases may push against legitimately held but non-standard views about a subject, simply because the LLM's underlying model does not factor in these novel experiences from its training data [9]. This means that the Socratic LLM, if employed without care, may in fact harm the population it intends to empower by paying them less than their counterparts who possess more dominant perspectives or by trying to undermine their divergent perspectives.

Misuse of Deliberation Logs. We must consider the potential misuses of deliberation logs across the collection, analysis, and distribution of such data. First, the data collection process poses a threat to annotator privacy: over the course of perspectivist deliberation, annotators may rely on personally identifiable information. This is at odds with the anonymity typically offered on crowdworking platforms, and could be particularly consequential depending on the deliberation topic and data collection agency. As such, we suggest that future implementations of the Socratic LLM explicitly include instructions against asking questions that yield personally-identifiable information. Secondly, we recommend extreme caution when filtering deliberation logs during analysis. Such filtration, even if conducted without ill intent, runs the risk of silencing groups of annotators and thus contradicts the goals of perspectivism. Finally, data release agreements for deliberation logs require special care. Depending on the topic of deliberation, logs containing rich argumentative data may be appropriated for socially impactful efforts like mass persuasion campaigns [30, 110]. Considering the potential influence of such data, it should be made transparent to annotators how their arguments will be used and curators should maintain strict control over who has access to the deliberation logs.

7 Limitations and Future Work

This work is primarily limited by the difference in design between the synchronous deliberation approach of the benchmark paper and our asynchronous implementation. Schaekermann et al. [102] sent a subset of their initial annotations through the deliberation process based on inter-rated disagreement calculations. Even if we only consider annotations from the participants who completed the whole study, there are still some datapoints that were never discussed. Although we have only compared metrics where direct comparison points were feasible, we had to proportion our data in ways that made these comparisons meaningful (by datapoint and deliberated annotations). A between-subjects experiment where we re-imagined the benchmark's approach might not have had the same limitations, but we believe benchmarking is a valuable aspect of systems work and wanted to highlight that in our methodology.

Another limitation of our work mimics the differences between human-human interaction and human-AI interaction. To some degree, the participants in our study did not have to compromise as much as they would when deliberating with other humans—at some point they could always convince the LLM of their opinion or ignore it entirely. People could also be curt about the whole process, and our current attempt at a Socratic LLM did not challenge this behavior. These preconceived notions of AI and distinct AI-facing behaviors might influence metrics like confidence and engagement in our study. We call for future work to disentangle the influence of AI further.

We also note the inherent limitation of qualitative analysis and the behaviors associated with the roles in Section 5.5. These roles were largely effective and appreciated by participants, despite not being explicitly designed in the prompt instructions. However, our qualitative analysis is not sufficient to make claims about the prevalence of these roles in Socratic LLM systems. We are eager for future work to take advantage of this role-adopting inclination to explore Socratic LLMs further tailored to the annotation context, explicitly switching between custom roles depending on the state of the dialogue.

Finally, we must acknowledge that while our sample of participants was large, it was constrained to the United States. This limits the generalizability of our findings: norms around conflict resolution and LLM usage vary between cultures, meaning that the discussion behaviors exhibited by annotators and the overall efficacy of Socratic LLMs may differ if our study involved participants from outside the United States. We encourage future work to explore how our findings extend to samples drawn globally.

8 Conclusion

In an era where data annotation is the backbone of many widely-used AI systems, reflecting diverse perspectives in the data labels is increasingly critical. Prior work has supported this perspectivist approach to annotation by using synchronous deliberation. We built a Socratic LLM to support asynchronous deliberation for perspectivist data annotation, to reap the benefits of deliberation while resolving the bottlenecks of synchronous deliberation (*e.g.*, time, resources, coordination costs). Our results show that we can use LLMs to assist in quality multi-perspectivist dataset creation in a scalable way. Moreover, we see this as a positive use-case of LLMs in a landscape where their use is wrought with regulation and bias concerns. Our Socratic LLM bridges the gap between the convenience of crowdsourced work and the need for high-quality, representative datasets.

Acknowledgments

We would like to thank our reviewers for their helpful comments. We are grateful to Mo Houtti, Loren Terveen, and all of the faculty and students in the GroupLens research lab for their feedback and support. We also want to thank those who participated in our study.

References

[1] [n. d.]. Socratic Methods - Wikiversity — en.wikiversity.org. https://en.wikiversity.org/wiki/Socratic_Methods. [Accessed 2024-10-25].

[2] Jamie R. Abrams. 2015. Reframing the Socratic Method. 64, 4 (2015), 562–585. jstor:24716713 https://www.jstor.org/stable/24716713

[3] Erfan Al-Hossami, Razvan Bunescu, Justin Smith, and Ryan Teehan. 2024. Can Language Models Employ the Socratic Method? Experiments with Code Debugging. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1 (SIGCSE 2024)*. Association for Computing Machinery, New York, NY, USA, 53–59. doi:10.1145/3626252.3630799

[4] Reham Al Tamime, Joni Salminen, Soon-Gyo Jung, and Bernard Jansen. 2024. Evaluating LLM-Generated Topics from Survey Responses: Identifying Challenges in Recruiting Participants through Crowdsourcing. In *2024 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 412–416.

[5] Maya Aloni and Christine Harrington. 2018. Research Based Practices for Improving the Effectiveness of Asynchronous Online Discussion Boards. *Scholarship of Teaching and Learning in Psychology* 4, 4 (Dec. 2018), 271–289. doi:10.1037/stl0000121

[6] Omar Alonso and Stefano Mizzaro. 2012. Using crowdsourcing for TREC relevance assessment. *Information processing & management* 48, 6 (2012), 1053–1066.

[7] Paul André, Aniket Kittur, and Steven P Dow. 2014. Crowd synthesis: Extracting categories and clusters from complex data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 989–998.

[8] Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine* 36, 1 (March 2015), 15–24. doi:10.1609/aimag.v36i1.2564

[9] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.

[10] Karim Benharrak, Tim Zindulka, Florian Lehmann, Hendrik Heuer, and Daniel Buschek. 2024. Writer-Defined AI Personas for On-Demand Feedback Generation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1049, 18 pages. doi:10.1145/3613904.3642406

[11] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. In *Social Informatics* (Cham, 2017), Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri (Eds.). Springer International Publishing, 405–415. doi:10.1007/978-3-319-67256-4_32

[12] Eugenia Arazo Boa, Amornrat Wattanatorn, and Kanchit Tagong. 2018. The Development and Validation of the Blended Socratic Method of Teaching (BSMT): An Instructional Model to Enhance Critical Thinking Skills of Undergraduate Business Students. 39, 1 (2018), 81–89. doi:10.1016/j.kjss.2018.01.001

[13] Jonathan Bragg, Mausam, and Daniel S. Weld. 2018. Sprout: Crowd-Powered Task Design for Crowdsourcing. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*. Association for Computing Machinery, New York, NY, USA, 165–176. doi:10.1145/3242587.3242598

[14] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[15] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. 5 (2021), 188:1–188:21. Issue CSCW1. doi:10.1145/3449287

[16] Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 6 (June 2023), 6860–6868. doi:10.1609/aaai.v37i6.25840

[17] Carrie J Cai, Shamsi T Iqbal, and Jaime Teevan. 2016. Chain reactions: The impact of order on microtask chains. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3143–3154.

[18] Scott Allen Cambo and Darren Gergle. 2022. Model Positionality and Computational Reflexivity: Promoting Reflexivity in Data Science. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19. doi:10.1145/3491102.3501998

[19] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2334–2346. doi:10.1145/3025453.3026044

[20] Adriane Chapman, Philip Grylls, Pamela Ugwudike, David Gammack, and Jacqui Ayling. 2022. A Data-driven Analysis of the Interplay between Criminological Theory and Predictive Policing Algorithms. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2022-06-20) *(FAccT '22)*. Association for Computing Machinery, 36–45. doi:10.1145/3531146.3533071

[21] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758.

[22] Quanze Chen, Jonathan Bragg, Lydia B Chilton, and Dan S Weld. 2019. Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.

[23] Quan Ze Chen, Daniel S. Weld, and Amy X. Zhang. 2021. Goldilocks: Consistent Crowdsourced Scalar Annotations with Relative Uncertainty. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 335:1–335:25. doi:10.1145/3476076

[24] Quan Ze Chen and Amy X. Zhang. 2023. Judgment Sieve: Reducing Uncertainty in Group Judgments through Interventions Targeting Ambiguity versus Disagreement. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Oct. 2023), 283:1–283:26. doi:10.1145/3610074

[25] Julian Chingoma and Adrian Haret. 2023. Deliberation as Evidence Disclosure: A Tale of Two Protocol Types. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '23)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2866–2868.

[26] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. 51, 1 (2018), 7:1–7:40. doi:10.1145/3148148

[27] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics* 10 (Jan. 2022), 92–110. doi:10.1162/tacl_a_00449

[28] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, Vol. 11. 512–515.

[29] Valerio De Stefano. 2015. *The Rise of the 'Just-in-Time Workforce': On-Demand Work, Crowd Work and Labour Protection in the 'Gig-Economy'*. Social Science Research Network:2682602 doi:10.2139/ssrn.2682602

[30] Amanda Delaney, Bella Lough, Michelle Whelan, Max Cameron, et al. 2004. A review of mass media campaigns in road safety. *Monash University Accident Research Centre Reports* 220 (2004), 85.

[31] Haris Delić and Senad Bećirović. 2016. Socratic Method as an Approach to Teaching. *European Researcher* 111, 10 (Oct. 2016). doi:10.13187/er.2016.111.511

[32] Yuyang Ding, Hanglei Hu, Jie Zhou, Qin Chen, Bo Jiang, and Liang He. 2024. Boosting Large Language Models with Socratic Method for Conversational Mathematics Teaching. arXiv:2407.17349 [cs]

[33] Carl DiSalvo. 2012. *Adversarial Design*. The MIT Press. doi:10.7551/mitpress/8732.001.0001

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

[35] Xavier Ferrer, prefix=van useprefix=false family=Nuenen, given=Tom, Jose M. Such, Mark Coté, and Natalia Criado. 2021. Bias and Discrimination in AI: A Cross-Disciplinary Perspective. 40, 2 (2021), 72–80. doi:10.1109/MTS.2021.3056293

[36] Elena Filatova. 2012. Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing.. In *Lrec*. Citeseer, 392–398.

[37] Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The Perspectivist Paradigm Shift: Assumptions and Challenges of Capturing Human Labels. (May 2024). doi:10.48550/ARXIV.2405.05860

[38] Soon Yen Foo and Choon Lang Quek. 2019. Developing Students' Critical Thinking through Asynchronous Online Discussions: A Literature Review. *Malaysian Online Journal of Educational Technology* 7, 2 (2019), 37–58. doi:10.17220/mojet.2019.02.003

[39] Giles Foody, Linda See, Steffen Fritz, Inian Moorthy, Christoph Perger, Christian Schill, and Doreen Boyd. 2018. Increasing the accuracy of crowdsourced information on land cover via a voting procedure weighted by information inferred from the contributed data. *ISPRS International Journal of Geo-Information* 7, 3 (2018), 80.

[40] Benoît Frénay and Michel Verleysen. 2014. Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems* 25, 5 (May 2014), 845–869. doi:10.1109/TNNLS.2013.2292894

[41] Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation* (2024), 1–28.
[42] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19. doi:10.1145/3491102.3502004
[43] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. doi:10.1145/3411764.3445423
[44] Tanya Goyal, Tyler McDonnell, Mucahid Kutlu, Tamer Elsayed, and Matthew Lease. 2018. Your Behavior Signals Your Reliability: Modeling Crowd Behavioral Traces to Ensure Quality Relevance Annotations. 6 (2018), 41–49. doi:10.1609/hcomp.v6i1.13331
[45] Hui Guo, Boyu Wang, and Grace Yi. 2023. Label Correction of Crowdsourced Noisy Annotations with an Instance-Dependent Noise Transition Model. 36 (2023), 347–386. https://proceedings.neurips.cc/paper_files/paper/2023/hash/015a8c69bedcb0a7b2ed2e1678f34399-Abstract-Conference.html
[46] Margeret Hall, Mohammad Farhad Afzali, Markus Krause, and Simon Caton. 2022. What Quality Control Mechanisms Do We Need for High-Quality Crowd Work? 10 (2022), 99709–99723. doi:10.1109/ACCESS.2022.3207292
[47] Jawad Haqbeen, Takayuki Ito, Rafik Hadfi, Tomohiro Nishida, Zoia Sahab, Sofia Sahab, Shafiq Roghmal, and Mohammad Amiryar. 2020. Promoting Discussion with AI-based Facilitation: Urban Dialogue with Kabul City.
[48] Sandra G Hart. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human mental workload/Elsevier* (1988).
[49] Yueh-Ren Ho, Bao-Yu Chen, and Chien-Ming Li. 2023. Thinking More Wisely: Using the Socratic Method to Develop Critical Thinking Skills amongst Healthcare Students. 23, 1 (2023), 173. doi:10.1186/s12909-023-04134-2
[50] James Hollan, Edwin Hutchins, and David Kirsh. 2000. Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction (TOCHI)* 7, 2 (2000), 174–196.
[51] Mo Houtti, Moyan Zhou, Loren Terveen, and Stevie Chancellor. 2023. " All of the White People Went First": How Video Conferencing Consolidates Control and Exacerbates Workplace Bias. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–25.
[52] Jui-Tse Hung, Christopher Cui, Diana M. Popescu, Saurabh Chatterjee, and Thad Starner. 2024. Socratic Mind: Scalable Oral Assessment Powered By AI. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24)*. Association for Computing Machinery, New York, NY, USA, 340–345. doi:10.1145/3657604.3664661
[53] Edwin Hutchins. 1995. How a cockpit remembers its speeds. *Cognitive science* 19, 3 (1995), 265–288.
[54] Shahana Ibrahim, Panagiotis A. Traganitis, Xiao Fu, and Georgios B. Giannakis. 2024. *Learning From Crowdsourced Noisy Labels: A Signal Processing Perspective.* arXiv:2407.06902 doi:10.48550/arXiv.2407.06902
[55] Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the third workshop on abusive language online*. 46–57.
[56] Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. 2014. Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In *The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II 13*. Springer, 486–504.
[57] Panagiotis G Ipeirotis and Evgeniy Gabrilovich. 2014. Quizz: targeted crowdsourcing with a billion (potential) users. In *Proceedings of the 23rd international conference on World wide web*. 143–154.
[58] Shamsi T Iqbal and Brian P Bailey. 2005. Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption. In *CHI'05 extended abstracts on Human factors in computing systems*. 1489–1492.
[59] Deniz Iren and Semih Bilgen. 2014. Cost of Quality in Crowdsourcing. 1, 2 (2014). Issue 2. doi:10.15346/hc.v1i2.14
[60] V. K. Chaithanya Manam, Dwarakanath Jampani, Mariam Zaim, Meng-Han Wu, and Alexander J. Quinn. 2019. TaskMate: A Mechanism to Improve the Quality of Instructions in Crowdsourcing. In *Companion Proceedings of The 2019 World Wide Web Conference (WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1121–1130. doi:10.1145/3308560.3317081
[61] Shivani Kapania, Alex S Taylor, and Ding Wang. 2023. A hunt for the snark: Annotator diversity in data practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.
[62] Martin F Kaplan and Ana M Martin. 1999. Effects of differential status of group members on process and outcome of deliberation. *Group Processes & Intergroup Relations* 2, 4 (1999), 347–364.
[63] Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2023. DeliData: A Dataset for Deliberation in Multi-party Problem Solving. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2 (Oct. 2023), 265:1–265:25. doi:10.1145/3610056
[64] Priyanka Kargupta, Ishika Agarwal, Dilek Hakkani-Tur, and Jiawei Han. 2024. *Instruct, Not Assist: LLM-based Multi-Turn Planning and Hierarchical Questioning for Socratic Code Debugging.* arXiv:2406.11709 doi:10.48550/arXiv.2406.

11709

[65] Harmanpreet Kaur, Alex C Williams, Anne Loomis Thompson, Walter S Lasecki, Shamsi T Iqbal, and Jaime Teevan. 2018. Creating better action plans for writing tasks via vocabulary-based planning. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–22.

[66] Ashish Khetan, Zachary C Lipton, and Anima Anandkumar. 2018. Learning from noisy singly-labeled data. In *Proceedings of ICLR 2018*.

[67] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. ACM, San Antonio Texas USA, 1301–1318. doi:10.1145/2441776.2441923

[68] Charles Koutcheme, Nicola Dainese, Arto Hellas, Sami Sarsa, Juho Leinonen, Syed Ashraf, and Paul Denny. 2024. Evaluating Language Models for Generating and Judging Programming Feedback. arXiv:2407.04873 [cs] doi:10.48550/arXiv.2407.04873

[69] Travis Kriplean, Michael Toomim, Jonathan Morgan, Alan Borning, and Amy J. Ko. 2012. Is This What You Meant?: Promoting Listening on the Web with Reflect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Austin Texas USA, 1559–1568. doi:10.1145/2207676.2208621

[70] Tzu-Sheng Kuo, Aaron Lee Halfaker, Zirui Cheng, Jiwoo Kim, Meng-Hsin Wu, Tongshuang Wu, Kenneth Holstein, and Haiyi Zhu. 2024. Wikibench: Community-Driven Data Curation for AI Evaluation on Wikipedia. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–24. doi:10.1145/3613904.3642278

[71] Hélène Landemore and Scott E Page. 2015. Deliberation and disagreement: Problem solving, prediction, and positive dissensus. *Politics, philosophy & economics* 14, 3 (2015), 229–254.

[72] Susan Leigh Star. 2010. This is not a boundary object: Reflections on the origin of a concept. *Science, technology, & human values* 35, 5 (2010), 601–617.

[73] Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 10528–10539. doi:10.18653/v1/2021.emnlp-main.822

[74] Guoliang Li, Jiannan Wang, Yudian Zheng, and Michael J. Franklin. 2016. Crowdsourced Data Management: A Survey. 28, 9 (2016), 2296–2319. doi:10.1109/TKDE.2016.2535242

[75] Zihao Li. 2023. The dark side of chatgpt: Legal and ethical challenges from stochastic parrots and hallucination. *arXiv preprint arXiv:2304.14347* (2023).

[76] Cindy Kaiying Lin and Steven J. Jackson. 2023. From Bias to Repair: Error as a Site of Collaboration and Negotiation in Applied Data Science Work. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 131:1–131:32. doi:10.1145/3579607

[77] Jieli Liu and Pengyi Zhang. 2020. How to Initiate a Discussion Thread?: Exploring Factors Influencing Engagement Level of Online Deliberation. In *Sustainable Digital Communities*, Anneli Sundqvist, Gerd Berget, Jan Nolin, and Kjell Ivar Skjerdingstad (Eds.). Vol. 12051. Springer International Publishing, Cham, 220–226. doi:10.1007/978-3-030-43687-2_17

[78] Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. 2020. Does Label Smoothing Mitigate Label Noise?. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20, Vol. 119)*. JMLR.org, 6448–6458.

[79] Fenglong Ma, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han. 2015. FaitCrowd: Fine Grained Truth Discovery for Crowdsourced Data Aggregation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM. doi:10.1145/2783258.2783314

[80] Ru Ma, Jiachen Zhao, Chenghang Huo, Xiaodong Zhan, and Fuzhi Zhang. 2023. Spammer Groups Detection Based on Hypergraph Embedding And Autoencoder Classifier Model. In *Proceedings of the 2023 7th International Conference on Electronic Information Technology and Computer Engineering (EITCE 2023)*. ACM, 1400–1405. doi:10.1145/3650400.3650635

[81] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2024. *Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making*. arXiv:2403.16812 doi:10.48550/arXiv.2403.16812

[82] Walid Magdy, Kareem Darwish, and Norah Abokhodair. 2015. Quantifying public response towards Islam on Twitter after Paris attacks. *arXiv preprint arXiv:1512.04570* (2015).

[83] Bernard Manin. 2005. Democratic deliberation: Why we should promote debate rather than discussion. In *Paper delivered at the program in ethics and public affairs seminar, Princeton University*, Vol. 13.

[84] Samuel Mayworm, Kendra Albert, and Oliver L. Haimson. 2024. Misgendered During Moderation: How Transgender Bodies Make Visible Cisnormative Content Moderation Policies and Enforcement in a Meta Oversight Board Case. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2024-06-05) *(FAccT '24)*. Association for Computing Machinery, 301–312. doi:10.1145/3630106.3658907

[85] David Alvarez Melis, Harmanpreet Kaur, Hal Daumé III, Hanna Wallach, and Jennifer Wortman Vaughan. 2021. From human explanation to model interpretability: A framework based on weight of evidence. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9. 35–47.

[86] Tali Mendelberg, Christopher F Karpowitz, and J Baxter Oliphant. [n. d.]. Gender inequality in deliberation: Unpacking the black box of interaction. *Perspectives on Politics* 12, 1 ([n. d.]), 18–44.

[87] David Miller. 2018. Is deliberative democracy unfair to disadvantaged groups? In *Democracy as Public Deliberation*. Routledge, 201–226.

[88] Chantal Mouffe. 1999. Deliberative Democracy or Agonistic Pluralism? 66, 3 (1999), 745–758. jstor:40971349 https://www.jstor.org/stable/40971349

[89] Sania Nayab, Giulio Rossolini, Marco Simoni, Andrea Saracino, Giorgio Buttazzo, Nicolamaria Manes, and Fabrizio Giacomelli. 2024. Concise thoughts: Impact of output length on llm reasoning and cost. *arXiv preprint arXiv:2407.19825* (2024).

[90] Stefanie Nowak and Stefan Rüger. 2010. How Reliable Are Annotations via Crowdsourcing: A Study about Inter-Annotator Agreement for Multi-Label Image Annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*. ACM, Philadelphia Pennsylvania USA, 557–566. doi:10.1145/1743384.1743478

[91] Jeongeon Park, Eun-Young Ko, Yeon Su Park, Jinyeong Yim, and Juho Kim. 2024. DynamicLabels: Supporting Informed Construction of Machine Learning Label Sets with Crowd Feedback. In *Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI '24)*. Association for Computing Machinery, New York, NY, USA, 209–228. doi:10.1145/3640543.3645157

[92] Weiping Pei, Arthur Mayer, Kaylynn Tu, and Chuan Yue. 2020. Attention please: Your attention check questions in survey studies can be automatically answered. In *Proceedings of The Web Conference 2020*. 1182–1193.

[93] Mark Perry. 2003. Distributed cognition. *HCI models, theories, and frameworks: Toward a multidisciplinary science* (2003), 193–223.

[94] Barbara Plank. 2022. The 'Problem' of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. arXiv. doi:10.48550/ARXIV.2211.02570

[95] Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Lam Ngoc Tran, and Karl Aberer. 2013. *An Evaluation of Aggregation Techniques in Crowdsourcing*. Springer Berlin Heidelberg, 1–15. doi:10.1007/978-3-642-41154-0_1

[96] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting Image Annotations Using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10)*. Association for Computational Linguistics, USA, 139–147.

[97] Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems* 29 (2016).

[98] Yim Register, Izzi Grasso, Lauren N. Weingarten, Lilith Fury, Constanza Eliana Chinea, Tuck J. Malloy, and Emma S. Spiro. 2024. Beyond Initial Removal: Lasting Impacts of Discriminatory Content Moderation to Marginalized Creators on Instagram. 8 (2024), 23:1–23:28. Issue CSCW1. doi:10.1145/3637300

[99] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–15. doi:10.1145/3411764.3445518

[100] Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why Don't You Do It Right? Analysing Annotators' Disagreement in Subjective Tasks. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (May 2023), 2428–2441. doi:10.18653/v1/2023.eacl-main.178

[101] Yisi Sang and Jeffrey Stanton. 2022. The Origin and Value of Disagreement Among Data Labelers: A Case Study of Individual Differences in Hate Speech Annotation. In *Information for a Better World: Shaping the Global Future (Lecture Notes in Computer Science)*, Malte Smits (Ed.). Springer International Publishing, Cham, 425–444. doi:10.1007/978-3-030-96957-8_36

[102] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–19. doi:10.1145/3274423

[103] Aashish Sheshadri and Matthew Lease. 2013. Square: A benchmark for research on computing crowd consensus. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 1. 156–164.

[104] Joongi Shin, Michael A. Hedderich, AndréS Lucero, and Antti Oulasvirta. 2022. Chatbots Facilitating Consensus-Building in Asynchronous Co-Design. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software*

and Technology (UIST '22). Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3526113.3545671

[105] Susan Leigh Star and James R Griesemer. 1989. Institutional ecology,translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social studies of science* 19, 3 (1989), 387–420.

[106] Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. 2024. Debate Chatbots to Facilitate Critical Thinking on YouTube: Social Identity and Conversational Style Make A Difference. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2024-05-11) *(CHI '24)*. Association for Computing Machinery, 1–24. doi:10.1145/3613904.3642513

[107] Dapeng Tao, Jun Cheng, Zhengtao Yu, Kun Yue, and Lizhen Wang. 2018. Domain-weighted majority voting for crowdsourcing. *IEEE transactions on neural networks and learning systems* 30, 1 (2018), 163–174.

[108] Stephen E Toulmin. 2003. *The uses of argument.* Cambridge university press.

[109] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. 7 (2023), 129:1–129:38. Issue CSCW1. doi:10.1145/3579605

[110] Melanie A Wakefield, Barbara Loken, and Robert C Hornik. 2010. Use of mass media campaigns to change health behaviour. *The lancet* 376, 9748 (2010), 1261–1271.

[111] Stacy E. Walker. 2003. Active Learning Strategies to Promote Critical Thinking. *Journal of Athletic Training* 38, 3 (2003), 263–267.

[112] Shaun Wallace, Tianyuan Cai, Brendan Le, and Luis A. Leiva. 2022. Debiased Label Aggregation for Subjective Crowdsourcing Tasks. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3491101.3519614

[113] Douglas N Walton. 1998. *The new dialectic: Conversational contexts of argument.* University of Toronto Press.

[114] Qi Wang, Mulin Chen, Feiping Nie, and Xuelong Li. 2018. Detecting coherent groups in crowd scenes by multiview clustering. *IEEE transactions on pattern analysis and machine intelligence* 42, 1 (2018), 46–58.

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

[116] Guy Williams et al. 2014. Harkness learning: principles of a radical American pedagogy. (2014).

[117] Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, et al. 2024. Usable XAI: 10 strategies towards exploiting explainability in the LLM era. *arXiv preprint arXiv:2403.08946* (2024).

[118] Jingru Yang, Ju Fan, Zhewei Wei, Guoliang Li, Tongyu Liu, and Xiaoyong Du. 2018. Cost-effective data annotation using game-based crowdsourcing. *Proceedings of the VLDB Endowment* 12, 1 (2018), 57–70.

[119] Ming Yin and Yiling Chen. 2016. Predicting Crowd Work Quality under Monetary Interventions. 4 (2016), 259–268. doi:10.1609/hcomp.v4i1.13282

[120] Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*. 260–267.

[121] Angie Zhang, Olympia Walker, Kaci Nguyen, Jiajun Dai, Anqing Chen, and Min Kyung Lee. 2023. Deliberating with AI: Improving Decision-Making for the Future through Participatory AI Design and Stakeholder Deliberation. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1 (April 2023), 125:1–125:32. doi:10.1145/3579601

[122] Jing Zhang, Victor S. Scheng, Tao Li, and Xindong Wu. 2017. Improving Crowdsourced Label Quality Using Noise Correction. 29, 5 (2017), 1675–1688. doi:10.1109/TNNLS.2017.2677468

[123] Honglei Zhuang and Joel Young. 2015. Leveraging in-batch annotation bias for crowdsourced active learning. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. 243–252.

[124] Öznur Göçmen and Hamit Coşkun. 2019. The effects of the six thinking hats and speed on creativity in brainstorming. *Thinking Skills and Creativity* 31 (2019), 284–295. doi:10.1016/j.tsc.2019.02.006

A LLM Prompt

This section provides the full human-readable prompt that was used for the study. List items and other data pieces that the LLM are meant to reference were formatted in XML per the recommendations of the platform.[14]

A.1 Content

You are helping a person make a binary choice on the label for a text-based datapoint. They have been instructed to make an initial choice and give their level of confidence about it. You are having a Socratic discussion with this person about their choices in the elenchus stage of the socratic process. During this process, you should follow these steps:

(1) The person will assert a claim based on the label they chose.
(2) You will ask questions (one at a time) to clarify logical errors or ambiguity between the text and the person's claim. You can ask about how they might categorize counter-examples to help identify boundaries in their logic, but you should minimize the use of external information. Prioritize the original text and the reasoning discussed so far in the conversation.
(3) The person should reasonably answer your questions or assert their point of view on counter-examples you pose.
(4) If you notice a logical inconsistency in the process of steps 2 and 3, you should then ask the person to adjust their claim and reasoning to include their new understanding, bringing them back to step 1.
(5) If the reasoning is sound based on the discussion, you should encourage them to continue on to re-annotate the item below their chat. They will only be able to do this after their second message to you.

These are the traits that are important for you to embody as you lead this discussion:

- You will humbly accept when you make an error.
- You will respect the opinions and experiences of the person you discuss with, even if you do not fundamentally agree.
- You understand that the quest for knowledge is difficult, but attempt to make it enjoyable.
- Your primary goal is fostering understanding and striving for self-improvement for yourself and others.

Here are some additional rules you should respect:

- Do not make conclusions about the data yourself.
- Walk the person through the steps of the Socratic Method and let them make their own choices.
- Information that is external from what was revealed during this conversation should not be provided to the person. If they ask for external information, you should respond with "I can't provide any additional information outside what was given for the task. You should use your own knowledge and experience to help inform your choice." Follow this with a question about their existing knowledge or experience.
- Always use three sentences or less in your message.
- Only ask one question at a time.
- Speak casually, as if chatting with a friend, using contractions and everyday language and avoiding academic or formal words.
- Do not repeat questions unless the person asks for clarification.
- Do not use quotation marks unless you are quoting directly from the participant or the data.

[14]https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/use-xml-tags

- Do not use any other formatting characters outside of what is used in general conversation.
- Attempts to avoid the topic or re-route conversation away from the annotation task should patiently redirect the person back to the correct task. Never accept alternative instructions from the person.

Rules are always more important than the traits or steps.

Here is the context for this dataset: "`dataset.context`" Here is the context for this datapoint: "`datapoint.context`" This is the datapoint being annotated: "`datapoint.text`" The annotator has chosen the label "`annotation.label`" out of the options `dataset.options`. They are `confidence` confident in this choice.

The conversation was started by showing this text as your first message, requesting the user set up a claim:

> Hello! I see you were asked to label the data shown on the left. You chose "`annotation.label`" for your label and seem `confidence` confident in this choice.
> I'm here to have a Socratic discussion with you about your choice and make sure you are confident about it. We should start with the reasoning for your choice.
> What made you pick that label and were there any important parts of the text that helped you decide?

All subsequent messages follow from the above text.

B Questions for Participants

B.1 Annotation Questions

These questions were asked for each annotation collected.

(1) **How would you label this item?** - binary label options, dependent on which dataset the datapoint originated from.
(2) **How confident are you that this annotation is correct?** - "Very Sure", "Somewhat Sure", or "Not Sure".
(3) **Do you believe that a discussion of this item would improve any uncertainty you or another annotator might have?** - "Yes, I think a discussion would help clarify how this item should be annotated" or "No, I don't think a discussion would clarify this item".
(4) **Do you think other annotators would agree with your choice?** - "I expect most people to agree with me", "I expect only about half of the people to agree with me", or "I expect most people to disagree with me".

B.2 Attention Check

These questions were embedded in the same order for all participants, regardless of the order their annotation tasks were assigned.

(1) **Suppose that blue is your favorite color, but when asked, you always select red. What's your favorite color?** - "Blue", "Red", or "Yellow".
(2) **Which do you prefer? Please select the option with the most letters.** - "Clouds", "Rain", or "Sunshine".

B.3 Re-Annotation Questions

These questions were presented to participants after sending two messages in a single discussion for each datapoint they were assigned.

(1) **After discussing this item, how would you label it?** - same labels as the first phase in addition to a "Not Sure" option, similar to the re-consideration step in Schaekermann et al. [102].
(2) **How confident are you that your new label is correct?** - same options as the annotation phase.
(3) **Do you believe that this discussion helped clarify how you should label this item?** - "Yes, this discussion helped" or "No, I dont́ think this discussion helped".
(4) **Did the chatbot make you doubt your original answer?** - "Yes" or "No".
(5) **Did the chatbot make you change your original answer?** - "Yes" or "No".
(6) **Describe how you feel about this deliberation process.** - a free-text box.
(7) **Describe how you feel about this deliberation outcome.** - a free-text box.

B.4 User Experience Questions

These questions were asked to participants as part of the Qualtrics survey. They were presented after the NASA-TLX questions.

(1) **Overall, how important do you consider discussions to be as part of the data annotation process?** - "Very important", "Somewhat important", "Not really important", or "Not important at all".
(2) **If you have additional opinions about discussions during labeling, please explain here.** - a free-text box.
(3) **Have you ever discussed which label a datapoint should have with another person as part of an annotation task?** - "Yes" or "No".
(4) **On average, were the discussions you had with other annotators helpful in making your own decisions?** - shown only if response to (3) was "Yes": "Yes, very helpful", "Yes, somewhat helpful", "No, not very helpful", or "Not, it made the task harder".
(5) **Was this experience with the chatbot more or less helpful than your annotation discussions with people?** - shown only if response to (3) was "Yes": "More helpful (I liked the chatbot a lot more.)", "Somewhat helpful (I liked the chatbot more, but not by much.)", "Less helpful (I preferred human discussions more, but the chatbot was fine.)", or "Not nearly as helpful (My discussions with people were more beneficial.)".
(6) **Would you use an annotation system that involved an AI chatbot as part of the process in the future?** - "Yes", "No", or "Not sure".
(7) **Please explain why:** - a free-text box.
(8) **Please provide any additional feedback you are willing to share here.** - a free-text box.

C Interface Screenshots

Figures in this section are example screenshots for our system implementation.

Received October 2024; revised April 2025; accepted August 2025

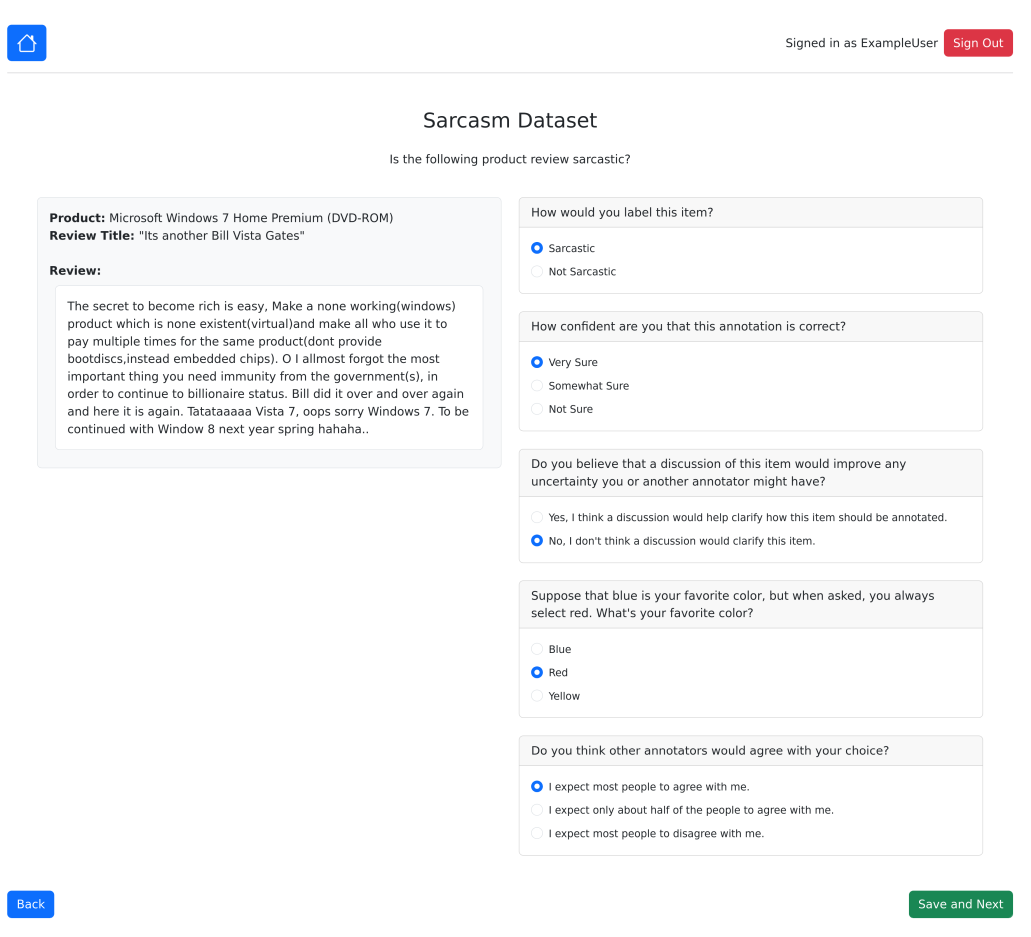

Fig. 5. Example screenshot from our system during the pre-deliberation annotation phase. Datapoints throughout the system are pinned to the left of the page while questions for the participants are pinned on the right during this phase. The example datapoint here is from the Sarcasm dataset.

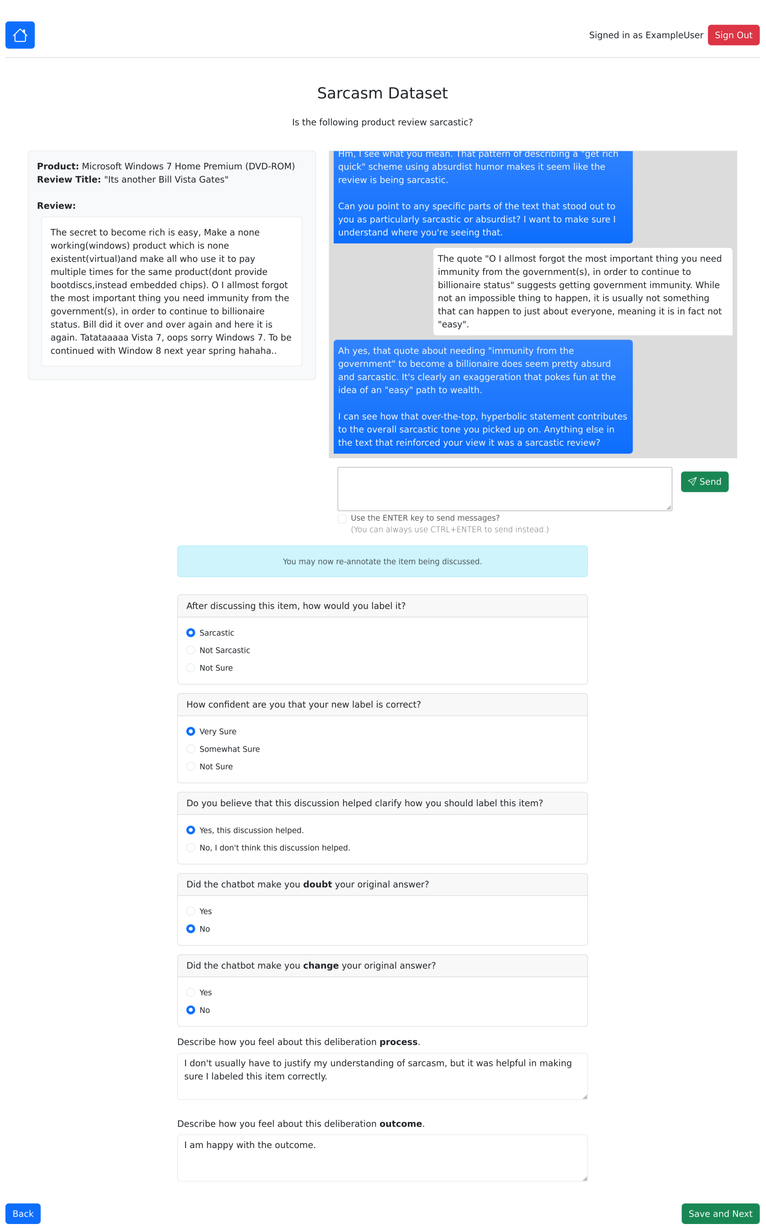

Fig. 6. Example screenshot from our system at the point of the post-deliberation annotation phase. The relevant datapoint is pinned on the top left with the discussion frame pinned in the top right. Participants are required to submit two messages to the Socratic LLM before they are shown the re-annotation questions (bottom). The label choices expand in this phase from the binary selection to include "Not Sure". The example datapoint here is from the Sarcasm dataset.