# How are hermeneutical injustices encoded in Reinforcement Learning from Human Feedback (RLHF)?

Author: Ieva Mockaitytė (i.mockaityte@student.tudelft.nl) Supervisor: Anne Arzberger Responsible professor: Jie Yang

## 1. Background and introduction

Large language models (LLMs) are increasingly integrated into everyday tools, including but not limited to search engines and messaging apps. Such increasing influence raises important ethical considerations regarding their outputs. While biases and fairness in LLMs are widely explored, less attention was given to possible hermeneutical injustice, especially in alignment processes such as RLHF.

Keywords:

- **Hermeneutical injustice** - "the injustice of having some significant area of one's social experience obscured from collective understanding owing to hermeneutical marginalization" [1]
- **Reinforcement Learning from Human Feedback (RLHF)** - a process of collecting feedback from humans in order to align LLM outputs to their preferences.

## 2. Research question

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

## 3. Methodology

- Design
  Semi-structured qualitative literature review examining how RLHF pipelines in large language models may embed hermeneutical injustice toward adults with ADHD.
- Search and selection
  - PRISMA-inspired screening.
  - **Include:** Peer-reviewed or organisational publications that describe at least one RLHF stage: Human feedback collection; Reward modelling or Policy optimisation.
  - **Exclude:** non-RLHF ethics papers or purely theoretical RLHF proposals.
- Analytical lens
  For each RLHF stage, mapped technical practices against three desiderata that guard against hermeneutical injustice:
  - **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?

## 4 Target group

Despite making up 2-4% of the adult population [2], people with ADHD are frequently misdiagnosed and do not receive proper treatment, [3] which leads to their different experiences being widely misunderstood. Key differences to consider include:

- **Differences in information processing** - Attention of adults with ADHD tends to deteriorate over time faster compared to neurotypical individuals. [4]
- **Differences in information conveying** - Adults with ADHD tend to use more words and a less coherent structure to convey a story. [5]

A 2025 study on neurodivergent users' interaction with LLMs identified common complaints and concerns expressed by people with ADHD, of which the most relevant to this study are:

- **Prompting** - difficulties phrasing prompts in ways that yield helpful responses.
- **Biased responses** - responses failing to capture neurodivergent thought processes.
- **Lack of authentic voice** - having to rephrase LLM responses to preserve their own voice. [6]

## 5.1 Hermeneutical injustices in human feedback collection

The tables below show the prominent practices of human feedback collection in the RLHF fine-tuning processes. The practical applications that were found to affect one or more of the desiderata are highlighted, and further explanation is provided below.

The papers collected during this stage of analysis include: [7, 8, 9, 10, 11, 12, 13]

| Feedback methods |
|---|
| Likert scale ratings (1–7) |
| Binary thumbs up/down |
| Binary preference |
| Response ranking |

| Feedback pools |
|---|
| 40 selected contractors |
| Users from 193 countries |
| Over 50 experts |
| US-based, master-qualified crowdworkers |

- **Representation** - the lack of efforts to include diverse human experiences.
- **Flexibility** - the exclusion of ADHD-typical communication traits.

## 5.2 Hermeneutical injustices in reward modelling

The table below shows the prominent reward modelling methods. The practical applications that were found to affect one or more of the desiderata are highlighted, and further explanation is provided below.

The papers collected during this stage of analysis include: [7, 9, 14, 15]

| Dominant Reward Modelling Practices |
|---|
| Pairwise preference modeling using cross-entropy loss |
| Rule-Based Reward Models (RBRMs) using hinge loss |
| Use of Bradley-Terry (Elo) models |

- **Authenticity** - The possible silencing of certain terms, particularly those used by neurodivergent people.

## 5.3 Hermeneutical injustices in policy optimization

The table below shows the prominent policy optimisation techniques. The practical applications that were found to affect one or more of the desiderata are highlighted, and further explanation is provided below.

The papers collected during this stage of analysis include: [16, 17, 18, 11]

| Dominant Policy Optimisation Practices |
|---|
| Group Relative Policy Optimization (GRPO) |
| Proximal Policy Optimisation (PPO) |

- **Authenticity** - The mode collapse phenomenon known to happen during PPO can lead to loss of the prompter's authentic voice.
- **Flexibility** - Similarly, due to difficulties of handling different communication styles, the mode collapse phenomenon could make LLMs generate homogeneous outputs that achieve high reward but lack diversity.

## 6. Conclusion

**The RLHF process is not hermeneutically epistemically neutral.** Hermeneutical injustice can get encoded in each of the stages of RLHF, even if care was taken to reduce it in previous stages. For example, even if the percentage of human evaluators diagnosed with ADHD roughly corresponds to the general population of adults diagnosed with ADHD, these individuals still represent a minority among evaluators, and their input may therefore be overwhelmed or suppressed by mode collapse effects.

## 6. Future recommendations

The first important step is to attempt to generalise these findings to other marginalised groups with unique needs - it is important to note that making an LLM accessible to a specific target group does not necessarily lead to improved experiences for all users.

- **Human feedback collection**. Conducting case studies from a significantly large group of adults with ADHD to find agreement and disagreement points.
- **Reward modelling**. Experimenting with combinations of different loss functions and different reward models.
- **Policy optimisation**. Building upon improving the dominant PPO method, drawing on proposed theoretical improvements.

Finally, our findings underscore the need for **interdisciplinary collaboration in LLM development**. Philosophical frameworks such as Fricker's epistemic injustice [1] and empirical insights from ADHD and neurodivergence research 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] Miranda Fricker. Hermeneutical injustice. In Miranda Fricker, editor, *Epistemic Injustice: Power and the Ethics of Knowing*, page 0. Oxford University Press.

[2] Weibel et al. Practical considerations for the evaluation and management of attention deficit hyperactivity disorder (adhd) in adults. *L'Encéphale*, 46(1):30–40, 2020.

[3] Ginsberg et al. Underdiagnosis of attention-deficit/hyperactivity disorder in adult patients: A review of the literature. *Primary Care Companion for CNS Disorders*, 16(3):23591, 2014.

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

[5] Martinset al. Network analysis of narrative discourse and attention-deficit hyperactivity symptoms in adults. *PLOS ONE*, 16:e0245113, 04 2021.

[6] 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.

[7] Ouyang et al. Training language models to follow instructions with human feedback, 2022.

[8] 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.

[9] OpenAI et al. Gpt-4 technical report, 2024.

[10] Yuntao Bai et al. Constitutional ai: Harmlessness from ai feedback, 2022.

[11] Yuntao Bai et al. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

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

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

[14] Mu et al. Rule based rewards for language model safety, 2024.

[15] Amelia Glaese et al. Improving alignment of dialogue agents via targeted human judgements, 2022.

[16] DeepSeek-AI et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[17] Shao et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

[18] Schulman et al. Proximal policy optimization algorithms. 07 2017.

[19] et al. Kay. Epistemic injustice in generative AI. 7(1):684–697. Number: 1.

[20] Casper et al. Open problems and fundamental limitations of reinforcement learning from human feedback, 2023.

[21] Steyvers et al. What large language models know and what people think they know. 7(2):221–231.

[22] Gallegos et al. Bias and fairness in large language models: A survey.

[23] Zheng et al. Secrets of rlhf in large language models part i: Ppo, 2023.

[24] Wang et al. Secrets of rlhf in large language models part ii: Reward modeling, 2024.

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