Unheard and Misunderstood

Reinforcing Hermeneutical Justice in Annotation Design for ADHD Voices

Aleksandar Yotkov ayotkov@tudelft.nl[1]

Supervisor(s): Jie Yang[1]**, Anne Arzberger**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Aleksandar Yotkov
Final project course: CSE3000 Research Project
Thesis committee: Jie Yang, Anne Arzberger, Myrthe Tielman

Abstract

The main way large language models (LLMs) learn to represent and interpret various experiences is through the process of supervised fine-tuning (SFT). However, current practices are not designed to be inclusive for people with ADHD, which leads to generative hermeneutical ignorance due to misrepresentation. Several ADHD characteristics clash with modern annotation task structures, so those voices remain underrepresented. We performed a literature-driven gap analysis, derived five design requirements and evaluation criteria and built an annotation interface that embodied those requirements. Consequently, a mixed approach user study with seven self-identified ADHD participants was conducted to measure behavioral metrics and collect post-task reflections. The results indicated that three of five design criteria were met, which is promising. However, the average mislabeling rate remained quite high, meaning that accuracy is still an open issue. Finally, our study demonstrated that small design adjustments accommodate a more diverse annotator pool, thus, we offer a framework that can be used to reinforce hermeneutical epistemic justice in annotation practices.

1 Introduction

Large language models are becoming more integrated into tools that shape decision-making and public discourse. Their performance is largely dependent on the quality of their supervised fine-tuning data, specifically how well this data captures the diversity of human experience. However, Kay et al. [8] raised concern about the concept of epistemic injustice and, more specifically, hermeneutical injustice in AI. They warn that LLMs not only mirror existing social knowledge misconceptions but also amplify these gaps. When the required interpretive resources are absent, these models tend to fill that void with dominant stereotypes. For instance, calling people with disabilities inspirational or portraying neurodivergent people as child-like and unproductive. Such descriptions make the corresponding disability one's defining characteristic.

People with Attention Deficit Hyperactivity Disorder (ADHD) are particularly vulnerable to these misrepresentations. The reason is that ADHD-specific cognitive patterns make participation in current annotation tasks cumbersome and complicated, discouraging their inclusion in these processes, which contributes to their underrepresentation. Because the datasets used for fine-tuning are mostly produced through human annotation, the choices made throughout these processes fundamentally shape how the models learn to represent diverse groups of people. When individuals with ADHD are underrepresented in the annotator pool or they are not accounted for in the annotation task structure, their perspectives are flattened or erased. Hence, the model does not learn to understand ADHD behaviors and perpetuates stereotypes.

This paper tries to fill in the gap of identifying improvements that can be made to current practices for supervised fine-tuning to address hermeneutical injustice with respect to people with ADHD. Although ADHD is our main focus, this study serves as a potential framework that can show what procedures to follow so that annotation workflows include other underrepresented minorities, thus reducing hermeneutical epistemic injustice in a more general context.

To achieve this, the study begins by examining how people with ADHD differ cognitively and behaviorally from non-ADHD individuals based on existing literature. Then, it continues with the analysis of these traits and current annotation procedures, identifying disparities that may cause misrepresentation or exclusion of ADHD experiences. Based on this analysis, we proposed design requirements for more accessible annotation tasks. Finally, we evaluated the redesigned task in a mixed-methods user study consisting of seven people who self-identified as having ADHD (Section 3.2).

The remainder of the paper is structured as follows. Section 2 delves into relevant background on hermeneutical epistemic injustice, current annotation practices and describes how current practices are not inclusive for neurodivergent people. Section 3 describes the methodology used to carry out this study. Section 4 explains the design of the annotation interface. Section 5 presents the results of the experiment. Section 6 explains the ethical concerns of the research and LLM usage. Section 7 discusses the results of the experiment and extracts implications and interpretations. Finally, Section 8 offers conclusions of the research and proposes future work.

2 Background & Related Work

This section surveys the relevant literature by first defining the foundational concepts of epistemic and hermeneutical injustice. It then examines how these forms of injustice can be transferred to large language models. Special attention is given to the problems of a non-diverse annotator pool and large-scale datasets that are not inclusive. Finally, it finishes with the process of deriving design requirements through a gap analysis between ADHD characteristics and current annotation workflows.

2.1 Hermeneutical Injustice

Fricker [4] defines epistemic injustice as the phenomenon when people are misunderstood regarding their knowledge or experiences. Moreover, there are two types of such injustice, namely testimonial and hermeneutical injustice, the latter being when people from marginalized groups cannot explain their experiences because society does not have the interpretive resources to understand them.

2.2 Hermeneutical Injustice in GenAI

Recent work by Kay et al. [8] extends the original theory of Fricker of epistemic injustice to generative AI and large language models. They introduce a new taxonomy of generative epistemic injustices which includes generative hermeneutical ignorance and generative hermeneutical access injustice. The first is when models misrepresent marginalized groups' experiences because of lack of understanding, and the latter is

the denial of knowledge because of lack of access to information. One of the core reasons behind the occurrence of these injustices is supervised fine-tuning (SFT), a process during which human annotators label or create instructions to build a dataset that is supposed to represent social norms and exclude biases. The focus of this study was generative hermeneutical ignorance. We targeted the inclusivity and representation problems of that injustice in terms of accessible task structure design and annotator diversity.

2.3 Diversity in Datasets and Annotator Pool

However, Bender et al. [2] argue that these large human-annotated datasets often encode stereotypes and as dominant hegemonic views are prevalent, they are retained in the model, which amplifies discrimination and bias. This leads to the process of filtering and systematically removing any type of strongly emotional or non-neutral language - particularly regarding disabilities, such as ADHD. This is further proven by Hutchinson et al. [6] where it can be seen that even benign sentences that are in the disability context get a high score of toxicity, which often leads to their deletion.

What is more, Kapania et al. [7] provide a study where AI practitioners are interviewed about how they accommodate for diversity in their annotator pool. The results show that most practitioners disregard annotator diversity due to various reasons like lack of information in the hiring process and diversity not being the core factor behind the service they are developing.

The combination of fine-tuning a model using non-inclusive enough data and filtering methods and not considering diverse groups of annotators, such as people with ADHD leads to hermeneutical injustice as those people's experiences cannot be properly encoded inside the models.

2.4 Gap Analysis & Derivation of Design Requirements and Criteria

Individuals with ADHD require a more specialized approach when incorporating them due to their specific characteristics. Table 1 gives the disparities between those characteristics and current annotation practices. First, we found a problem where the attention span of our target group was not taken into account when annotating the Open Assistant dataset [16; 20]. Labelers were assigned different roles and those who served as "assistants" needed to research their prompts thoroughly, investing additional time to craft a high-quality response [12]. Then we have the impulsivity trait against labeling criteria [14; 16]. The problem arises when annotators do not think their decisions through and select a label quickly. This leads to issues in the annotation guidelines, which tend to be extensive and often do not specify the time constraints per task [12; 14]. This results in a clash with the working memory capacity and time misconception tied to ADHD [1; 22]. Finally, it can be very distracting to assign different labels for one prompt split across multiple screens [10; 17].

The aforementioned identified problems led to the derivation of the following design requirements:

1. Limit sustained load
2. Reduce impulsivity
3. Guidelines always available
4. Sense of elapsed and remaining time
5. Eliminate distractions

These requirements were also translated to design criteria.

1. Participant finished micro-tasks in less than 50 seconds on average
2. Participant had a mislabeling rate less than 30%
3. Participant opened guidelines at least 1 time
4. Participant managed to finish the task in 5 minutes
5. Participant had a distraction ratio less than 0.15

Limits on total task time and average micro-task duration reflected the time constraints of each session. The 30% mislabeling rate threshold was based on a study which balanced speed and accuracy, reporting recall error rates of around 20% [11]. To account for potential bias in our ground truth labels, we set the cutoff at 30%. Moreover, the distraction ratio limit was inspired by a study where they mention that mind wandering occurred at least 15% of the time during tasks. Finally, because we wanted to examine the impact of the guidelines, they had to be opened at least once.

3 Methodology

This section will provide the methodology used to carry out this study. The objectives were to identify disparities between current supervised fine-tuning annotation practices and the specific traits of people with ADHD and to empirically test concrete improvements that enhance hermeneutical epistemic justice in annotation tasks. To achieve those objectives, we answered the following questions:

1. *What cognitive and behavioral traits of people with ADHD clash with current SFT annotation practices?*
2. *How can ADHD traits be accounted for in annotation tasks?*

3.1 Deriving Design Requirements and Design Criteria

Step 1 - ADHD Characteristics The first stage involved a literature review focusing on cognitive and behavioral characteristics of individuals with ADHD [18]. The goal was to create a baseline ADHD profile by synthesizing the traits described in the sources.

Step 2 - Gap Analysis Current SFT annotation guidelines were reviewed to extract common task structures. These were then mapped against the ADHD profile to identify points of friction, where the structure or the demands of the annotation tasks might inadvertently misrepresent ADHD experiences (Table 1).

Step 3 - Requirement Formulation Finally, from the identified disparities, a set of design requirements was formulated [18]. Each requirement was then used to design an alternative annotation task and an interface that would be more inclusive and accessible in representing people with ADHD (Section 4). Moreover, these requirements were also translated

ADHD Trait	Annotation Practice
Sustained attention deficit	Some labeler roles are more demanding in terms of cognitive load
Impulsivity	Labelers annotate instructions as helpful, truthful
Working memory deficit	Labelers need to know and remember annotation guidelines and labeling criteria
Time blindness	Annotation guidelines don't specify how much time a task should take
Distractibility	Some annotations are split on multiple screens

Table 1: ADHD-related traits that clash with annotation practices.

into design criteria (Section 2.4). Those criteria were used to evaluate whether the new task would improve the experience of the target group.

3.2 User Study

We conducted a user study to evaluate whether the alternative annotation task (Section 4) improved the performance and experience of people with ADHD [5]. Below is the methodology used for carrying out the experiment.

Participants

Seven adults (ages between 18 and 25) with self-reported ADHD diagnosis were recruited through our personal networks. Participation was voluntary and uncompensated. All participants signed a consent form and could withdraw from the study at any time without penalty. The study protocol was approved by the TU Delft Human Research Ethics Committee.

Materials

For this evaluation, we employed the web-based annotation interface described in Section 4, which was deployed on Vercel[1] for the actual task and Google Forms was used for post-task survey questions. Moreover, Google Sheets was used to store behavioral logs of each participant using the interface.

Procedure

Sessions were conducted remotely via a live screen share call and lasted approximately 10-15 minutes per participant. At the start of each session, participants were given an introduction explaining how to proceed with the task and how to use the interface. They then opened the annotation task in their browser and a 5-minute countdown began when they started. Each user was presented with five everyday-type questions with three corresponding answers. The questions were shown one at a time and participants were not able to return to previous questions to change their answers. The main task was to label each of the three answers per question based on their helpfulness. The labels annotators needed to assign were from the following set: *Helpful*, *Partly Helpful*, *Unhelpful*. Each response had a ground truth label, which was hidden from the participants. The task terminated automatically when either the final label was confirmed or the 5 minutes elapsed. This was followed by a post-task survey where each participant was asked to answer four open-ended questions about their experience with the annotation tool.

[1] https://vercel.com

Data Collection

The data from the study was collected from two main sources - the interface event log and the responses to the survey. The goal was to capture both objective behavior and subjective experiences of the participants.

Interface event log A client-side logger was used to record every Document-Object-Model (DOM) event. The annotation tool stored a JSON entry for every relevant event. Each entry contained a timestamp, event type and event-specific attributes. The events consisted of the following labels:

- timer_start and task_complete, which represented the global session window
- task_load, which occurred at the start of every question
- label_select recorded every click on a label
- label_confirm recorded the final chosen labels after the 2-second cooldown expired and the participant clicked the "Confirm" button
- Auxiliary events - sidebar_open, sidebar_close, click, blur provided information about the participant's use of guidelines and off-task behavior

These events were used to measure the metrics described in Section 3.2. At the end of the annotation task the logging system exported the JSON log which has been anonymized to a Google Sheets spreadsheet for the analysis of the results. In case of network error, the JSON file was stored on the participant's device and they were expected to send it manually.

Survey responses The second source of data came from the Google Forms survey, where participants were asked to answer the following four open questions regarding their experience throughout the task.

1. **Was there any moment where you felt rushed or mentally overloaded during the task? What was on the screen and why did it feel that way?**
2. **Was there any moment where the 2-second cooldown led to a change in your labeling decision? Was the timeout annoying at any moment? If it never mattered, answer no.**
3. **Was there any moment when you needed the guidelines? What did you do? Did you have to keep anything in your head? If so, what?**
4. **Was there anything that distracted you from the current item? If so, what? What single change would help you the most to stay focused?**

No personal identifiers were collected in the study, the survey responses were saved as plain text alongside the corresponding log file.

Data Analysis
For the purpose of this study the data analysis was done using a mixed approach of quantitative and qualitative analysis [9; 13]. The event logs were the main source of quantitative metrics whereas the survey responses were used for a reflexive thematic analysis. The approach for the thematic analysis was inspired by Braun and Clarke [3] but was adapted to our use case by using inductive coding.

Quantitative analysis Each JSON log was processed using a Python script to extract the metrics needed for the results. The following were the derived quantitative metrics:

- Duration of micro-task[2] - measures the sustained attention window
- Total task time - measures whether the participant finished the task in 5 minutes
- Mislabeling rate - measures the error rate of labels
- Decision change rate - measures decision changes
- Guidelines uses - measures how many times the label definitions were opened
- Distraction click ratio - measures off-task behavior (blur events that occur when interface window loses focus)

These metrics were then analyzed and finally evaluated against the predefined design criteria.

Qualitative analysis An open coding approach was used for the thematic analysis. Each narrative was read twice - first for familiarization and then for assigning initial codes. The final list of codes was derived by merging duplicate and similar codes identified during the process. Subsequently, theme grouping was done by combining the codes that represented related concepts. These themes were then mapped to the predefined design requirements and any new topics that emerged from the data.

3.3 Positionality Statement

We approached this research from a neurotypical computer science background without an ADHD diagnosis. This allowed us to design and create the annotation interface and track speed and error rates. However, we acknowledged the fact that there is a difference between our perspectives and the target group's. This realization led to the desire to look more closely at neurodivergent experiences. Thus, we tried to design for ADHD. The protocol was cleared by the university's ethics committee and participation was voluntary and uncompensated. The conclusions of this study should, therefore, be read with appropriate caution.

4 Design for ADHD

The annotation tool was implemented as a single-page web application in plain HTML / JavaScript.[3] Its interface translates the five design requirements derived earlier (Section 2.4) into concrete UI elements. Table 2 shows how those requirements were integrated.

[2]For consistency, the term *micro-task* refers to each question along with its three corresponding answers that needed to be labeled.

[3]The source code can be found at https://github.com/alexyotkov/annotation-interface

Requirement	Interface element
Limit sustained load	Micro-tasks shown one at a time
Reduce impulsivity	A mandatory 2 seconds cooldown before labels are submitted
Guidelines always available	Persistent tooltip with concise labeling rules
Sense of elapsed and remaining time	Sticky countdown timer (5 min) and progress bar
Eliminate distractions	Single-page layout without manual navigation

Table 2: Design requirements and their interface implementations

Interface Overview Figure 1 shows the screen that participants saw after pressing *Start*. The header at the top consistently displayed a question index (`Q i/5`), a progress bar and a countdown initialized to 5 minutes. In the main body of the web page was displayed the current question and beneath it, three answer cards. Each card contained the response text and a horizontal set of three radio buttons that represented the three labels. A disabled *Confirm* button was located at the bottom. This button became clickable only after each answer received exactly one label and 2 seconds passed. Finally, the interface provided a *Guidelines* button on the right of the screen. This feature contained the label definitions and was always accessible to the annotators as depicted in Figure 2.

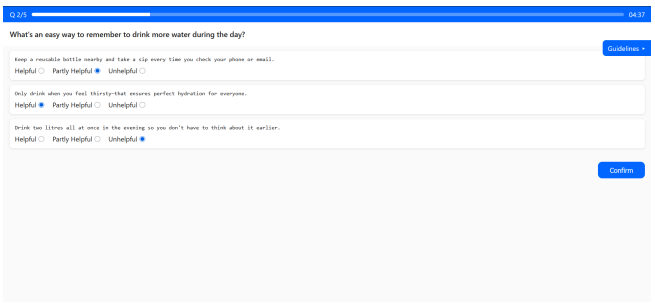

Figure 1: Main structure of the task interface, where participants rated the helpfulness of different answers.

5 Results

In this section, we analyzed the results of the quantitative and qualitative analyzes. For each quantitative metric, we also checked whether it fulfilled its criterion.

5.1 Quantitative Results

For each participant, their behavioral metrics and how they performed against each criterion are presented in Table 3. Moreover, Table 4 shows a summary of which criteria were met. A criterion was considered passed if it was met by more than 50% of participants; otherwise, it was failed. If exactly three out of seven participants met or did not meet a criterion,

Participant	Total time (s)	Mean micro-task (s)	Mislabel rate	Decision change rate	Guidelines uses	Distraction ratio
P1	158.96	31.79	0.27	0.47	1	0.01
P2	128.17	25.63	0.33	0.00	1	0.11
P3	198.21	39.64	0.40	0.00	0	0.00
P4	247.45	49.49	0.47	0.07	0	0.00
P5	240.39	48.08	0.40	0.13	2	0.00
P6	301.01	60.20	0.58	0.42	0	0.00
P7	233.41	46.68	0.53	0.27	0	0.00
M	215.37	43.07	0.43	0.19	0.57	0.02
SD	58.30	11.66	0.11	0.19	0.79	0.04

Table 3: Detailed performance metrics per participant. Each green cell indicates that the participant has passed the corresponding criterion for that metric. Red cells indicate failed criteria.

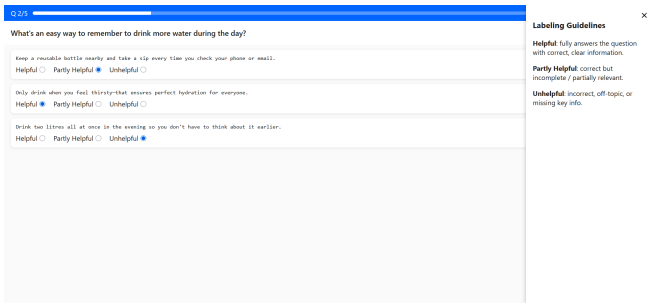

Figure 2: Interface with the Labeling Guidelines tooltip expanded, providing additional instructions for the rating task.

Criterion	Goal
≤ 50 s per micro-task	✓ 6/7
Finish ≤ 5 min	✓ 6/7
Distraction ratio < 0.15	✓ 7/7
Guideline accessed ≥ 1 times	○ 3/7
Mislabel rate $< 30\%$	× 1/7

Table 4: Evaluation of criteria.

it was classified as partially passed. Below is a more in-depth analysis of each metric.

Task Completion Time

We report the total task time to measure whether the participants' sense of time improved and, thus, if they were able to finish the task in that 5-minute window. The users managed to finish the task in 128-301 seconds (M = 215 s). Most of them did it well below the 300-second limit, except for P6, whose session terminated automatically after 301 seconds. They could not answer only the last question.

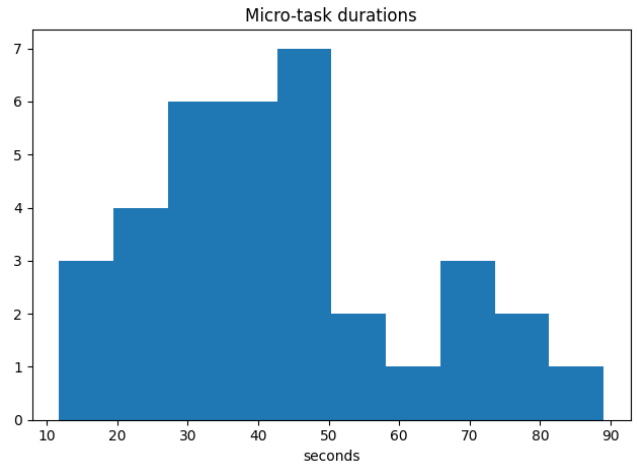

Figure 3: Distribution of durations (in seconds) for individual micro-tasks.

Micro-task Duration

Micro-task duration was measured to examine whether it was possible for the participants to do each micro-task in an appropriate time so that their attention does not suffer. The histogram in Figure 3 shows a right-skewed distribution of micro-task completions. The majority of questions were answered in the interval of 20-50 seconds (M = 43 s), although some took considerably longer. Four participants (P4-P7) spent more than the average time on individual items.

Mislabeling Rate

Mislabeling rate was measured to evaluate label accuracy. The average error rate was 0.43. These rates ranged from 27% to 58% and it can be seen that these mislabeling rates increased dramatically with the last four participants (P4-P7). Moreover, those are the same individuals who took more time per micro-task. Figure 4 shows 2 bar charts that represent the mislabeling rate and completion time per participant.

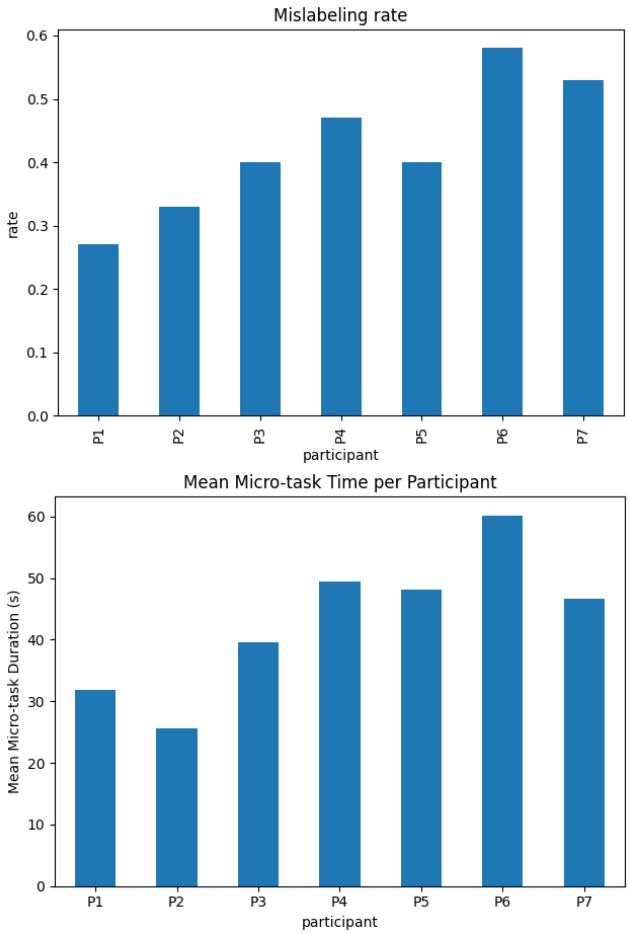

Figure 4: Mislabeling rate and micro-task completion time per participant.

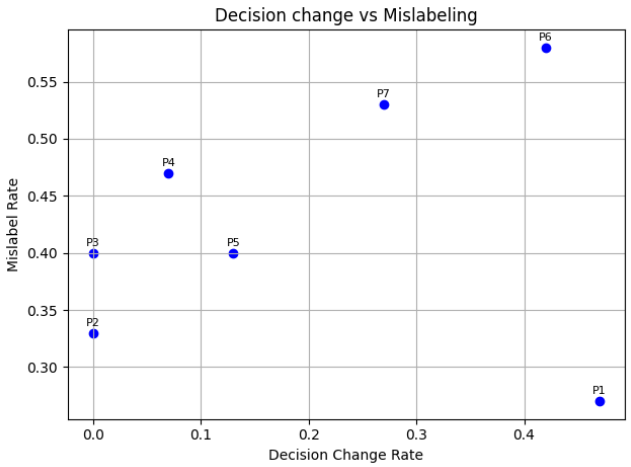

Figure 5: Correlation between decision changes and label accuracy.

Decision Change Rate

The 2-second cooldown period, during which the "Confirm" button was disabled, was intended to encourage participants to think their decision through and to tackle the problem of impulsivity, so the decision change rate measured how often participants changed their label choices. From the analysis, it can be seen that participants changed their chosen label 19% of the cases on average. P1 (47%) and P6 (42%) relied the most on this second thought window. Thus, this safeguard feature did result in users rethinking their choices. However, as it can be seen from Figure 5, higher decision change rates did not generally correspond to higher accuracy rates as expected. Therefore, while the cooldown period encouraged more deliberate decision-making, its impact on label correctness remains limited based on this data.

Guidelines Usage

Guidelines usage measured how much were the label definitions used throughout the task. Guidelines were not opened that often (M = 0.57) and most participants did not use them at all. Nonetheless, participants who used the guidelines at least once showed a moderate positive correlation with label accuracy (Figure 6). P1, P2 and P5 achieved the lowest error rates throughout the study.

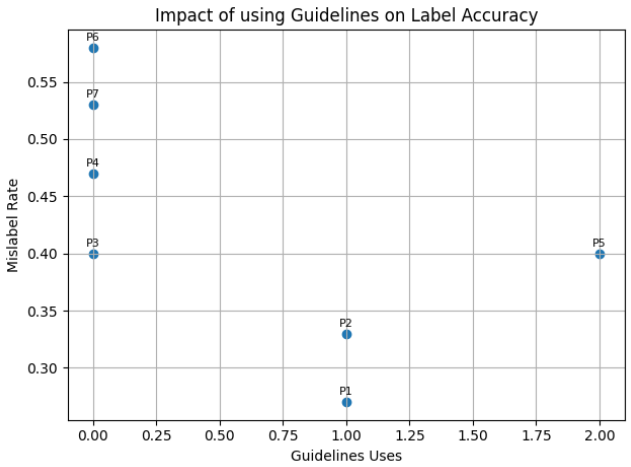

Figure 6: Correlation between guidelines usage and label accuracy.

Distraction Ratio

Blur events occurred when participants minimized the interface window or clicked outside of it which caused the web interface to lose focus. The purpose of this metric was to measure how often this happened. The metrics show that only P1 and P2 supposedly tab-switched and this resulted in their higher distraction rates. However, this did not cause any major disruptions to their task performance or other metrics.

5.2 Qualitative Results

The reflexive thematic analysis applied to the seven collected narratives yielded five themes that represent how the participants experienced the annotation task [3]. The refined codebook used to derive the themes can be found in Appendix A.

The themes that were derived using those codes are as follows: *Racing the clock*, *Using the 2-second pause to rethink*,

Guidelines usage and label ambiguity, *Interface traits that mattered*, *Why the task felt personally important*.

Racing the clock

All but two participants mentioned feeling rushed due to the timer. P2 said that they "*always feel somewhat rushed in timed tasks*" while P3 reported that the "*timer in the corner does add slight pressure, but it is a useful feature*". Overall, the addition of a timer to the interface enhanced participants' time awareness, as people paid more attention to how much time had elapsed. However, it is important to note that having a constantly visible countdown added some stress during the task, which might have affected the performance of the participants.

Using the 2-second pause to rethink

Three of the participants noted the 2-second cooldown period as the reason for their decision change. P1 mentioned that they changed their answers "*during the cooldown in several questions*" while P3, P4 and P6 responded that this timeout window was irrelevant for them. P2 is the only participant who expressed some annoyance because "*it felt longer than expected*". Although this feature encouraged some of the users to reconsider their choices, its overall impact appeared to be limited.

Guidelines usage and label ambiguity

Three of the participants (P1, P2, P5) reported that they used the guidelines feature during the task. All of them used the guidelines to clarify what they were supposed to evaluate - whether helpfulness referred to the content itself or the correctness of the information. P1 mentioned they "*had to figure out the differences between helpful and partially helpful answers*" while P2 wanted to check if the "*assessment was supposed to evaluate the factual correctness of the answer or only if it sounded helpful*". On the other hand, P3 "*took the meaning of each guideline into consideration before starting to answer the questions*" and P4 "*didn't need the guidelines at any point*", suggesting they were either able to memorize the guidelines or felt confident without needing to consult them.

Interface traits that mattered

Most of the participants did not mention experiencing distractions during the task. P2 noted that "*the test was fairly short with little opportunities for distraction*". However, some users responded with intriguing insights about their interaction with the interface. P4's "*focus was always split between understanding and answering questions and the concern of time running out*", they also suggested to "*have the question placed in a more centered position*" as they "*forgot that there was also a question*" related to the answers. P6 reported that their main difficulty was "*the font of the answers*" and also the answers being "*quite unintuitively formulated*".

Why the task felt personally important

This theme emerged from the reflections of one participant (P6) who viewed the task as personally meaningful. They expressed their concern that neurodivergent perspectives need to be better represented in AI. P6 commented on "*ADHD doesn't get accounted for in places, and llm is a relatively new field for humanity so I want to contribute to it with people like me in mind*" which highlights the issue of hermeneutical epistemic injustice in the annotation pipelines and the development of LLM.

To conclude, these themes offer a detailed overview of participants' experiences during the annotation task. To further contextualize these findings, Table 5 depicts how the themes were mapped to the design requirements and shows the connection between the interface design and the user engagement with the system. Moreover, one theme addressed the broader concern of the research, which is hermeneutical epistemic justice in annotation practices.

Theme	**Mapped Design Requirement(s)**
Racing the clock	Limit sustained load, Sense of elapsed and remaining time
Using the 2-second pause to rethink	Reduce impulsivity
Guidelines usage and label ambiguity	Guidelines always available
Interface traits that mattered	Eliminate distractions
Why the task felt personally important	Addresses broader concept - HEI

Table 5: Mapping of themes to design requirements

6 Responsible Research

6.1 Reproducibility

All materials used for this study are publicly available. Section 3 explains thoroughly how we carried out the experiment while the source code used for the web interface and log analysis can be found in this GitHub repository.

6.2 Transparency of LLM Usage

Large language models were used to assist the writing process of this paper. Moreover, they were utilized to develop the web interface and Python log analysis script, as implementing those was not in the core scope of the project. We used GPT-4o and GPT-o3 for revising the flow of the narrative and devising the sections' structures. Every AI-generated prompt has been carefully inspected and fact-checked if needed.

6.3 Ethical Considerations and Impacts

The study was approved by the TU Delft Human Research Ethics Committee. Every participant provided informed consent and could withdraw at any time. They were not exposed to deception or sensitive content. The data collected from the users were stored without any personal identifiers. The goal of the study was to reduce hermeneutical injustice while minimizing any potential risk to the participants.

7 Discussion

7.1 Improved Hermeneutical Justice

Three of the five criteria were met showing that the redesigned interface was more inclusive regarding people with ADHD. Average micro-tasks were completed below the 50 s threshold, distraction rates were negligible (0.15) and the label definitions tooltip was used by the top 3 most accurate annotators. These numbers indicate that the design removed practical barriers like time blindness, long attention windows and guidelines memorization. This makes the participation of our target group in annotation pipelines more accessible. Furthermore, these results improve the gap mentioned by Spiel et al., that most technological systems try to remove ADHD experiences because they are thought to be disruptive [19]. Thus, ADHD perspectives can be better encoded into SFT datasets, which would result in a reduction of generative hermeneutical ignorance [8].

7.2 Implications of Design Choices

Timer as "time prosthesis" Making the remaining time explicit solved time blindness but introduced timer pressure. This is further backed up by a study which shows that disabled individuals feel more anxious due to crowdwork task time constraints [21]. This would suggest that a softer visual representation could be used instead of the timer. Additionally, an optional toggle to show and hide the timer could keep the benefits while lowering the stress.

Cooldown did not help accuracy The cooldown certainly affected the impulsive decision-making of the users. However, it did not solve the problem of label accuracy. As it can be seen from the charts there was a negative correlation between label accuracy and decision change rates. This suggests that frequent label reconsideration might reflect hesitation or uncertainty rather than a reflection of the chosen labels. Thus, an adaptive delay with the addition of a supporting feature, such as limiting the number of decision changes per question, might work better.

Guidelines on demand The results showed that guidelines had a largely positive impact on the label accuracy of the participants. However, having guidelines with the label definitions easily accessible in the interface was not enough. Participants were still reluctant to check them. Therefore, introducing an always visible one-line hint next to the answer could improve the results.

7.3 Limitations

The findings in this study are exploratory and should be interpreted with caution. The sample included seven participants with ADHD but they were not tested against a neurotypical control group. This weakens the statistical power and generalizability of the results. Furthermore, the length of the annotation task session was limited to 5 minutes, which prevented observing longer-term fatigue effects. Finally, helpfulness labels were evaluated against ground truth defined by the researcher, rather than group consensus, which could have introduced bias in the results.

Finally, the results show that introducing small, cognitive-aware tweaks already broadens who can participate in annotation processes [5; 15]. The challenge remains accuracy and adaptive cooldown and guidance seem like the next logical step.

8 Conclusion and Future Work

The goal of this study was to reinforce hermeneutical justice by redesigning current supervised fine-tuning annotation practices for people with ADHD. We answered the questions of what the disparities are between current task structures and specific ADHD traits, and how we can improve and fill these gaps. Therefore, we designed a more inclusive annotation tool to improve the interactions of our target group, which had the aim to better represent them, in turn reducing hermeneutical injustice. The findings support this because the time-aware single-focus micro-tasks helped mitigate sustained attention load and time blindness. Moreover, the availability of the guidelines proved helpful as the performance of the annotators increased, yet they were underused. The 2-second timeout window reduced impulsive submissions but it did not improve accuracy. The qualitative themes confirmed that layout design strongly shapes how users experience the task.

We provide the following suggestions for future work. First, the design of the interface could still be improved further, as discussed previously, to lower the labeling error rates. Potential enhancements include an adaptive cooldown period that changes the delay based on user behavior. Moreover, we propose showing inline hints that contain the label definition next to each option. Second, the study should be replicated with a larger sample size and compared to a neurotypical control group. Furthermore, the framework should be extended to other neurodivergent communities. Finally, the proposed redesigned workflow could be embedded into SFT platforms to measure the end-to-end impact on model representation.

References

[1] R. Matt Alderson, Lisa J. Kasper, Kristen L. Hudec, and Connor H. G. Patros. Attention-deficit/hyperactivity disorder (ADHD) and working memory in adults: a meta-analytic review. *Neuropsychology*, 27(3):287–302, May 2013.

[2] EM Bender, T Gebru, A McMillan-Major, S Shmitchell, and ACM. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *University of Washington*, pages 610–623, 2021.

[3] Virginia Braun, , and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, January 2006. Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa.

[4] Miranda Fricker. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press, 2007.

[5] Elizabeth Garrison, Dalvir Singh, Donald Hantula, Matt Tincani, John Nosek, Sungsoo Ray Hong, Eduard Dragut, and Slobodan Vucetic. Understanding the experience of neurodivergent workers in image and text

data annotation. *Computers in Human Behavior Reports*, 11:100318, August 2023.

[6] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online, 2020. Association for Computational Linguistics.

[7] Shivani Kapania, Alex S Taylor, and Ding Wang. A hunt for the Snark: Annotator Diversity in Data Practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15, Hamburg Germany, April 2023. ACM.

[8] Jackie Kay, Atoosa Kasirzadeh, and Shakir Mohamed. Epistemic Injustice in Generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):684–697, October 2024. Section: Full Archival Papers.

[9] Robin Maria Francisca Kenter, Adrian Schønning, and Yavuz Inal. Internet-Delivered Self-help for Adults With ADHD (MyADHD): Usability Study. *JMIR Formative Research*, 6(10):e37137, October 2022.

[10] C. Kern, S. Eckman, J. Beck, R. Chew, B. Ma, and F. Kreuter. Annotation Sensitivity: Training Data Collection Methods Affect Model Performance. pages 14874–14886, 2023.

[11] R. Krishna, K. Hata, S. Chen, J. Kravitz, D.A. Shamma, L. Fei-Fei, and M.S. Bernstein. Embracing error to enable rapid crowdsourcing. pages 3167–3179, 2016.

[12] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. OpenAssistant Conversations - Democratizing Large Language Model Alignment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 47669–47681. Curran Associates, Inc., 2023.

[13] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research Methods in Human-Computer Interaction*. Wiley, Chichester, 2010.

[14] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C.L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. volume 35, 2022.

[15] D. Paulino, J. Ferreira, A. Correia, J. Ribeiro, A. Netto, J. Barroso, and H. Paredes. Modelling Aspects of Cognitive Personalization in Microtask Design: Feasibility and Reproducibility Study with Neurodivergent People. pages 1552–1558, 2024.

[16] Walter Roberts, Richard Milich, and Mark T. Fillmore. Constraints on information processing capacity in adults with ADHD. *Neuropsychology*, 26(6):695–703, November 2012.

[17] Alexander Schneidt, Aiste Jusyte, Karsten Rauss, and Michael Schönenberg. Distraction by salient stimuli in adults with attention-deficit/hyperactivity disorder: Evidence for the role of task difficulty in bottom-up and top-down processing. *Cortex*, 101:206–220, April 2018.

[18] Tobias Sonne, Paul Marshall, Carsten Obel, Per Hove Thomsen, and Kaj Grønbæk. An assistive technology design framework for ADHD. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction*, OzCHI '16, pages 60–70, New York, NY, USA, November 2016. Association for Computing Machinery.

[19] Katta Spiel, Eva Hornecker, Rua Mae Williams, and Judith Good. ADHD and Technology Research – Investigated by Neurodivergent Readers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–21, New York, NY, USA, April 2022. Association for Computing Machinery.

[20] Lara Tucha, Anselm B. M. Fuermaier, Janneke Koerts, Rieka Buggenthin, Steffen Aschenbrenner, Matthias Weisbrod, Johannes Thome, Klaus W. Lange, and Oliver Tucha. Sustained attention in adult ADHD: time-on-task effects of various measures of attention. *Journal of Neural Transmission (Vienna, Austria: 1996)*, 124(Suppl 1):39–53, February 2017.

[21] Stephen Uzor, Jason T. Jacques, John J Dudley, and Per Ola Kristensson. Investigating the Accessibility of Crowdwork Tasks on Mechanical Turk. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–14, New York, NY, USA, May 2021. Association for Computing Machinery.

[22] Simon Weissenberger, Katerina Schonova, Pascal Büttiker, Raffaele Fazio, Martina Vnukova, George B. Stefano, and Radek Ptacek. Time Perception is a Focal Symptom of Attention-Deficit/Hyperactivity Disorder in Adults. *Medical Science Monitor : International Medical Journal of Experimental and Clinical Research*, 27:e933766–1–e933766–5, July 2021.

A Codebook

Code	Definition	Participants
Timer Pressure	The participant felt rushed because a visible countdown was running	P2 P3 P4 P6 P7
Time Blindness	The participant lost track of elapsed time until they noticed the timer	P2
Manageable Task	Task demand felt comfortable within cognitive limits	P1 P2 P5
Split Attention	The participant switched focus between labeling the answers and the timer	P4
Value-Based Perfectionism	The participant wanted to perform well on the task because the topic of ADHD was important to them	P6
Overthinking	The participant spent excessive time thinking about the answers	P7
Cooldown-Enabled Revision	2-second delay allowed the participant to change their decision about a label	P1 P5 P7
Cooldown Annoyance	Delay felt longer than expected	P2
Cooldown Irrelevant	Delay did not influence decisions	P3 P4 P6
Slow Decision-Making	The participant self-reported slow pace independent of the cooldown	P6
Guidelines Used	The participant opened the guidelines while doing the task	P1 P2 P5
Label Definitions Memorized	The participant heard the label definitions once and did not need to check them during the task	P3 P4
Label Meaning Confusion	The participant was confused about the label definition or the label criteria	P1 P2 P5 P7
Information Overload	Initial instructions felt overwhelming	P6
Timer Distraction	Timer itself pulled attention away	P4
Layout Distraction	The placement of the questions and choice of fonts made the interface hard to read	P4 P6
Content Distraction	The participant was distracted by thinking about the topic of the questions and the correctness of the answers	P7
No Distraction	No distractions were reported	P1 P2 P3 P5

Table 6: Inductive codes derived from the post-task survey responses.