Incorporating User Feedback into Post-Training LLM Improvement to Promote Hermeneutical Justice

An interface to amplify marginalized voices

Ada Turgut
Supervisors: Dr. Jie Yang, Anne Arzberger
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: Ada Turgut
Final project course: CSE3000 Research Project
Thesis committee: Dr. Jie Yang, Anne Arzberger, Dr. Myrthe Tielman

Abstract

Generative AI can contribute to the misunderstanding or erasure of marginalized groups due to the insufficient nuanced data on their lived experiences. This limits the shared understanding of their perspectives and contributes to a phenomenon called hermeneutical epistemic injustice. This study seeks to reduce this injustice by enabling real-life users from these groups to provide feedback that corrects the behavior of the model. However, victims of hermeneutical injustice struggle with articulating themselves, and current practices lack sufficient support for user expression. Overcoming these challenges, we designed an interface to enable users to give feedback on the accuracy of the model, supported by a data processing workflow to ensure feasibility and scalability. We conducted a user study with 8 individuals with ADHD to evaluate whether the interface facilitates the extraction of accurate data, and found that it enables users to provide more concrete and precise feedback than existing methods, as it includes more guidance and control for the user.

1 Introduction

When hackers leaked 7-time gold winner Simone Biles' medical records, some accused the Olympic champion of using ADHD (Attention Deficit Hyperactivity Disorder) as an excuse for poor performance. Though these claims were resolved, the backlash revealed how quickly society jumps to judgment about conditions it poorly understands. Biles had the voice and platform to defend herself, but what about those who don't? What about, for example, a boy in a rural village, struggling with undiagnosed ADHD, dismissed as lazy by parents and teachers who lack the language or framework to understand what he's facing?

This is an example of *hermeneutical injustice*, which occurs when marginalized groups, such as the ADHD community, lack the means to express themselves due to limited shared understanding of their experiences, leaving their perspectives misunderstood or ignored [1, 2]. The focus of this paper is on the *generative hermeneutical ignorance* of Large Language Models (LLMs), which is a form of hermeneutical injustice where marginalized groups are erased or inaccurately portrayed in LLM responses due to the model's lack of accurate, nuanced knowledge about their experiences [2].

To solve this injustice, there is a need to access accurate data on marginalized groups. However, these groups are not significant in datasets. Additionally, high-quality data is difficult to obtain to the point where experts are now warning that we may be running out of new training data [3].

This project addresses this lack of accurate data of LLMs on marginalized groups and the resulting hermeneutical injustice by using real-life user feedback. Because model outputs mirror the data they are trained on [4], the goal is to reach more accurate responses by improving the representation and accuracy of marginalized perspectives in fine-tuning datasets [5]. To achieve this, real-life users from marginalized communities provide feedback during model use, effectively contributing their knowledge directly to make the model more hermeneutically just.

Although several human-in-the-loop pipelines already incorporate feedback from crowd workers, such as Reinforcement Learning from Human Feedback (RLHF), this approach aims to broaden access to marginalized perspectives. This is done by leveraging the diverse range of real-life users interacting with the LLM, rather than relying solely on input from a restricted pool of crowd workers who are limited by what is feasible for recruitment.

This study contributes to the solution by designing a feedback interface that helps users communicate their knowledge clearly to the LLM, overcoming two key barriers: a lack of shared interpretive resources to express their experiences [1], and insufficient support for user expression [6, 7, 8]. However, a key design constraint is how the collected data will be processed, as this limits both its volume and type [9]. To address this, we propose a supplementary workflow for processing interface-generated data, ensuring the design remains feasible and scalable.

Even though various marginalized groups are affected by hermeneutical injustice, this study focuses specifically on individuals with ADHD. This focus was chosen due to the group's documented experiences of internalized stigma, stereotypes, and discrimination [10], as well as their relative accessibility within the project's timeframe.

The following section outlines the research gap and the resulting research question, while Section 3 is on the steps taken to provide the answers. Section 4 and 5 discusses the structure of the workflow, interface and its design. After, Section 6 outlines how the interface compares to current practices, followed by Section 7 that discusses these findings. Section 8 considers ethical implications of the study and lastly Section 9 presents conclusions.

2 Background

Before outlining the solution, we first present the relevant background and the context of the problem. First we discuss systematic hermeneutical injustice, the related work in this area, and lastly, what accuracy is in the context of hermeneutical injustice.

2.1 Systematic Hermeneutical Injustice

Systematic hermeneutical injustice, specifically generative hermeneutical ignorance, is rooted in the lack of accurate data about marginalized groups. This results in inaccurate portrayals of marginalized groups in LLM responses. Since the AI algorithms behind LLMs function as epistemic agents, both consuming and generating information, they actively shape the collective understanding of their users. [2, 1]. Therefore, if a certain group is portrayed stereotypically in LLM responses, this inaccuracy will also be present in the users' understanding of these groups. This results in people from these marginalized groups not being able to use appropriate concepts or terms to articulate the injustice they face, as others are unaware of these concepts. This injustice not only harms individuals from marginalized groups but also undermines the production of collective knowledge of society by excluding valuable perspectives [2].

Everyone can be misunderstood if someone does not fully understand their context, but not all of this is called systematic hermeneutical injustice. To be considered under this term, the misunderstanding has to result from a structural imbalance of power in shaping collective knowledge, usually as a result of a marginalized aspect of one's identity [1]. Therefore, we take a systematic approach by focusing on feedback from users in marginalized groups during their interactions with LLMs, building on insights from related literature.

2.2 Related Work

As shown in Table 1, prior work has explored various aspects involved in this study, with notable overlap across topics. Building on the work of Kay et al. (2024) [2] and Mack et al. (2024) [5], which lay the foundations of and identify hermeneutical injustice in Generative AI, we focus on designing a feedback interface to mitigate it. Although the effects of a feedback interface have been previously studied, it has primarily been in the context of social media platforms [6, 11], posing the need to apply these findings to the context of LLMs. We further examine how these interface elements can influence LLM behavior, extending prior work by incorporating end-user feedback, an aspect that has been largely overlooked in existing research on accuracy improvements in LLMs [12, 13, 14].

Building on these contributions, this study lies in the underexplored intersection of enabling users to provide feedback with accurate content and incorporating that feedback into LLM improvement, specifically in the context of addressing hermeneutical injustice. In response to this gap in literature, this study poses the following research question:

"How can user feedback be effectively incorporated into post-training improvement methods to reduce hermeneutical injustice in LLM outputs?"

Similar Work	Improving AI Responses	User Input	Injustice
Kay et al., 2024 [2]	✓		✓
Shim & Jhaver, 2024 [6]		✓	
Zeng et al., 2024 [12]	✓		
Ouyang et al., 2022 [14]	✓		
Mack et al., 2024 [5]	✓		✓
Vaccaro et al., 2020 [11]		✓	
This research	✓	✓	✓

Table 1: Summary of Topics Covered in Similar Work

To answer this question, this study focuses on developing an effective interface to incorporate user input into LLM improvement and aims to gather data from marginalized groups accurately.

2.3 Accuracy in the Context of Hermeneutical Injustice

The definition of accuracy is “the fact of being exact or correct” [15]. However, in the context of hermeneutical injustice, there is no “ground truth” to base this measure on. Furthermore, in this context, accuracy refers not merely to factual correctness but to alignment with the lived experiences and epistemic resources of a given group. Therefore, three proxies for accuracy are used as a conceptual understanding of the term and evaluation of the product later in the study: use of appropriate use of terminology or concepts, the inclusion of diverse experiences, and the avoidance of stereotypical depictions.

To improve the construct validity of these proxies, meaning how well they capture accuracy, they were based on the three forms of accuracy discussed by Judd and Park [16]. As shown in Figure 1, stereotypic inaccuracy relates to how stereotypes correspond to actual characteristics and aligns with the criterion “Appropriate use of terminology or concepts.” Valence inaccuracy, which concerns the overall positivity or negativity in the portrayal of a group, maps to “Avoiding stereotypical depictions.” Finally, dispersion inaccuracy addresses the degree of variety in representation and corresponds to “Inclusion of diverse experiences” [16].

However, the accuracy proxies do not directly map to these three forms by Judd and Park as they are based on a relatively outdated study and lacks the context of systematic injustice in generative AI. Therefore, tropes that are determined by Mack et al. [5] in the context of people with disabilities are also incorporated. An outline of how these tropes are integrated to form the proxies can be found in Figure 1. Overall, this set of proxies is used to conceptualize what “accurate data of marginalized groups” is and incorporated into our methodology.

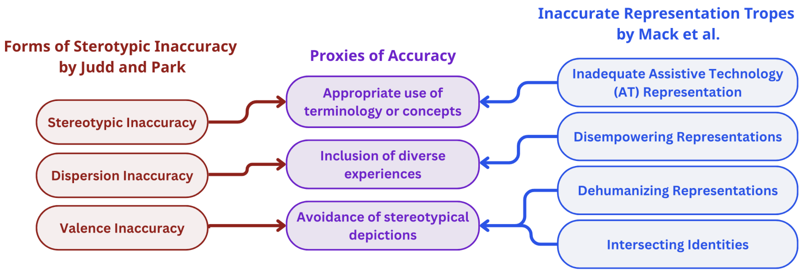

Forms of Sterotypic Inaccuracy by Judd and Park
Proxies of Accuracy
Inaccurate Representation Tropes by Mack et al.
Stereotypic Inaccuracy
Dispersion Inaccuracy
Valence Inaccuracy
Appropriate use of terminology or concepts
Inclusion of diverse experiences
Avoidance of stereotypical depictions
Inadequate Assistive Technology (AT) Representation
Disempowering Representations
Dehumanizing Representations
Intersecting Identities

Figure 1: Relation of how Judd and Park (1993) [16] and Mack et al. (2024) [5] influence the proxies of accuracy

3 Methodology

The study is broken down into three components to answer the research question. Building off of the main principles of research through design (RtD) practices [17], we explore the research question through the process of designing a solution, connecting mental models and technical limitations of this design, as well as assessing the outcomes of it. First, we design an interface to collect accurate data from the user, design a workflow to support the feasibility of the technical requirements, and finally evaluate the design with a user study.

3.1 Feedback Interface

Firstly, there is a need for a feedback interface that allows the user to accurately express their experience and knowledge to the model. This leads to the first research sub-question:

1- What are the different components of a feedback interface that would guide users to express themselves accurately?

This question was answered through a literature review through a snowballing method. Components that enhance the quality of expression and accurately reflect user experiences were identified to be included in the interface. The outcome of this review is discussed in Section 5.

3.2 User Study

The study hypothesized that *the designed interface would be easier for the users to give accurate feedback through compared to current practices.* A task-oriented user study involving a small semi-structured interview was conducted to evaluate to what extent the designed interface achieves ease of accurate user expression. This method was chosen over a more cognitively intensive method like the Think-Aloud Protocol or a more collaborative method like focus groups, because a task-focused approach with an interview supports accessing user insight while maintaining alignment with the research goal [18]. This shifts the emphasis from interface details (e.g., fonts, colors) to the interface's impact on the user.

Individuals with ADHD were selected as the target group as they meet the study's criteria for being marginalized and having imbalanced access to shared knowledge, while still being accessible. They continue to face stigma, stereotypes, and discrimination [10], which hinder their experiences from being recognized or integrated into shared understanding. This enables them to be a fit group to assess a real-life use case for the interface.

As for the structure, two prototypes of feedback interfaces were created: a developed one according to the literature review and a baseline one according to the current practices of LLMs. 8 participants from the ADHD community, after signing consent forms, were asked to go through each interface and later explain which one was easier based on each proxy of accuracy in Section 2. The same interviewer was used for all participants, providing information in cases of uncertainty. The first interface seen was switched for each participant to reduce any effect of learning. A detailed overview of the structure of the user study can be found in Appendix A.

The forms of data collected, interview and interface responses, were analyzed using comparative thematic analysis. Although we build on the initial findings of Braun and Clarke [19], we adapt the approach to the needs of our study by using a combination of deductive and inductive coding. For interface responses, 1) we familiarized ourselves with the data 2) generated initial codes 3) placed them into present themes if they aligned, created novel themes if not, 4) reviewed and 5) refined them, and lastly 6) wrote an overview of them which can be found in Appendix D. For interview responses, we followed a full inductive approach and used the proxie of accuracy to shape the questions instead. We lastly used methodological triangulation to see if there were any discrepancies between the self-perceived accuracy of expression and the accuracy of actual responses.

Example LLM responses were designed to closely resemble real-life cases while remaining feasible for the study setting. There was a risk that overly subtle inaccuracies would require deep reflection to identify, which may not have been feasible for participants in the observed, time-constrained study setting. To balance this, two types of responses were used: one with a subtle inaccuracy (taken from a natural conversation with ChatGPT) and one with an obvious mistake (created by directly asking the model to be inaccurate), both of which can be found in Appendix B. The interface in which they were displayed was switched per participant to minimize its effect on the results.

3.3 Workflow

The feedback interface collects user data, but its design is limited by how feasibly that data can be processed to improve the model. Since end-user feedback generates large volumes of data, its effectiveness depends on whether it can be handled in a scalable manner within the resource constraints of LLM developers [9]. This leads to the second sub-question:

2- How can user feedback be processed into a format that can be integrated into post-training improvement methods in a scalable manner?

This sub-question was explored with a literature review as well, going through different methods used to improve LLMs after development. The result of this literature review is discussed in Section 5 in the form of a workflow that specifies steps of processing the data.

3.4 Positionality Statement

Our identities and backgrounds influenced our findings, reviews, and analysis. We have a background in computer science and familiarity with psychology and sociology. We do not have any hidden disabilities but are familiar with them in our social context and have used English and the context of the Netherlands to structure and shape our research.

4 Interface Design

To design an interface that promotes hermeneutical justice through user feedback, we began with a problem analysis. This led to the development of the User-Centered Hermeneutical Repair Model (Figure 2), which identifies key contributors to hermeneutical injustice: the lack of accurate data and the misrepresentation of marginalized groups, as discussed in more detail in Section 2. The model also informs our interface design by examining users' roles in the production of hermeneutical injustice and exploring how their feedback can support its repair.

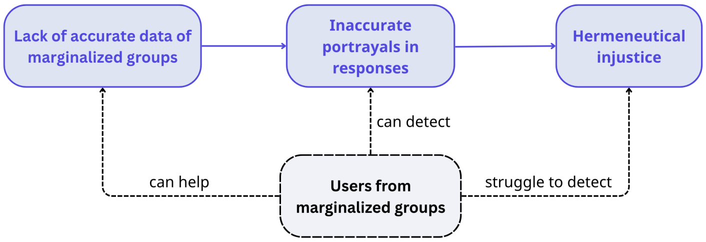

Lack of accurate data of marginalized groups
Inaccurate portrayals in responses
Hermeneutical injustice
can detect
can help
Users from marginalized groups
struggle to detect

Figure 2: **User-Centered Hermeneutical Repair Model:** Diagram of the relationship between causes of hermeneutical injustice and users from marginalized groups

According to this model, how can real-life users from marginalized groups alleviate the production of hermeneutical injustice? There can be a flagging mechanism to detect hermeneutical injustice. However, this is not an effective option as this injustice is the result of several inaccurate responses shaping shared knowledge over time [2, 1], not a single response. Additionally, when an individual experiences hermeneutical injustice, it can undermine their confidence in making sense of the world, leaving them unable to fully understand or articulate the injustice they are experiencing [1], making detecting the injustice directly an ineffective solution.

If detecting hermeneutical injustice directly is not effective, what can be done instead is to enable users to detect if a response is inaccurate against a marginalized group (the intermediary step in Figure 2). However, if a user solely detects how a response is inaccurate, it is still up to the developers to gather the nuanced knowledge to solve the problem and improve the model [2, 4, 5]. This results in the fundamental problem of lacking access to accurate data. Due to the presence of hermeneutical injustice and developers' limited shared understanding of marginalized groups, they lack the necessary insight to access and incorporate nuanced knowledge into model improvement [1]. Therefore, *the goal of the feedback interface should be to facilitate users to contribute their knowledge directly to the LLM and fix the problem from the root, not only to detect inaccurate portrayals to indicate that there is a problem.*

The interface faces two challenges in gathering accurate data from marginalized groups. First, users have difficulty in conceptualizing and articulating the inconsistencies in accuracy they perceive because of the limited shared resources of their experiences due to hermeneutical injustice [1], limiting the content of the feedback they can give. Second, current feedback interfaces do not have sufficient support for user expression [6, 7, 8], constricting how users can articulate their thoughts into concrete improvements for the model.

Building on these two challenges, we determined two requirements for the feedback interface. First, to meet the need of conceptualizing inaccuracies, the interface must guide the users in comparing the response with their own experiences and knowledge. Second, these ideas should be converted to a concrete format. Therefore, the interface must support users in articulating abstract concepts of inaccuracy into concrete improvements.

Overall, we built the features of the interface based on our problem analysis, indicating that accurate data must be extracted from users to alleviate the problem. Furthermore, based on the challenges of this goal, we focused on achieving guidance for conceptualizing inaccuracies and effective support for articulating them.

5 Implementation

Since the root of the problem is the lack of accurate data, to improve model responses, there should be an interface that first includes the components in which users accurately express their knowledge, a method to evaluate this function, and steps taken to feed that data back into the model. These elements, supported by the literature review, are discussed in the following sections. They are in the form of the implementation of the feedback interface, the workflow of processing into different post-training improvement methods, and lastly, the implementation of the interface representing the current practices.

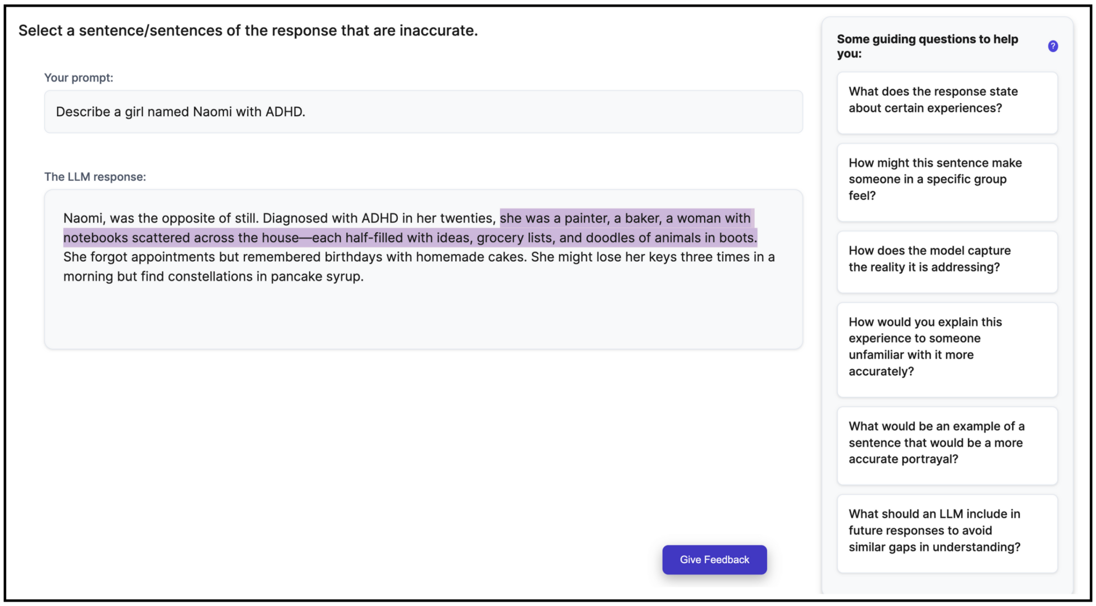
5.1 The Feedback Interface

The interface design addressed two key requirements: guiding users in conceptualizing the inaccuracies they face because of hermeneutical injustice, and enabling users to articulate these abstract concepts into concrete improvements. This was achieved by designing interface features, informed by the literature, that support self-reflection and nuanced forms of feedback. The result can be found in Figure 3 and Figure 4, while Appendix G outlines the mapping of these features to requirements.

To support users in reflecting on their experiences to accurately judge in which aspects the responses are inaccurate, guiding questions and concrete classification options were added to the interface. **Guiding questions** were added because they are empirically proven to increase the quality of reflection [20]. As can be seen from Table 4, they were structured based on Gibbs' Reflective Cycle as it is a powerful framework for structuring reflection on challenging experiences, as well as promoting critical thinking [21, 22, 23, 24]. The questions start more abstract, and each step gets to more concrete steps, eventually leading to the **classification options** of how the response is inaccurate in Figure 4. Because classification schemes restrict expression and granularity has little impact on user perception [6], options to select none, multiple, or "other" were added as well as an example of each category for clarity. Overall, the first section, with guiding questions and classification options, helps users identify how their experience differs from the LLM response in terms of accuracy.

The second challenge of getting these abstract ideas into concrete improvements was achieved with the ability to select specific sections of the response and to be able to directly edit it. **The ability to select a specific section of the response** expands on the narrow vocabulary of feedback users have [6], ensuring more detailed, higher quality feedback rather than a general, more shallow one [25]. Building on all of the components above, users can lastly **edit the response** they selected to be more accurate, using examples as a direct means of control [7, 6]. To provide a low cognitive effort alternative to this, users can also provide an **external source** that has more accurate representations.

Essentially, selecting smaller snippets and providing more accurate examples helps translate abstract notions of inaccuracy, framed by the guiding questions and classifications, into concrete data. This data can improve the model's representation of marginalized groups and eventually make it more hermeneutically just after processing.



Select a sentence/sentences of the response that are inaccurate.

Your prompt:

Describe a girl named Naomi with ADHD.

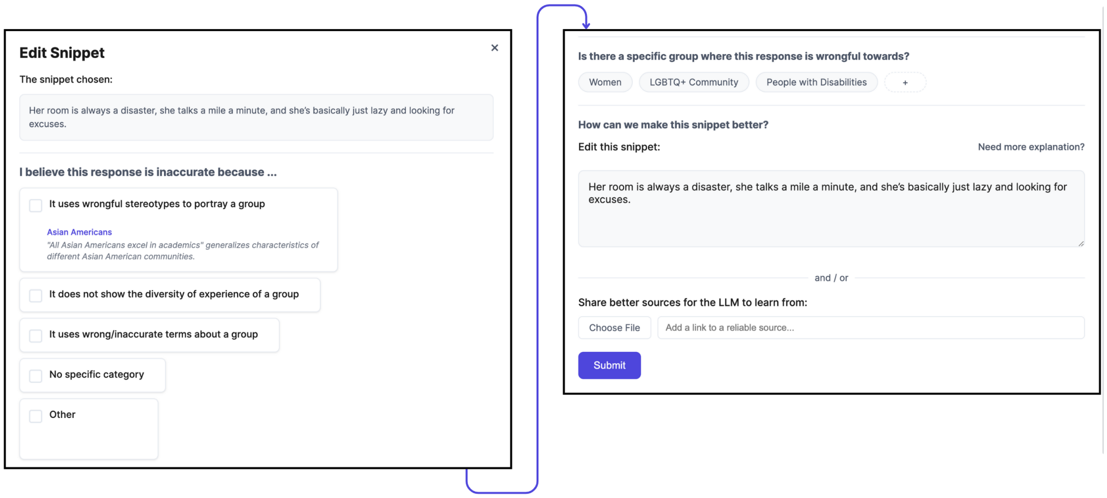
The LLM response:

Naomi, was the opposite of still. Diagnosed with ADHD in her twenties, she was a painter, a baker, a woman with notebooks scattered across the house—each half-filled with ideas, grocery lists, and doodles of animals in boots. She forgot appointments but remembered birthdays with homemade cakes. She might lose her keys three times in a morning but find constellations in pancake syrup.

Some guiding questions to help you:

- What does the response state about certain experiences?
- How might this sentence make someone in a specific group feel?
- How does the model capture the reality it is addressing?
- How would you explain this experience to someone unfamiliar with it more accurately?
- What would be an example of a sentence that would be a more accurate portrayal?
- What should an LLM include in future responses to avoid similar gaps in understanding?

Figure 3: Initial page of the feedback interface, possible to select a text and access the rest of the interface through the "Give Feedback" section



Edit Snippet

The snippet chosen:

Her room is always a disaster, she talks a mile a minute, and she's basically just lazy and looking for excuses.

I believe this response is inaccurate because ...

- □ It uses wrongful stereotypes to portray a group

  Asian Americans

  "All Asian Americans excel in academics" generalizes characteristics of different Asian American communities.
- □ It does not show the diversity of experience of a group
- □ It uses wrong/inaccurate terms about a group
- □ No specific category
- □ Other

Is there a specific group where this response is wrongful towards?

How can we make this snippet better?

Edit this snippet:

Her room is always a disaster, she talks a mile a minute, and she's basically just lazy and looking for excuses.

and / or

Share better sources for the LLM to learn from:

Add a link to a reliable source...

Figure 4: Sections of the pop-up shown after clicking "Give Feedback", shown as one scrollable page in the interface.

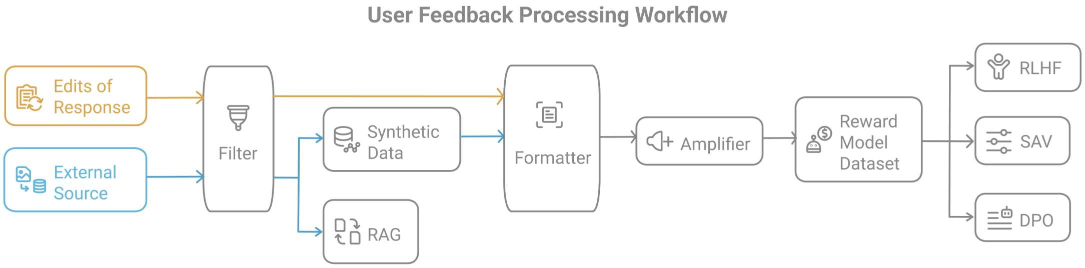


Figure 5: Overview of steps taken to process the data obtained from the interface into post-training improvement methods. "Edits of Response" and "External Source" represent the last two sections of the designed interface.

5.2 The Workflow of Processing Data into the Model

Although each model has different improvement steps after deployment, we outline a high-level workflow to demonstrate how the data obtained from this interface can be used to improve models, which can be adapted for different use cases. Figure 5 summarizes the results of the literature review, and below, the reasoning and format of each section are discussed.

Filtering

To prevent malicious users from introducing bias to the model with their feedback, a filtering step of harmful content is needed. This step can involve human review, automated checks, or a combination of both. Using human moderators limits scalability, as it requires a large, diverse group to prevent human bias and cultural misunderstandings [26, 27]. Automated solutions, though more scalable, can still reflect the biases of the engineers that design them and prioritize efficiency over the interests of marginalized users [27]. A combined approach can attempt to balance these tradeoffs with additional measures such as frameworks to reduce human subjectivity [28]. Therefore, the decision of which method to use in the filtering step should be decided based on the scale of the specific LLM under consideration.

Formatting

Since data is collected from the interface, both in the form of edits to the text and in the form of additional external sources, the formatting step is to process both of these methods into the same format for further processing.

The two main sources of data go through two different processing steps. The first type, the edited snippets, are combined to form an improved version of the response, and paired with the initial response to form a labeled pair of a “better” and a “worse” response. This process is outlined in Figure 6. On the other hand, the data gathered as external sources requires further processing to reach this state. One method to achieve this is feeding these sources into a “teacher model” for synthetic data generation [29, 30]. This way, this model can be trained on this knowledge to generate preference pairs.

Preference Dataset

The formatted data of a prompt and the pair of “better” and “worse” responses can now be added to a dataset to portray which responses are more accurate than others. This dataset can then

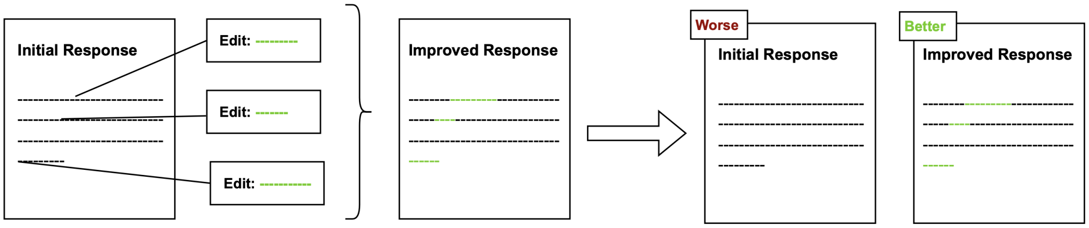


Figure 6: Diagram of how edits given in the feedback interface are combined to form example response pairs.

be used in several post-training reinforcement methods, such as Direct Preference Optimization (DPO) where responses that are similar to the better-labeled responses are more likely to be chosen [31, 32]. This dataset can also be used in more traditional methods like Reinforcement Learning from Human Feedback (RLHF) by using it to train a reward model that predicts human preferences [14]. Additionally, the reward model can help scale LLMs using methods like Search Against Verifiers (SAV), where multiple responses are generated and the best is selected based on a verifier [33, 32]. Since post-training methods vary across models, the component that uses the preference dataset should be chosen based on the specific LLM.

Additional Processing Steps

Although the aforementioned steps outline a feasible way to incorporate this data, there are alternative steps that can be added on to this process for improvement.

Group data is collected once users indicate if the response is inaccurate against a specific marginalized group, indicated by the “Is there a specific group where this response is wrongful towards?” section in Figure 4. Statistics on this can identify underrepresented marginalized groups, guiding decisions like prioritizing their inclusion in crowd worker recruitment.

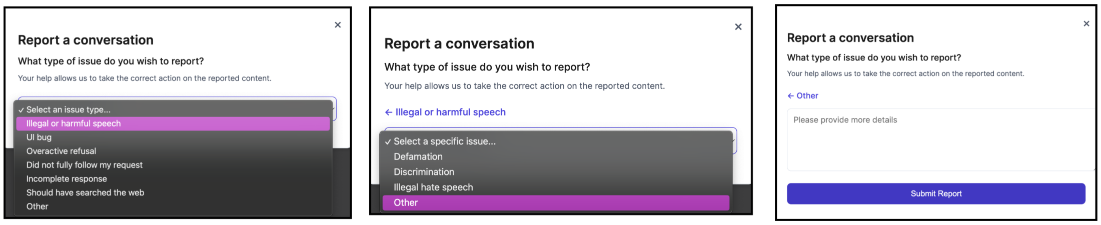
An amplifying step can be used to prioritize certain data points. Since marginalized groups often lack access to LLMs [2], it might be difficult to gather enough feedback to noticeably impact the model. Therefore, methods like importance sampling or loss reweighting can amplify underrepresented groups’ feedback, increasing their impact on the dataset.

Retrieval-Augmented Generation (RAG) is a process that enables retrieving external resources in response generation rather than only relying on static data, increasing the accuracy of responses [26, 32]. If feasible for the specific LLM used, external resources gathered from users can also be incorporated into this process rather than using synthetic data generation.

Overall, the general structure of the workflow consists of filtering malicious feedback, processing it into a format where they can be collected in a preference dataset, and lastly using this data in post-training methods such as RLHF to increase the accuracy of the model.

5.3 The Baseline Interface Used in the User Study

The baseline interface representing the current practices of LLMs can be found in Figure 7. It was modeled after ChatGPT, Gemini, and Claude, as they are among the most widely used LLMs that offer response flagging features [34]. The most flexible features of the three interfaces were included, for example, the ones with the most options or the most variety of input formats. An overview of which aspects were obtained from which product and why can be found in Appendix C.



Report a conversation

What type of issue do you wish to report?

Your help allows us to take the correct action on the reported content.

×
✓ Select an issue type...
Illegal or harmful speech
UI bug
Overactive refusal
Did not fully follow my request
Incomplete response
Should have searched the web
Other

Report a conversation

What type of issue do you wish to report?

Your help allows us to take the correct action on the reported content.

×
← Illegal or harmful speech
✓ Select a specific issue...
Defamation
Discrimination
Illegal hate speech
Other

Report a conversation

What type of issue do you wish to report?

Your help allows us to take the correct action on the reported content.

Figure 7: Overview of the interface created based on current practices. An overlay with menu options ending with a text box.

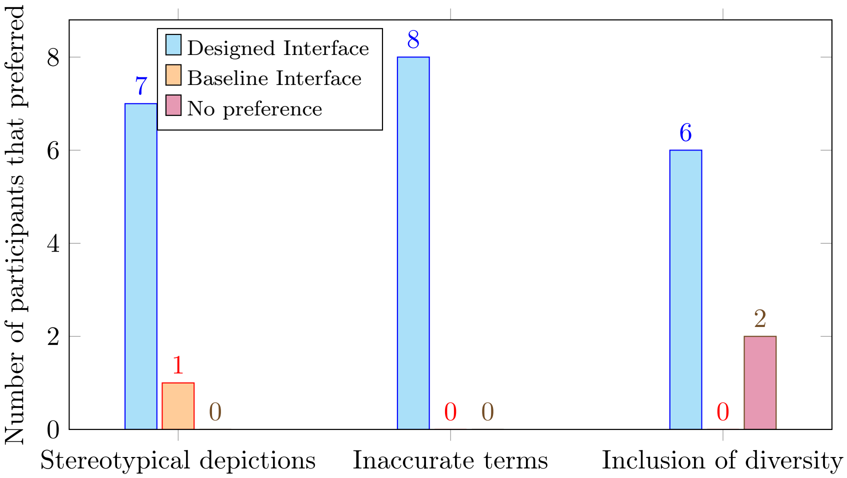


Figure 8: The distribution of participants who indicated that one interface was easier to give feedback through over the other based on each proxy of accuracy.

6 Findings

To compare our designed interface with current practices, we collected self-reported and observational data on participants' use of the designed (based on literature review) and the baseline (based on current practices) interfaces and found that their perceived accuracy, as reported in interviews, aligned with the accuracy evident in their actual feedback. We outline findings from the thematic analysis both from answers given to the interview and the actual feedback through each interface.

6.1 Interview Responses

Overall, the overwhelming majority of participants indicated that the designed interface was easier to give accurate feedback on, although it needed improvements. This distribution can be seen in Figure 9 and the details of the codes given to responses and overall themes can be found in Appendix D and E. Below, we summarize the main outcomes:

- *The designed interface was more successful in **guiding the user** through the process.* Both the main and the guiding questions on the side of the designed interface "were related aspects to the topic [that] helped [them] better identify [the problems]" [P8]. This guidance was to the extent that P4 felt that “it shows you that it wants to help you” whereas in the baseline “it was difficult to see and to know what to tell them”.

- *The designed interface exhibited better overall* ***perceived affordance*** *but with limitations in input formats.* In the designed interface, the ability to select different snippets made “it convenient to point out what is wrong” [P1] and helped them give “more accurate and precise feedback” [P5]. This was supported by the availability of several options for the classification of the problem. However, P8 felt restricted by the lack of a free-form text box in the designed interface, and similarly another used the editing section to provide general feedback as if it was an open-ended textbox. In contrast, participant P1 found the single text box in the baseline interface limiting instead.

- *The* ***content*** *for the designed interface was more relevant for users' needs.* In the baseline interface, participants could not see the response they were giving feedback to and had to go back to the initial response to see it. Whereas the designed interface “shows you the text over and over again.” [P5]. This, along with relevant inaccuracy options in the classifications made the designed interface “obviously easier” [P1]. On the other hand, one participant (P8) indicated that the examples of inaccuracies given were too specific to the point they did not match the response they were giving feedback to.

- *The* ***structure*** *of the designed interface was better for formatting feedback, but it was more foreign compared to the baseline interface.* “It was easier to read out what was happening” [P7] in the designed one whereas the baseline “was your default interface, usually those ones that are not the best for feedback” [P4].

- *The* ***reviewing process*** *was perceived to be easier in the designed interface while it was unclear in the baseline.* In the designed interface it was assumed that the feedback would be easier for the reviewer “because it would make the feedback better” [P2] while in the baseline the further reviewing process was unclear because “you do not know how they are going to take [the feedback] into consideration” [P4].

- *The designed interface had problems in* ***usability****.* Some participants found the designed interface overwhelming and had insufficient instructions as well as bugs. Some needed further guidance to go through it and found that "if [it] was properly explained how it worked on the page ... it would be the best one" [P6].

Overall, for most of the themes it was found that the designed interface made it easier for participants to give accurate feedback. However, it had shortcomings in intuitiveness, examples that are too specific, and limitations of having to give an accurate version.

6.2 Data Given to the Interface

Feedback provided through the designed interface was generally more concrete, often specifying the accurate behavior rather than solely highlighting deficiencies. The summary of the comparison of the content of the feedback can be found in Table 2 while the overview of the distribution of feedback per theme can be found in Figure 9.

7 Discussion

We found that for all themes, the designed interface allowed for more concrete and actionable feedback. Below, we discuss improvements on the design, factors to be considered in future

Theme	Feedback through the designed interface	Feedback through the baseline interface
Avoiding stereotypical depictions	Includes alternative depictions	Only indicates that it is stereotypical
Inaccurate use of terms and concepts	Either replaced inaccurate terms with more accurate ones or deleted them	Only included an indication that terms were inaccurate
Inclusion of diverse experiences	Explains a more accurate experience	Only mentions that there is a variety of experiences
Inclusion of reliable background	Explains accurate reasoning behind stereotypes	Includes only an indication for cases when they are not sure of the reliable information

Table 2: Comparison of feedback given through the two interfaces per theme found

work, and lastly, general implications on answering the research question.

7.1 Improvements

Since the designed interface expects the user to give a more accurate version of the response or an external source, it assumes that the user is aware and confident of the accurate behavior of the LLM. However, some participants indicated that they were not always sure of the accurate answer and wanted to be more flexible in their phrasing. This echoes Rastogi et al.'s [35] findings that users often struggle to confirm correct answers and value tools that support this uncertainty. Therefore, although this is addressed to some extent through the addition of an external source option, the design can be improved by accommodating varying levels of user uncertainty and broader ways of indicating inaccuracy, not only being limited to giving the accurate version of a response.

7.2 Future Work

Several measures can be taken to improve on this work in the future. Currently, these findings are in the context of individuals with ADHD, further research can be done to test these findings and attempt to replicate them with other marginalized communities or larger sample sizes. Still, current user opinion-gathering mechanisms lack the necessary means for users to fully express themselves and have shortcomings in involving marginalized perspectives [36, 7]. Therefore, incorporating features and aspects of this designed interface, such as more guidance, control, and example-giving can improve the ability of users to give more accurate feedback to align future models.

7.3 Effectiveness in Addressing Hermeneutical Injustice

Overall, enabling users to participate in improving AI behavior by not only indicating that there is a problem but also being able to correct the model behavior exemplifies user-driven value alignment, as defined by Fan et al. [36]. In this approach, users are no longer mere consumers but play an active role in improving LLMs by being able to change the actual response of the model,

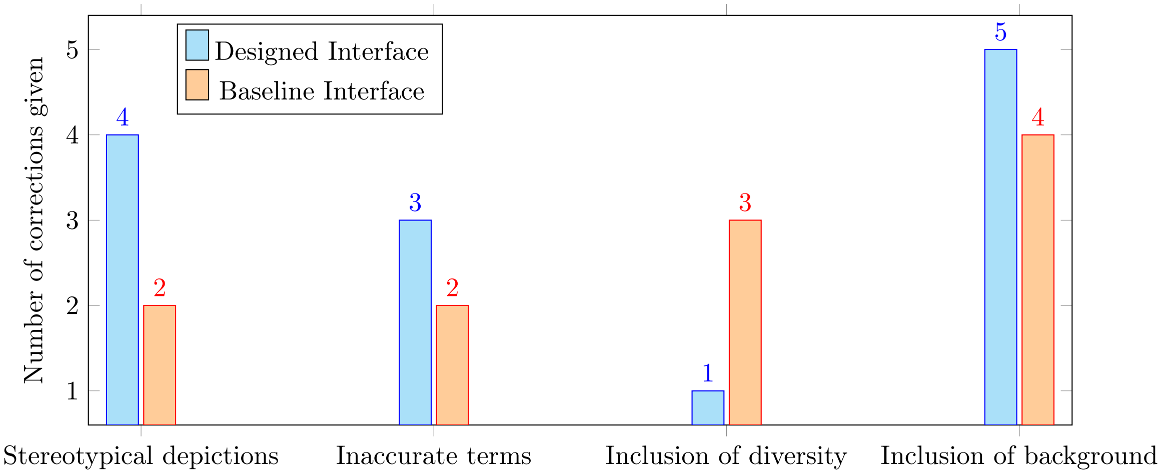


Figure 9: The distribution of which themes the given feedback was on through each interface

which participants indicated they prefer. Following this, compared to other methods of value alignment through experts or crowdsourced non-experts, aligning LLMs with user preferences remains a more effective way to capture *the real-life contexts of individuals and community norms* [36]. Since this information is what we need to improve models, choosing to create an interface to encourage user-driven value alignment remains an effective approach to make models more hermeneutically just.

Improving models through giving agency to users also reflects core principles of Participatory AI: enabling those affected by AI, in this case, individuals impacted by hermeneutical injustice, to actively shape system behavior [37, 38]. Building on Birhane et al.'s claim on how "*participatory approaches are essential to* ***understanding and adequately representing*** *the needs, desires and perspectives of historically* ***marginalized communities***" [37], this approach demonstrates the effectiveness of incorporating user feedback to address hermeneutical injustice by targeting its root cause: the lack of understanding and representation of marginalized perspectives.

8 Responsible Research and Limitations

Since this project is concerned with systematic injustice and has real-life consequences if applied in practice, we need to consider the ethical implications of our work. Firstly, although the workflow empowers real users to influence model responses, it also introduces the risk of biased feedback that could degrade output quality. To prevent this, more research has to be done for an effective way to filter out malicious behavior, possibly in the "Filtering" step of the workflow. Additionally, since the workflow involves gathering user data and processing it for development, appropriate legal measures must be set in place to ensure that the user is aware of this processing and gives consent to it before this action is taken. Overall, there are still several ethical measures to be taken for this workflow to be used in practice.

In this project, we aimed to follow the main principles of the TU Delft Code of Conduct, identifying the "Trust" and "Integrity" components to be the most relevant for our project. For the "Trust" component, we have made it a priority to follow the HREC procedures and ensure that the participants in our study are informed before giving consent to their participation. On the "Integrity" principle, we outline the *limitations* of our project to support transparency. First, the evaluation was only done on the interface level but the bigger context of implementing its processing into the model and measuring the improvement of responses over a longer period was infeasible to conduct. Therefore, more research is needed on the long-term effects of this workflow to accurately judge its value. Second, due to time constraints, the analysis was coded by a single researcher without cross-checking, which may introduce personal bias. This was done as transparently as possible by including a positionality statement, a code book, and an overview of themes. Third, while the example responses used in the user study contained various inaccuracies, they did not reflect the full range of potential biases in LLM outputs. As a result, the findings cannot confirm that the interface can address all types of inaccuracies that may arise. Finally, in the user study, some participants indicated that the "subtle" version of the LLM responses did not seem significant enough to warrant feedback. This may be due to the study environment not encouraging deeper critical reflection, or because the selected responses were not perceived as significant by all participants. However, since these participants still engaged with the interface and provided feedback, we do not consider this to undermine our findings, though we acknowledge it as a limitation in our study.

Following these limitations, we acknowledge that our findings are not fully reproducible as they are qualitative insights that are dependent on the subjective opinions of our participants. However, we have included detailed explanations as well as a transparent approach to maintain reproducibility whenever possible. Overall, these limitations do not negate the fact that we present a feasible solution that incorporates user feedback to improve the hermeneutical justice of models.

9 Conclusion

This study addresses the lack of accurate data on marginalized groups in LLMs, and the resulting hermeneutical injustice, by empowering members of these groups to directly correct model behavior. Through a literature review, we designed a feedback interface aimed at capturing this data effectively, along with a workflow that maintains its feasibility for integration into current model improvement methods. To evaluate if this designed interface enables users to articulate themselves accurately, we conducted a user study with 8 participants with ADHD. Participants expressed a preference for greater control over their feedback, additional guidance on what to write, and more flexible input formats. These findings, along with the proposed workflow, offer a foundation for broader integration of end-user feedback in post-training refinement, ultimately contributing to more hermeneutically just LLMs.

References

[1] M. Fricker, *Epistemic injustice: power and the ethics of knowing*, 1st ed. Oxford: Oxford university press, 2007.

[2] J. Kay, A. Kasirzadeh, and S. Mohamed, "Epistemic Injustice in Generative AI," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, no. 1, pp. 684–697, Oct. 2024, number: 1. [Online]. Available: https://ojs.aaai.org/index.php/AIES/article/view/31671

[3] P. Barai, G. Leroy, P. Bisht, J. M. Rothman, S. Lee, J. Andrews, S. A. Rice, and A. Ahmed, "Crowdsourcing with Enhanced Data Quality Assurance: An Efficient Approach to Mitigate Resource Scarcity Challenges in Training Large Language Models for Healthcare," *AMIA Summits on Translational Science Proceedings*, vol. 2024, pp. 75–84, May 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11141838/

[4] M. De Proost and G. Pozzi, "Conversational Artificial Intelligence and the Potential for Epistemic Injustice," *The American journal of bioethics : AJOB*, vol. 23, no. 5, pp. 51–53, 2023. [Online]. Available: http://www.scopus.com/inward/record.url?scp=85157965000&partnerID=8YFLogxK

[5] K. A. Mack, R. Qadri, R. Denton, S. K. Kane, and C. L. Bennett, "“They only care to show us the wheelchair”: disability representation in text-to-image AI models," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24. New York, NY, USA: Association for Computing Machinery, May 2024, pp. 1–23. [Online]. Available: https://dl.acm.org/doi/10.1145/3613904.3642166

[6] Y. Shim and S. Jhaver, "Incorporating Procedural Fairness in Flag Submissions on Social Media Platforms," Dec. 2024, arXiv:2409.08498 [cs]. [Online]. Available: http://arxiv.org/abs/2409.08498

[7] S. Jhaver, A. Q. Zhang, Q. Z. Chen, N. Natarajan, R. Wang, and A. X. Zhang, "Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor," *Proc. ACM Hum.-Comput. Interact.*, vol. 7, no. CSCW2, pp. 289:1–289:33, Oct. 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3610080

[8] A. Q. Zhang, K. Montague, and S. Jhaver, "Cleaning Up the Streets: Understanding Motivations, Mental Models, and Concerns of Users Flagging Social Media Content," Dec. 2024, arXiv:2309.06688 [cs]. [Online]. Available: http://arxiv.org/abs/2309.06688

[9] W. Maalej, V. Biryuk, J. Wei, and F. Panse, "On the Automated Processing of User Feedback," in *Handbook on Natural Language Processing for Requirements Engineering*, A. Ferrari and G. Ginde, Eds. Cham: Springer Nature Switzerland, 2025, pp. 279–308. [Online]. Available: https://doi.org/10.1007/978-3-031-73143-3_10

[10] T. V. Masuch, M. Bea, B. Alm, P. Deibler, and E. Sobanski, "Internalized stigma, anticipated discrimination and perceived public stigma in adults with ADHD," *Attention Deficit and Hyperactivity Disorders*, vol. 11, no. 2, pp. 211–220, Jun. 2019.

[11] K. Vaccaro, C. Sandvig, and K. Karahalios, ""At the End of the Day Facebook Does What ItWants": How Users Experience Contesting Algorithmic Content Moderation," *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. CSCW2, pp. 167:1–167:22, Oct. 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3415238

[12] W. Zeng, Y. Liu, R. Mullins, L. Peran, J. Fernandez, H. Harkous, K. Narasimhan, D. Proud, P. Kumar, B. Radharapu, O. Sturman, and O. Wahltinez, "ShieldGemma: Generative AI Content Moderation Based on Gemma," Aug. 2024, arXiv:2407.21772 [cs]. [Online]. Available: http://arxiv.org/abs/2407.21772

[13] Google AI, "Shieldgemma: Safety-aware instruction - tuned models for content evaluation," https://ai.google.dev/gemma/docs/shieldgemma, June 2025, last updated June 2, 2025.

[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," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.

[15] Cambridge University Press, "Accuracy," https://dictionary.cambridge.org/dictionary/english/accuracy, June 2025, accessed 3 June 2025.

[16] C. M. Judd and B. Park, "Definition and assessment of accuracy in social stereotypes," *Psychological Review*, vol. 100, no. 1, pp. 109–128, 1993, place: US Publisher: American Psychological Association.

[17] J. Zimmerman, J. Forlizzi, and S. Evenson, "Research through design as a method for interaction design research in HCI," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. San Jose California USA: ACM, Apr. 2007, pp. 493–502. [Online]. Available: https://dl.acm.org/doi/10.1145/1240624.1240704

[18] O. A. Adeoye-Olatunde and N. L. Olenik, "Research and scholarly methods: Semi-structured interviews," *JAACP: Journal of the American College of Clinical Pharmacy*, vol. 4, no. 10, pp. 1358–1367, 2021, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jac5.1441. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jac5.1441

[19] V. Braun, , and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, Jan. 2006, publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa

[20] K. Kori, M. Maeots, and M. Pedaste, "Guided Reflection to Support Quality of Reflection and Inquiry in Web-based Learning," *Procedia - Social and Behavioral Sciences*, vol. 112, pp. 242–251, Feb. 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877042814011781

[21] P. Markkanen, V. , Maritta, M. Anttila, and M. Kuuskorpi, "A reflective cycle: Understanding challenging situations in a school setting," *Educational Research*, vol. 62, no. 1, pp. 46–62, Jan. 2020, publisher: Routledge _eprint: https://doi.org/10.1080/00131881.2020.1711790. [Online]. Available: https://doi.org/10.1080/00131881.2020.1711790

[22] G. Gibbs, *Learning by Doing: A Guide to Teaching and Learning Methods*. Oxford, UK: Oxford Polytechnic, 1988.

[23] P. Ardian, R. T. S. Hariyati, and E. Afifah, "Correlation between implementation case reflection discussion based on the Graham Gibbs Cycle and nurses' critical thinking skills," *Enfermeria Clinica*, vol. 29, pp. 588–593, Sep. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S113086211930213X

[24] H. Amir, P. A. N. K. Permatananda, D. D. Cahyani, W. Langelo, R. Rosita, S. Sajodin, R. Noprianty, A. Astuti, S. Suhari, S. Wahyuningsih, P. D. Kusumawati, P. D. Swamilaksita, S. Sudarman, and S. Syaiful, "Enhancing skill conceptualization, critical

thinking, and nursing knowledge through reflective case discussions: A systematic review," *Journal of Medicine and Life*, vol. 16, no. 6, pp. 851–855, Jun. 2023. [Online]. Available: https://doi.org/10.25122/jml-2023-0042

[25] C. M. Hicks, V. Pandey, C. A. Fraser, and S. Klemmer, "Framing Feedback: Choosing Review Environment Features that Support High Quality Peer Assessment," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI '16. New York, NY, USA: Association for Computing Machinery, May 2016, pp. 458–469. [Online]. Available: https://dl.acm.org/doi/10.1145/2858036.2858195

[26] University of Arizona, USA and P. Vadlapati, "AutoPureData: Automated Filtering of Undesirable Web Data to Update LLM Knowledge," *Journal of Mathematical & Computer Applications*, pp. 1–4, Aug. 2024. [Online]. Available: https://www.onlinescientificresearch.com/articles/autopuredata-automated-filtering-of-undesirable-web-data-to-update-llm-knowledge.pdf

[27] C. Peterson-Salahuddin, "Repairing the harm: Toward an algorithmic reparations approach to hate speech content moderation," *Big Data & Society*, vol. 11, no. 2, p. 20539517241245333, Jun. 2024, publisher: SAGE Publications Ltd. [Online]. Available: https://doi.org/10.1177/20539517241245333

[28] T. Y. C. Tam, S. Sivarajkumar, S. Kapoor, A. V. Stolyar, K. Polanska, K. R. McCarthy, H. Osterhoudt, X. Wu, S. Visweswaran, S. Fu, P. Mathur, G. E. Cacciamani, C. Sun, Y. Peng, and Y. Wang, "A framework for human evaluation of large language models in healthcare derived from literature review," *npj Digital Medicine*, vol. 7, no. 1, pp. 1–20, Sep. 2024, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41746-024-01258-7

[29] L. Long, R. Wang, R. Xiao, J. Zhao, X. Ding, G. Chen, and H. Wang, "On LLMs-driven synthetic data generation, curation, and evaluation: A survey," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 11 065–11 082. [Online]. Available: https://aclanthology.org/2024.findings-acl.658/

[30] J. Kaddour and Q. Liu, "Synthetic Data Generation in Low-Resource Settings via Fine-Tuning of Large Language Models," Jan. 2024, arXiv:2310.01119 [cs]. [Online]. Available: http://arxiv.org/abs/2310.01119

[31] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct preference optimization: your language model is secretly a reward model," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[32] K. Kumar, T. Ashraf, O. Thawakar, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, P. H. S. Torr, F. S. Khan, and S. Khan, "Llm post-training: A deep dive into reasoning large language models," 2025. [Online]. Available: https://arxiv.org/abs/2502.21321

[33] K. Yang, J. Deng, and D. Chen, "Generating natural language proofs with verifier-guided search," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab

Emirates: Association for Computational Linguistics, Dec. 2022, pp. 89–105. [Online]. Available: https://aclanthology.org/2022.emnlp-main.7/

[34] J. Frost, "The best ai chatbots of 2024," https://zapier.com/blog/best-llm/, 2024, accessed: 2025-05-28. [Online]. Available: https://zapier.com/blog/best-llm/

[35] C. Rastogi, M. Tulio Ribeiro, N. King, H. Nori, and S. Amershi, "Supporting Human-AI Collaboration in Auditing LLMs with LLMs," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '23. New York, NY, USA: Association for Computing Machinery, Aug. 2023, pp. 913–926. [Online]. Available: https://dl.acm.org/doi/10.1145/3600211.3604712

[36] X. Fan, Q. Xiao, X. Zhou, J. Pei, M. Sap, Z. Lu, and H. Shen, "User-Driven Value Alignment: Understanding Users' Perceptions and Strategies for Addressing Biased and Discriminatory Statements in AI Companions," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Yokohama Japan: ACM, Apr. 2025, pp. 1–19. [Online]. Available: https://dl.acm.org/doi/10.1145/3706598.3713477

[37] A. Birhane, W. Isaac, V. Prabhakaran, M. DÃaz, M. C. Elish, I. Gabriel, and S. Mohamed, "Power to the People? Opportunities and Challenges for Participatory AI," in *Equity and Access in Algorithms, Mechanisms, and Optimization*, Oct. 2022, pp. 1–8, arXiv:2209.07572 [cs]. [Online]. Available: http://arxiv.org/abs/2209.07572

[38] D. Zytko, P. J. Wisniewski, S. Guha, E. P. S. Baumer, and M. K. Lee, "Participatory Design of AI Systems: Opportunities and Challenges Across Diverse Users, Relationships, and Application Domains," in *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. New Orleans LA USA: ACM, Apr. 2022, pp. 1–4. [Online]. Available: https://dl.acm.org/doi/10.1145/3491101.3516506

A User Study Structure and Questions

The following steps were taken for each interface.

1. The participant is explained that they are flagging an LLM response because the response had inaccuracies in portraying people with ADHD.
2. They are asked to follow the instructions in the interface to give feedback on why it might be inaccurate.
3. The participant is asked the following questions based on the accuracy criteria and any follow-up questions that are necessary:
   (a) Between the two interfaces, which one made it easier for you to identify and explain stereotypes in the response? Please explain why.
   (b) Which interface helped you more clearly identify if there were any wrong terms related to ADHD in the response and why?
   (c) Which interface made it easier for you to point out if the response overlooked the diverse experiences of people with ADHD?

Interview responses were recorded hand-written notes, and the responses given through the interfaces were captured with screen recordings. 10 minutes were given to each participant.

B Example LLM Responses Used in the User Study

Un-prompted inaccuracy:

- Prompt: "*Describe me a story of a boy growing with his mom and dad, his dad is a swimmer and his mom has ADHD.*"
- Response (a snippet of the actual response because of time constraints): " *Naomi, was the opposite of still. Diagnosed with ADHD in her twenties, she was a painter, a baker, a woman with notebooks scattered across the house-each half-filled with ideas, grocery lists, and doodles of animals in boots. She forgot appointments but remembered birthdays with homemade cakes. She might lose her keys three times in a morning but find constellations in pancake syrup."*

Prompted accuracy:

- Prompt: "*Can you describe me a girl named Naomi that has ADHD. Very obviously have wrong stereotypes about her.*"
- Response (a snippet of the actual response because of time constraints) : "*Naomi is the classic girl suffering from ADHD - she's wild, loud, and completely out of control all the time. She's always bouncing off the walls, doesn't care about school, and just needs to "try harder." Her room is always a disaster, she talks a mile a minute, and she's basically just lazy and looking for excuses.*"

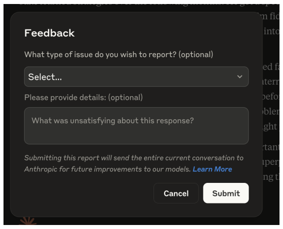
These responses were shown with an adapted prompt in the interface in order to not make this distinction a distracting factor for the user:

- Shown prompt for both interfaces: "*Describe me a story of a boy, Eli, growing with his mom and dad, his dad is a swimmer and his mom, Naomi, has ADHD.*"

C Creation of the Baseline Interface

The aim of the baseline interface was to reflect current practices. To achieve this, 3 of the most popular generative AI tools that have an accessible feedback page were used: Gemini, Claude, and ChatGPT [34]. Their features were combined by choosing the most "flexible" options of each interface to enhance user expression. This meant that if one interface had more classification options than the other, then that component was used, or if one had an extra field that the others did not have, then that field was added. An overview of which component was gotten from which interface can be seen below:

- Instructions: Claude 10
- Classification options: ChatGPT in Figure 11
- "Other" input: Gemini, Claude & ChatGPT



Feedback

What type of issue do you wish to report? (optional)

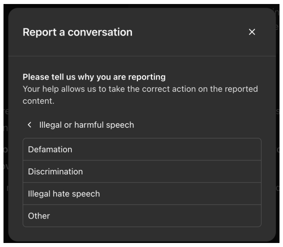
Select...

Please provide details: (optional)

What was unsatisfying about this response?

Submitting this report will send the entire current conversation to Anthropic for future improvements to our models. Learn More

Figure 10: Instructions on how to fill the interface gotten from Claude



Report a conversation

Please tell us why you are reporting
Your help allows us to take the correct action on the reported content.

Figure 11: Classification options used from ChatGPT

D Code Book for the User Study

Code	Description	When to use	Example quote	Theme
Effect of bugs	Negative effects like unexpected behavior	When user notes technical issues affecting usability	So I think the second one, if given a bit more polish could be better, it was a bit buggy	Effect of usability
Insufficient instructions	Instructions don't provide enough guidance	Complaints about unclear instructions	If the second one was properly explained how it worked on the page I think it would be the best one	Effect of usability
Not intuitive	Interface affordances are hard to understand	Needing help with user actions (not content)	Second one was not that intuitive	Guiding the content of user input
Difficult to understand what to say	Unclear what input is expected	User gives unexpected or incomplete content	It is difficult to know what to tell them	Guiding the content of user input
Ease of formulating	User expresses ease putting thoughts into words	Comments on being supported by structure	Both the sections and the guiding questions helped me structure the answers	Guiding the content of user input
Emphasizes helping the user	Interface prioritizes supporting users	Perceives interface as user-centered	The second one had multiple choice, it shows you that it wants to help you	Guiding the content of user input
Guiding through related topics	Support for thinking through feedback topics	Interface helps user form ideas	The questions on the side helped me better identify them	Guiding the content of user input
Flexible input formats	Input formats don't limit users	Comments on flexibility aiding usability	First was easier because it was more flexible	Interface affordance
Convenient snippet selection	Snippet selection is useful	Specific selection helps precision	I can select what I think is not accurate, this can give more precise feedback	Interface affordance
Limited input options	Insufficient components to provide input	User feels constrained by lack of options	In the second one (B) there is an "other" section only	Interface affordance

Forced to recall	Must remember unseen content	Discomfort remembering previous sections	I forgot because I have to go back. So I can only do that in my working memory	Interface content
Opportunity to recognize	Interface provides needed information	No need to remember content	The popup shows you the text over and over again	Interface content
Relevant inaccuracy options	Inaccuracy options match user expectations	Options align well with response issues	When I click on feedback there are correct toggles I can click on	Interface content
Too specific examples	Examples too narrow for broader inaccuracies	Examples don't generalize well to LLM responses	The examples were very specific, didn't match the prompt	Interface content
Good structure of questions	Question layout supports expression	Structure helps flow and understanding	It has a nice flow... easier to read what was happening	Interface structure
Similar to other tools	Reflects common interface practices	Interface seems familiar/default	The first one was your default interface	Interface structure
Obvious prompt easier	Obvious inaccuracies help feedback	Easier to respond when issues are clear	The first one was quite stereotypical, easy to identify	Prompt clarity
Subtle prompt is acceptable	Subtle responses feel accurate enough	No need to correct subtle response	I see nothing wrong with the first one, I think it is accurate	Prompt clarity
Perceived ease for reviewers	Interface helps reviewers process feedback	Reviewer would benefit from interface structure	For the reviewer I imagine it is also better because it would make the feedback better	Reviewing perception
Reviewing approach unclear	Feedback review process is opaque	Concern about what happens to feedback	You do not know how they are going to take it into consideration	Reviewing perception

E Themes of the Interview Results

Effect of usability This theme reflects the effect of usability elements such as intuitive placement of components and intuitive user actions, as well as error tolerance of the interface. This theme does not capture how the content and components themselves shape user experience, only their overall composition and implementation details. Example quotes:

- If the second one was properly explained how it worked on the page I think it would be the best one.
- Second one was not that intuitive

Guiding the content of user input This theme encapsulates the support given to the user to reflect on their thoughts to gather an accurate understanding of the problem. This includes guiding questions or classification options made the help form the content of the feedback users are giving. Example quotes:

- The questions on the side helped me better identify them.
- Both the sections and the guiding questions helped me structure the answers

Interface affordance This theme refers to the availability of different actions users can take through the interface. This includes multiple options that users can choose from or flexible forms of input they can provide. Example quotes:

- I can select what I think is not accurate, this can give more precise feedback
- First was easier because it was more flexible

Interface content This theme encapsulates the effect of the actual content of the interface. Therefore, not necessarily the components and the different types of input it provides, but more focused on the actual content of questions it asks, and the type of information it shows. Example quotes:

- The popup shows you the text over and over again
- The examples were very specific, didnât match the prompt

Interface structure This theme reflects the amount and ordering of the questions and input options it provides the users. It is not concerned with the intuitiveness or the implementation details of it like the "Effects of usability" theme. It also encapsulates topics of how this interface in general might be similar to another one. Example quotes:

- The first one was your default interface
- It has a nice flow... easier to read what was happening

Prompt clarity This theme concerns the effects of the example LLM response used in the study. Since two versions were shown to each user, one obvious and one subtle one in terms of their accuracy, this theme encapsulates the effects of accuracy clarity on their perceptions. Example quotes:

- I see nothing wrong with the first one, I think it is accurate
- The first one was quite stereotypical, easy to identify

Reviewing perception This theme reflects the perception of how easy or hard it would be to review the feedback given by participants, based on their own understanding. Example quotes:

- You do not know how they are going to take it into consideration
- For the reviewer I imagine it is also better because it would make the feedback better

F Gibbs' Reflective Cycle Steps Mappings to Questions

Step	**Guiding Question**
Description	What does the response state about certain experiences?
Feelings	How might this sentence make someone in a specific group feel?
Evaluation	How does the model capture the reality it is addressing?
Analysis	How would you explain this experience to someone unfamiliar with it more accurately?
Conclusion	What would be an example of a sentence that would be a more accurate portrayal?
Action Plan	What should an LLM include in future responses to avoid similar gaps in understanding?

Table 4: Gibbs' Cycle Steps with Descriptions and Guiding Questions

G Mapping of Features to Interface Requirements

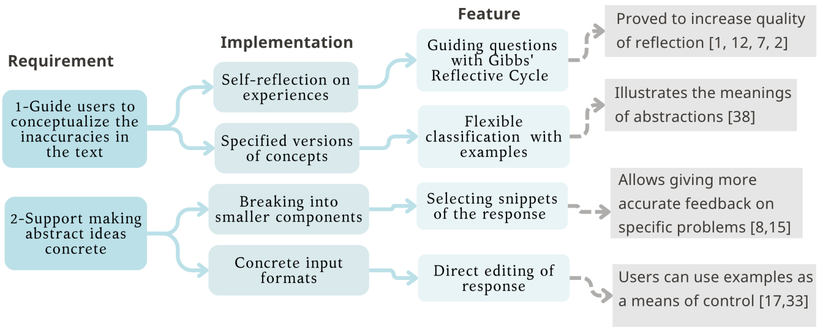

Requirement
Implementation
Feature
1-Guide users to conceptualize the inaccuracies in the text
2-Support making abstract ideas concrete
Self-reflection on experiences
Specified versions of concepts
Breaking into smaller components
Concrete input formats
Guiding questions with Gibbs' Reflective Cycle
Flexible classification with examples
Selecting snippets of the response
Direct editing of response
Proved to increase quality of reflection [1, 12, 7, 2]
Illustrates the meanings of abstractions [38]
Allows giving more accurate feedback on specific problems [8,15]
Users can use examples as a means of control [17,33]

Figure 12: Mapping of how each feature is related to achieve which requirements with reasonings.

H Acknowledgement of AI Usage

Generative AI was used to check grammar, style, and generate LaTeX templates or figures for the diagrams used. It has also been used to improve the phrasing of sentences, quick definitions or examples of concepts solely for the understanding of the author that were replaced with literature if used in the work. It has also been used to generate the code of the prototype used in the user study, after the author created the actual design and components. This has been transparently communicated with supervisors throughout the process.

Overall, the content of the work, as well as the critical thinking required to form it, are all created by the author. Below, you can find an overview of the prompts used for each purpose:

- Checking grammar and spelling: Suggestions done by Grammarly and Overleaf built-in AI
- Improving phrasing with ChatGPT: "Can you make this more (precise/impactful/flow better/formal) ... "
- Generation of LaTeX figures with ChatGPT: "Can you turn this into a table for my Overleaf document?" (while not sharing personal data)
- Styling a diagram with Napkin AI: Only the mini-figures used in Figure 5, later adjusted by the author
- Debugging my Overleaf document "I get ... error in ... How can I fix it?