AUTHOR AND SUBERVISORS

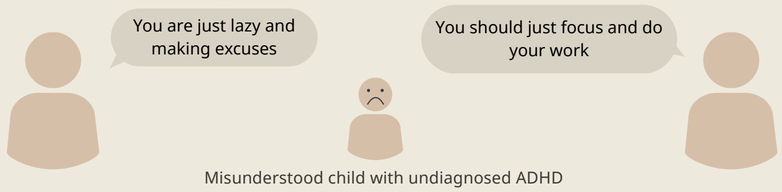
*Author: Ada Turgut*
*aturgut@tudelft.nl*
*Supervisors: Dr. Jie Yang, Anne Arzberger*


# An Interface to Amplify Marginalized Voices

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

REFERENCES

## 01. What is the problem?




Misunderstood child with undiagnosed ADHD

**Hermeneutical injustice:** when marginalized groups *lack the means to express themselves* due *to limited shared understanding of their experiences*, leaving their perspectives misunderstood or ignored [1, 2].

This study focuses on Large Language Models (LLMs) reinforcing this injustice:

**Generative hermeneutical ignorance:** 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 specific marginalized groups* [2].

**Problems with accessing this knowledge:**

- Marginalized groups are already not significant in datasets
- Good quality data is difficult to obtain [3]
- Experts warn that we may soon lack new data for training [3]

There should be way to gather this data. Marginalized users are more present as real-life users compared to crowdworkers → we can use their help!

## 02. Related Work and the Research Gap:

| Similar Work | Improving AI Responses | User Input | Injustice |
|---|---|---|---|
| Kay, Kasirzadeh, & Mohamed, 2024 | ✓ | | ✓ |
| Shim & Jhaver, 2024 | | ✓ | |
| Zeng et al., 2024 | ✓ | | |
| Ouyang et al., 2022 | ✓ | | |
| Mack, Qadri, Denton, Kane, & Bennett, 2024 | ✓ | | ✓ |
| Vaccaro, Sandvig, & Karahalios, 2020 | | ✓ | |
| This research | ✓ | ✓ | ✓ |

Table 1: Summary of Topics Covered in Similar Work

This study covers the research gap in improving AI responses with user input in the context of injustice.

## 03. Methodology

**Research Question**:

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

We focus on:

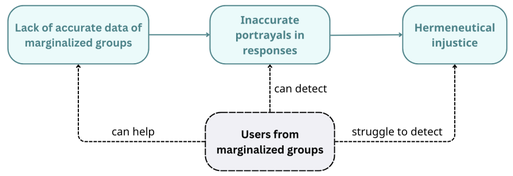
- **An interface** that supports users articulating their experiences accurately

Evaluation → A **user study** to evaluate if the interface enables ease of accurate expression

Support feasibility → A **workflow** for processing this data

Target focus group: Individuals with ADHD → Facing stigma & accessible

## 04. Interface Design




**User-Centered Hermeneutical Repair Model**

If the users only detect inaccurate portrayals:

→ Developers still need to find accurate data to fix it [2, 4, 5]
→ Results in the initial problem of limited access to data [1]
→ Then, the solution is for users to provide accurate data instead of just detecting inaccuracies

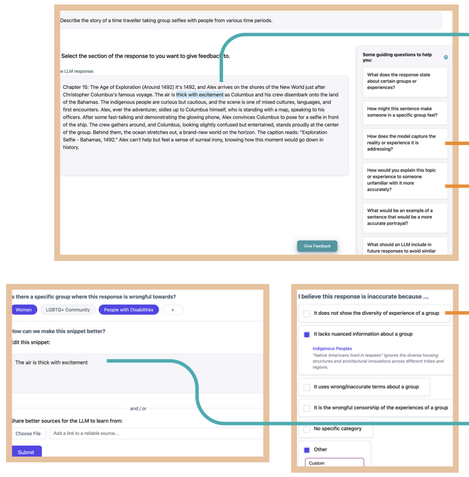
**What are the challenges of users providing accurate data on their experiences?**

**Therefore, the interface should:**

- Difficulty in conceptualizing and articulating the inconsistencies in accuracy due to hermeneutical injustice [1] → **Guide** the users to conceptualize and compare the inaccuracies in the text with their knowledge
- Lack of sufficient support for user expression in current systems [6, 7, 8] → Enable users to **articulate** the abstract concepts of inaccuracy into concrete improvements

## 05. Interface Implementation

The interface was based on requirements from the problem analysis.




**Splitting text into sections**
Allows giving more accurate feedback on specific problems [6, 9] (Articulation)

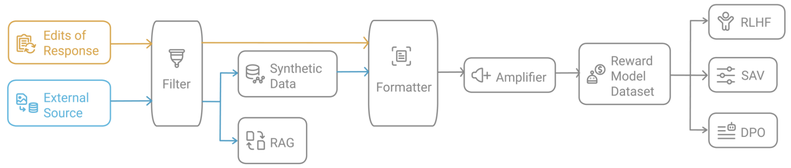
**Guiding questions** (Guiding)
Proved to increase quality of reflection [10]

**Question Structure** (Guiding)
Gibbs' Reflective Cycle is used to formulate as it is a powerful framework for structuring reflections [11, 12, 13, 14]

**Flexible classification with examples** (Guiding)
Illustrates the meanings of abstractions [6]

**Editing a response** (Articulation)
using examples as a means of control [6, 7]

## 06. Workflow Implementation



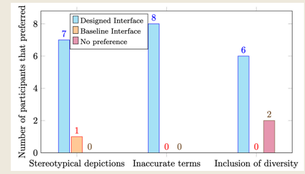

**User Processing Workflow:** How data gotten from the interface can be used in post-training improvement.

## 07. User Study Findings

Does the designed interface make it easier for users to express themselves accurately compared to the current practices?

**What did users think?**

The designed one was "obviously easier" [P1]
except for the cases where the example or the input format was too limiting.




Participant preferences from the interview

**Breaking into components**

"Because I can select what I think is not accurate, **this can give more precise and accurate** feedback. " [P5]

**Guidance through reflections**

"The questions on the side were related aspects to the topic. **These helped me better identify them.**" [P8]

**What did users do?**

**LLM Response:**
"Naomi is the classic girl suffering from ADHD - she's wild, loud, and completely out of control all the time..."

**Correction through the designed interface:**
"She can experience bursts of energy which may make her come off as loud" [P4]

**Input through the baseline interface:**
"The message portrays people with ADHD using harmful stereotypes" [P2]

Feedback provided through the designed interface was more concrete.

## 08. Discussion

Participants were not sure and wanted more flexible inputs. → Users often struggle to confirm correct answers and value tools that support uncertainty [15]. → Need for broader formats of input.

**Then is involving users to this extent even effective?**

**Participatory AI**: "essential to understanding and adequately representing the needs, desires and perspectives of historically marginalized communities" [16].

**User-driven value alignment:** aligning LLMs with user preferences remains a more effective way to capture the real-life contexts of individuals [17].

## 09. Conclusion

To validate this solution, more research is needed to involve other marginalised groups, bigger sample sizes, evaluation of the long-term effects of the workflow, and preventing malicious behavior.

However, the insights from this study show that incorporating more guidance, control, and example-formatted inputs can improve the ability of users to give more accurate feedback and be used to make models more hermeneutically just in the future.