Unheard and Misunderstood: Addressing Injustice in LLMs

How are hermeneutical injustices encoded in Reinforcement Learning from Human Feedback (RLHF) in the context of LLMs?

Ieva Mockaitytė
Supervisor(s): Jie Yang, Anne Arzberger
[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: Ieva Mockaitytė
Final project course: CSE3000 Research Project
Thesis committee: Jie Yang, Anne Arzberger, Myrthe Tielman

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

This study investigates how hermeneutical injustices can become encoded in the Reinforcement Learning from Human Feedback processes used to fine-tune large language models (LLMs). While current research on fairness in LLMs has focused on bias and fairness, there remains a significant gap concerning subtler harms such as hermeneutical injustice. Using adults diagnosed with ADHD as a case study, this research explores how their unique communication and cognitive patterns may be misrepresented or excluded from the RLHF pipeline.

The research adopts a qualitative literature review methodology, focusing specifically on real-world RLHF implementations by AI companies. The RLHF pipeline was divided into stages of human feedback collection, reward modeling, and policy optimization. Then, these stages of the RLHF were analyzed through the lens of hermeneutical injustice using interpretive desiderata: representation, flexibility, and authenticity.

The findings highlight several conceptual risks. Limited annotator diversity and restrictive feedback formats may exclude neurodivergent voices. Reward models can unintentionally suppress atypical expressions, while policy optimization strategies, especially those prone to mode collapse, can erase some communication styles. Overall, the study shows that without deliberate attention to epistemic inclusion, RLHF processes may perpetuate hermeneutical injustices and undermining the epistemic fairness of LLMs.

1 Introduction

The usage of large language models (LLMs) worldwide is increasing, and so is their influence on people's daily lives. Studies show that users tend to overestimate the accuracy of LLM outputs, especially when the responses are more verbose. [1]. This poses a significant risk: if users accept such responses without checking their validity, they may contribute to spreading factually incorrect, harmful, or biased information. Hence, it is crucial to ensure that the LLM outputs reflect factual accuracy, fairness and inclusivity of marginalised groups.

One critical concern is the potential for LLMs to perpetuate hermeneutical injustice, a specific form of epistemic injustice defined by Miranda Fricker (2007) as "the injustice of having some significant area of one's social experience obscured from collective understanding owing to hermeneutical marginalization" [2]. Unlike mere misinformation, hermeneutical injustice involves systemic gaps in understanding marginalized lived experiences. Since LLM training is highly dependent on data reflecting dominant discourses and narratives, experiences of marginalised groups may not be adequately represented.

Currently, research on justice in LLMs focuses on bias and fairness, especially with respect to race, gender, age, and religion [3]. However, the subtler phenomenon of hermeneutical injustice remains underexplored. For example, as of May 2025, the search query "hermeneutical AND injustice* AND LLM*" on Scopus returns zero results. This absence strongly indicates a research gap: while hermeneutical injustice itself is a well-established philosophical concept, its possible manifestation in LLMs remains underexplored.

A good starting point to identify hermeneutical injustices in LLM responses is to look into processes designed to align these responses to human preferences. One of the techniques used in the training of LLMs is Reinforcement Learning from Human Feedback (RLHF). Following the classification proposed in a paper by Casper et al. (2023), we define the RLHF process by three stages: human feedback collection, reward modelling, and policy optimisation. [4], Each of these stages has specific vulnerabilities that could unintentionally marginalize certain experiences and perspectives. To illustrate this concretely, the paper employs a case study of adults with Attention Deficit/Hyperactivity Disorder (ADHD), a group which is often subject to misunderstanding and marginalization due to their unique communication patterns and cognitive processing styles.

Therefore, this paper aims to answer the following research question: **How are hermeneutical injustices encoded in Reinforcement Learning from Human Feedback in the context of LLMs?** First, we will define the concept of hermeneutical injustice, and relate it to LLMs. Next, we will identify and define the core mechanisms and stages of RLHF. Then, we will look into how hermeneutical injustice might be encoded at each of these stages.

The paper is structured as follows. Section 2 provides the necessary background information and related work on hermeneutical injustice and ADHD experiences. The methodology and the precise scope of the project is formalised in Section 3. Section 4 covers the objective findings of the literature survey. The view of the findings through the analytical lens can be found in Section 5. The ethical aspects of this research are covered in Section 6. A discussion and broader context of the results can be found Section 7, and the final conclusions along with future recommendations are provided in Section 8.

2 Background and related work

This section introduces essential background information, defines key concepts clearly, reviews related work, and explicitly identifies the research gap addressed in this study.

2.1 Hermeneutical injustice

Hermeneutical injustice, as conceptualised by Miranda Fricker (2007), refers to a specific form of epistemic injustice where significant experiences of certain social groups are obscured or misunderstood due to gaps in collective interpretive resources. Unlike mere misinformation, this type of injustice is caused by the structural marginalisation of people's perspectives, leading to misunderstandings in dominant discourses. Commonly, this marginalisation can result in individuals or groups being unable to articulate certain aspects of their own experiences. The book *Epistemic Injustice. Power & the Ethics of Knowing* illustrates this well with the example of a homosexual boy growing up in America in the 1950s: the collective consensus referred to homosexuality as "just a stage", "a sickness", or "a sin". The only options for the boy are to either try to challenge such a deeply rooted view on homosexuality, or try to fit himself into these boxes, which are very inaccurate and constrict his sense of identity [2].

While hermeneutical injustice is well-theorised in philosophy, its application to LLMs is understudied. A study by Kay et al. (2023) introduces the concept of *generative algorithmic epistemic injustice*. This framework highlights how LLMs can marginalise groups through biased training data or feedback loops, providing concrete examples of problematic LLM behaviour. However, the focus is more on identifying cases of epistemic injustice - a detailed technical analysis on LLM training stages and is still missing, and the exact methods that may encode hermeneutical injustice are not pinpointed. [5].

2.2 Case study: ADHD

Affecting between 2 and 4 percent of the adult population, ADHD is considered one of the most common psychiatric disorders. [6] Despite this, research shows that it tends to be significantly misdiagnosed due to the inherent comorbidity of the disorder and a tendency for medical professionals to focus on the other coexisting illnesses [7]. Additionally, a significant part of research is focused on childhood diagnosis and early intervention, which further contributes to misunderstood adult experiences. This is a clear example of hermeneutical injustice - due to a prominent gap of insights into ADHD experiences, people do not get the proper diagnosis, and, by extension, proper treatment. [8].

ADHD serves as a suitable case study for this research due to the inherent distinct communication patterns and cognitive processes that deviate notably from neurotypical norms. A 2025 study analysed Reddit communities of neurodivergent people and collected the main use cases of LLMs among neurodivergent people, as well as their main concerns and complaints about these LLMs. This study found that the majority of LLM discussions in the ADHD community express frustration over prompting difficulty and receiving responses different than desired. Additionally, 20% of the discussions brought up complaints about neurotypical biases in received LLM responses. Although more prominent in autism and social anxiety communities, complaints about LLM responses struggling to maintain the authentic voice of the prompter were also noticeable in ADHD community. [9]

Furthermore, studies have identified distinctions between neurotypical individuals and those diagnosed with ADHD, particularly in the ways information is processed and communicated. For analytical clarity, this study categorizes these distinctions into two dimensions: information processing and information conveying.

Information processing

Adults with ADHD often encounter challenges with sustained attention, showing a faster deterioration of focus over time compared to neurotypical individuals [10].

Information conveying

Adults with ADHD tend to use more words to convey the same narrative than individuals without ADHD [11].

2.3 RLHF process

Reinforcement Learning from Human Feedback (RLHF) is a method that has recently emerged as a way to align LLM outputs with human preferences. Casper et al. (2023) outline three main stages involved in RLHF:

1. **Human feedback collection** - in this first step, examples from a base model are taken, with humans providing feedback in the form of preferences between a set number of examples.

2. **Reward modelling** - in the second step, the collected feedback is used to fit a reward model, whose goal is to approximate human preferences.

3. **Policy optimisation** - finally, the base model is fine-tuned using reinforcement learning guided by the trained reward model.

The main focus of this categorisation in the paper is to pinpoint the exact problems and limitations in each of these stages [4]. While this provides a great foundation, the implications of possible encodings of hermeneutical injustice in RLHF are not addressed. Consequently, this research specifically aims to bridge the gap between philosophical analyses of hermeneutical injustice and technical analyses of RLHF, by investigating how RLHF processes could unintentionally encode hermeneutical injustices, particularly concerning adult ADHD experiences.

3 Methodology

This research adopts a semi-structured qualitative literature review methodology to investigate how hermeneutical injustices could potentially be encoded in the RLHF processes used in LLMs. Given the interdisciplinary and conceptual nature of the topic, a qualitative literature review is particularly appropriate, since it allows for an in-depth conceptual exploration and synthesis of the technical nuances within RLHF processes, specifically from the lens of hermeneutical injustice experienced by ADHD adults.

3.1 Literature selection procedure

The literature selection followed a structured and transparent approach to ensure clarity and reproducibility. Given the interpretive and conceptual nature of this investigation, it would be inaccurate to call this a fully systematic review - nevertheless, the selection process was inspired by the PRISMA guideline checklist of 2020.

Databases

The literature was gathered primarily from the following databases and sources:

- Scopus
- Google Scholar
- Official reports from reputable LLM companies, notably OpenAI and Anthropic.

Inclusion criteria

The selected literature had to meet all of the following criteria:

- The focus must be on RLHF as it is implemented in practice within the context of LLMs.
- The paper must provide a detailed description or evaluation of at least one of the three RLHF stages:
  - Human feedback collection
  - Reward modelling
  - Policy optimisation
- The paper must be properly peer-reviewed or published by an established and reputable organisation.

Exclusion criteria

The sources that were explicitly excluded from this research were:

- Papers that did not directly address RLHF processes (for example - general LLM ethics papers, papers that describe other LLM training methods)
- Papers that described purely theoretical RLHF implementations without clear practical application or evidence of use.

These papers were specifically analysed to identify and extract detailed real life applications of the RLHF process in the context of LLMs. The following step was to look at these findings through the lens of hermeneutical injustice of ADHD adults.

3.2 The analytical approach

The analysis of the papers was based on a structured interpretive lens of hermeneutical injustice. The theories and findings of the RLHF papers were used to identify points of intersection between RLHF stages and the risks of hermeneutical injustice relating to individuals with ADHD. We have decided to form a desiderata list after considering the complaints expressed by ADHD users from the previously mentioned study [9], with the aim to cover the entire process of RLHF. After careful phrasing, we have formed the following list of desiderata guiding this analysis:

- **Representation** - does the RLHF method allow for the representation of diverse human experiences and perspectives, including those of marginalised groups?
- **Flexibility** - is the RLHF approach capable of handling a variety of communication and cognitive traits, specifically when they deviate from neurotypical norms?
- **Authenticity** - can the voices and experiences of neurodiverse groups be accurately maintained throughout the RLHF process?

Precisely, the known cognitive and communicative characteristics of the case study group (adults with ADHD) were systematically mapped onto identified technical limitations of the RLHF process. Then, using the previously defined hermeneutical injustice desiderata, each RLHF limitation was evaluated to identify conceptual intersections where such injustices may be encoded.

4 Results

This section outlines the findings from the literature survey. The three stages of the RLHF pipeline are separated into different subsections for clarity.

4.1 RLHF pipeline: Human feedback stage

The human feedback collection stage is the first component of the RLHF process. During this stage, human annotators are asked to evaluate outputs generated by a language model according to specific criteria, with the aim to guide the model towards more helpful, relevant, and correct responses in the future. A study by Kaufmann et al. presents a classification of feedback types used in RLHF. These include: Binary trajectory comparisons, trajectory rankings, state preferences, action preferences, binary critique, scalar feedback, corrections, action advice, implicit feedback, and natural language [12]. However, many of these are not used in practice of finetuning LLMs.

Next, we will address the publicly available information about the RLHF feedback collection process of publicly available LLMs. In order to critically assess the process of human feedback collection, it is important to ask two questions - who was providing feedback and how were they asked to express their preferences. We are looking into the methods of two LLM companies: OpenAI and Anthropic.

OpenAI

- **GPT-3 and InstructGPT**
  The study by Ouyang et al. (2022) provides insights into the pool of annotators whose feedback was used to finetune the GPT-3 model. They report hiring a team of **40 carefully selected contractors** who worked in the labelling process. The paper acknowledges the limitations of such an annotator pool - for example, the group consisted of primarily English speakers, admitting that this group is not an accurate representation of the distribution of people using this LLM. Regarding the feedback types, **a Likert scale of 1-7** was used to evaluate the responses given by GPT-3 [13]. Additionally, the annotators were instructed to provide a **ranking** of the responses from best to worst, including any possible ties [14].

- **Evolution of ChatGPT**
  Barman et al. (2025) published a study detailing the human feedback collection stage of OpenAI, pointing out the differences between the early InstructGPT models and later ChatGPT models. The study mainly notes that after the expansion of OpenAI's user base, the feedback collection has significantly changed. Firstly, **users from 193 countries** were now able to provide feedback while they were using the tool, drastically expanding the diversity. However, the feedback type was less detailed and expressive as the previously used Likert scale - the options only included a possibility of a **thumbs up/down** rating; a chance of the model providing two responses, asking the user to **choose the preferred one**; and in case of a request to regenerate a response, the user could indicate if the new one was **better, worse or similar in quality** [14].

- **GPT-4**
  OpenAI's *GPT-4 Technical Report* confirms that this later model also utilised RLHF for the post-training alignment. Interestingly, the exact methods are not publicly available with the study citing "the competitive landscape and the safety implications" as the reasons. However, this report mentions that **over 50 experts** were hired to test the behaviour of the model, focusing on dangerous topics and jailbreaking. This collected expertise was used for later improvements of GPT-4. Additionally, a few vague mentions of using similar feedback collection techniques that were used on GPT-3 can be found [15].

Anthropic

- **Early models**
  A paper by Bai et al. (2022) details the early work of Anthropic, focusing on the utilisation of RLHF. They describe the human feedback process as follows. The human evaluators, of which there were around 20, would write a prompt or a question, the

model would generate two responses, and then the workers would choose **which of two responses was better**. An opposite red-teaming strategy was also used, in which case the evaluators would have to choose the more harmful response. However, as identified in this paper, **the crowdworkers were all US-based and master-qualified** [16].

- **Anthropic Claude 2 and Constitutional AI**
  A study by Bai et al. (2022) introduced a new method with the aim to eliminate the need for human feedback, refered to as Constitutional AI (CAI). The main idea of this method is to defined a set of guidelines and principles, referred to as the "constitution", which would later be used for the AI to engage in self-critique [17]. Anthropic has stated that the CAI method was used together with RLHF and unsupervised learning in the development of Claude 2 and its previous versions. The human feedback continued being in a **binary preference** format, which was later used to calculate Elo scores. However, no information about changes in crowdworker pools is published [18].

- **Anthropic Claude 3**
  The Claude 3 model card states that **binary preference** format was still used for fine-tuning. However, Claude 3's documentation does not list annotator demographics or any inclusivity efforts. Overall, it can be seen that Claude 3's RLHF was an iterative improvement on Claude 2 [19].

4.2 RLHF pipeline: Reward modelling stage

In this stage, a reward model is trained to predict human-preference ratings for model outputs. The RM converts qualitative judgments into a scalar signal that the policy later optimises. Below, we summarise publicly documented approaches by two LLM companies: OpenAI and Google DeepMind.

OpenAI

- **GPT-3 and InstructGPT**
  Ouyang et al. (2022) create many pairwise comparisons by showing labelers four to nine model outputs at once and requesting a ranking. These rankings are broken down into ordered pairs, and the RM is trained with a **pairwise cross-entropy** loss. [13]

- **GPT-4 and later models**
  The *GPT-4 Technical Report* confirms that later models have moved on from relying solely on human feedback and started augmenting the human preferences with **Rule Based Reward Models (RBRMs)** that encode explicit safety heuristics [15]. Mu et al. (2024) formalise the objective as a **hinge loss** penalising outputs that violate those rules [20].

Google DeepMind

- **Sparrow**
  A paper by Glaese et al. (2022) details the RLHF process used by Google DeepMind's Sparrow - their approach consists of training two separate reward models. The first one - **Preference Reward Model** - is based on user's expressed preferences between

several possible responses. The preference RMs are reportedly **Bradley-Terry (Elo)** models. The second one - **Rule Reward Model** - is based on adversial probing, which uses a simple **cross-entropy loss**. [21]

4.3 RLHF pipeline: Policy optimisation stage

Once a reliable reward model is in place, the final stage fine-tunes the language model to maximise the learned reward. Below we summarise the publicly documented strategies adopted by two prominent organisations: DeepSeek and Anthropic.

DeepSeek

- **DeepSeek-R1-Zero and DeepSeek-R1**
  DeepSeek report using the **Group Relative Policy Optimization (GRPO)** algorithm for training. [22] Introduced by Shao et al. (2024), this method is a variant of a Proximal Policy Optimisation (PPO) algorithm aimed to save computational resources. It does this by estimating the baseline from group scores while foregoing the critic model. [23].

Anthropic

- **Early models**
  A paper by Bai et al. (2022), previously mentioned in the Human Feedback Collection section notes that Anthropic utilised a **Proximal Policy Optimisation (PPO)** algorithm [16], a policy gradient method introduced by Schulman et al (2017) [24].

5 Practices through the analytical lens

5.1 Hermeneutical injustices in human feedback collection

The human feedback collection stage is foundational to RLHF. It forms the basis on which the models are taught to align with human preferences. Therefore, this stage is also particularly vulnerable to encoding hermeneutical injustice, specifically when the feedback sources are limited in diversity or the feedback mechanisms restrict nuance. These risks are particularly important considering our target group of adults with ADHD, whose communication and interpretation styles differ from dominant norms.

Firstly, the demographics of human annotators raise concerns. For example, the early versions released by OpenAI were finetuned using feedback from only 40 contractors. Similarly, Anthropic's early work relied on a small crowdworker pool of US-based, master-qualified contractors. While these decisions ensure annotation quality, they also systematically exclude a wide range of lived experiences. This concerns the **Representation** desideratum due to the lack of efforts to include diverse human experiences.

Secondly, the feedback format restricts the ways in which annotators can express themselves. While InstructGPT used more expressive Likert scales and ranking systems, many other models relied on binary thumbs up/down ratings or a forced choice between two responses. For example, in case both provided responses contain hermeneutical injustice and the model uses a forced choice strategy, the user does not have a possibility to reject both outputs. Similarly, a long LLM output may contain accurate marginalised experiences along with inaccurate ones, but the thumbs up/down method does not allow for the user to articulate such nuanced concerns. This is particularly a concern for the ADHD community, which was already identified as preferring to communicate using more words. This concerns the **Flexibility** desideratum by excluding ADHD-typical communication traits.

5.2 Hermeneutical injustices in reward modelling

The second stage of the RLHF process is the reward modelling. Even if the process ensures enough representation of minority groups in the previous stage, this one may introduce hermeneutical injustices in different ways.

Firstly, the hinge loss used by OpenAI is of a thresholded nature, often underpinning an allowed vs. disallowed classification of content [20]. This raises a concern of hermeneutical injustice. A hinge-based safety model might, for example, block or heavily penalise content that includes certain keywords or phrases. This can disproportionately affect certain communities or ways of speaking. For instance, marginalized groups reclaiming or reusing terms that are flagged as slurs could find an LLM unwilling to discuss their issues, because the safety model has learned with a hard margin that those terms are unsafe. The system might reject or heavily filter the output that contains those words, even if the context is important to the user. This can particularly concern the **Authenticity** desideratum by possibly silencing certain terms, particularly those used by neurodivergent people.

5.3 Hermeneutical injustices in policy optimisation

Lastly, the policy optimisation stage is the final step of the RLHF process. While further stages tend to amplify hermeneutical injustices encoded in the previous stages, it is important to discuss this last stage separately to identify the precise points where hermeneutical injustice may be encoded.

The main danger of this stage is that PPO can suffer from mode collapse, where the policy converges to generating repetitive or homogeneous outputs that achieve high reward but lack diversity [25] [26]. This specifically relates to **Flexibility** and **Authenticity** desiderata in ways that are important for heterogeneous user groups such as adults with ADHD. For example, if the majority prefers concise, focused communication styles, the mode collapse phenomenon will cause the LLM to produce only such responses, which systematically excludes ADHD-typical communication styles, such as previously identified tendencies of adults with ADHD to convey a narrative in a way that uses many words.

6 Responsible Research

This research was conducted with awareness of ethical responsibilities and limitations that are inherent to interpretative, literature-based work. This methodology has strengths in exploring intersections between philosophical concepts and technical processes, but it also carries some risks.

First key limitation is the focus on official documentation and publications from major AI developers. While this was a deliberate choice to ensure that the papers are grounded in real world applications of RLHF in LLMs, it introduces a potential bias by excluding contributions from small, less known organisations that were potentially missed during the literature gathering process. Although this decision ensures methodological clarity and reproducibility, it also limits the breadth of captured methods.

Secondly, it is important to acknowledge the limitations of hermeneutical injustice analysis. The literature on RLHF reviewed in this paper does not explicitly engage with the philosophical concept of hermeneutical injustice. Therefore, the connections made between RLHF practices and this concept are interpretive and should not be treated as a direct empirical finding, but rather as an exploratory, conceptual basis.

Additionally, while care was taken to include only peer-reviewed or reputable industry publications, it is important to note that the LLM industry is a rapidly evolving field, with many contributions being very recent. As a result, some of the sources cited in this study are first released on preprint platforms such as arXiv. Despite having possibly not undergone formal peer review processes at the time of writing, such sources were included due to their technical importance. However, extra care was taken in evaluating these sources, for instance by assessing citation practices and the quality of argumentation.

Finally, this work focuses on improving alignment between LLMs and users with an ADHD diagnosis. However, it does not assume that alignment improvements for this group will also benefit all others. There is an ethical risk that prioritising one communicative or cognitive style may unintentionally diminish performance or comfort for users whose styles differ significantly. While this trade-off is difficult to eliminate entirely, it reinforces the importance of inclusive design practices that can accommodate a wide range of user needs.

In conclusion, this research prioritises transparency and reproducibility. The literature selection process was explicitly described in section 3, all interpretative claims are situated separately and not presented as empirical conclusions.

7 Discussion

This study investigated how hermeneutical epistemic injustices that obscure marginalised lived experiences can become encoded in the technical process of Reinforcement Learning from Human Feedback in large language models. Using adults with ADHD as a case study, we examined how specific RLHF stages may suppress or distort neurodivergent communication styles. The analysis identified threats to three desiderata (Representation, Authenticity, and Flexibility), each of which is important for preserving the hermeneutical justice of marginalised groups. These findings are summarised in Table 1.

The human feedback stage was found to affect the **Representation** and **Flexibility** desiderata. Our findings align with Carik et al. [9], who documented that ADHD users report neurotypical biases in LLM responses. Limited diversity in feedback pools means that neurodivergent styles may be underrepresented. Moreover, constrained feedback formats (such as binary ratings or pairwise choices) limit the ability to capture nuanced reactions to model outputs, a critical flaw for users with more expressive, context-sensitive communication preferences. Casper et al. [4] also identify feedback bias as a RLHF limitation. Our findings build on this by showing how such bias is not only as a statistical skew, but also a risk for hermeneutical injustice, where entire ways of speaking and understanding are systematically excluded.

In the reward modelling stage, the **Authenticity** desideratum is most at risk. However, the final RLHF stage, policy optimisation, raises risks to both **Flexibility** and **Authenticity**. For ADHD users, who often express ideas with more verbosity or indirect structure [11], this leads to systematic filtering of their preferred style. Casper et al. again identify this risk through the lens of technical performance issues [4], whereas our analysis reframes it as a harm on hermeneutical justice. Additionally, the overlap of the **Authenticity** desideratum between both of the two latter stages implies that the adverse effects of these stages are

Stage	Common Practices	Affected Desiderata
1. Human Feedback	• Likert scale • Thumbs up/down rating • Binary preference • Carefully selected contractors • Users from 193 countries	• Representation • Flexibility
2. Reward Modelling	• Pairwise cross-entropy loss • Hinge loss • Bradley-Terry (Elo) model	• Authenticity
3. Policy Optimisation	• Group Relative Policy Optimization (GRPO) • Proximal Policy Optimisation (PPO)	• Authenticity • Flexibility

Table 1: How each RLHF stage risks encoding hermeneutical injustice through common practices.

similar. Thus, the limitations of both of these stages can also explain the complaints by the ADHD community identified by Carik et al., where the users claim that the LLM responses struggled to maintain the users' authentic voice [9].

Overall, this research extends existing work on fairness and alignment by introducing hermeneutical injustice as a lens for evaluating RLHF stages. It demonstrates that some design choices such as loss functions, feedback format, or optimiser type can indirectly dictate the narrative, possibly excluding certain minority groups.

8 Conclusions and Future Work

This research investigated how hermeneutical injustice can become encoded in the RLHF processes of large language models, using ADHD as a case study. By examining real-world RLHF implementations and analysing them through the lens of three desiderata (Representation, Flexibility, and Authenticity) this study identified several conceptual risks that could systematically marginalize ADHD-typical communication styles.

The main conclusion is that the RLHF process is not epistemically neutral. Each stage (human feedback, reward modeling, and policy optimization) can introduce or amplify hermeneutical injustice, depending on how data is collected, processed, and optimized. Specifically:

- The feedback collection stage often relies on limited annotator pools and limiting rating formats, undermining Representation and Flexibility.
- The reward modeling stage introduces Authenticity risks, especially through hinge loss functions that penalize specific types of language.
- The policy optimization stage, through mechanisms like PPO, can result in mode collapse that disproportionately filters out ADHD-typical expression, affecting both Flexibility and Authenticity.

While this study is conceptual and based on literature review, it identifies opportunities for future work. For instance, allowing open text or nuanced feedback during human feedback collection could capture marginalised voices more effectively. Experimenting with different loss functions or combinations of them is another important step into reducing hermeneutical injustice in practice.

Finally, our findings underscore the need for interdisciplinary collaboration in LLM development. Philosophical frameworks such as Fricker's epistemic injustice [2] and empirical insights from ADHD and neurodivergence research [10, 11, 9] provide a richer and more just foundation for model alignment. After all, hermeneutical justice should not be seen as a philosophical add-on, but rather as a core requirement in responsible LLM development.

References

[1] M. Steyvers, H. Tejeda, A. Kumar, C. Belem, S. Karny, X. Hu, L.W. Mayer, and P. Smyth. What large language models know and what people think they know. 7(2):221–231.

[2] Miranda Fricker. Hermeneutical injustice. In Miranda Fricker, editor, *Epistemic Injustice: Power and the Ethics of Knowing*, page 0. Oxford University Press.

[3] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey.

[4] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback, 2023.

[5] Jackie Kay, Atoosa Kasirzadeh, and Shakir Mohamed. Epistemic injustice in generative AI. 7(1):684–697. Number: 1.

[6] S. Weibel, O. Menard, A. Ionita, M. Boumendjel, C. Cabelguen, C. Kraemer, J.-A. Micoulaud-Franchi, S. Bioulac, N. Perroud, A. Sauvaget, L. Carton, M. Gachet, and R. Lopez. Practical considerations for the evaluation and management of attention deficit hyperactivity disorder (adhd) in adults. *L'Encéphale*, 46(1):30–40, 2020.

[7] Ylva Ginsberg, Javier Quintero, Ernie Anand, Marta Casillas, and Himanshu P. Upadhyaya. Underdiagnosis of attention-deficit/hyperactivity disorder in adult patients: A review of the literature. *Primary Care Companion for CNS Disorders*, 16(3):23591, 2014.

[8] Elena Even-Simkin. Assessment of pragmatic skills in adults with adhd. *Language and Health*, 2(1):66–78, 2024.

[9] Buse Carik, Kaike Ping, Xiaohan Ding, and Eugenia H. Rho. Exploring large language models through a neurodivergent lens: Use, challenges, community-driven workarounds, and concerns. *Proc. ACM Hum.-Comput. Interact.*, 9(1), January 2025.

[10] Sustained attention in adult adhd: time-on-task effects of various measures of attention. *Journal of Neural Transmission*, 124(1):39–53, February 2017.

[11] Rafael Martins, Cláudia Drummond, Natália Mota, Pilar Erthal, Gabriel Bernardes Pacheco de Moraes, Gabriel Lima, Raquel Molina, Felipe Sudo, Rosemary Tannock, and Paulo Mattos. Network analysis of narrative discourse and attention-deficit hyperactivity symptoms in adults. *PLOS ONE*, 16:e0245113, 04 2021.

[12] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback, 2024.

[13] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[14] Kristian González Barman, Simon Lohse, and Henk W. de Regt. Reinforcement learning from human feedback in LLMs: Whose culture, whose values, whose perspectives? 38(2):35.

[15] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan,

Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[16] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark,

Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

[17] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022.

[18] Anthropic. Model card and evaluations for claude models, 2023. Accessed: 2025-06-02.

[19] Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. Accessed: 2025-06-02.

[20] Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. Rule based rewards for language model safety, 2024.

[21] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements, 2022.

[22] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun

Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[23] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

[24] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. 07 2017.

[25] Yufeng Yuan, Yu Yue, Ruofei Zhu, Tiantian Fan, and Lin Yan. What's behind ppo's collapse in long-cot? value optimization holds the secret, 2025.

[26] Banghua Zhu, Hiteshi Sharma, Felipe Vieira Frujeri, Shi Dong, Chenguang Zhu, Michael I. Jordan, and Jiantao Jiao. Fine-tuning language models with advantage-induced policy alignment, 2023.