Enhancing RLHF with Human Gaze Modeling

Karim Galliamov,[1] Ivan Titov,[2] and Ilya Pershin[1]

[1]Research Center of the Artificial Intelligence Institute, Innopolis University, Innopolis, Russia
[2]University of Edinburgh, United Kingdom

k.galliamov@innopolis.university, ititov@inf.ed.ac.uk, i.pershin@innopolis.ru

Abstract

Reinforcement Learning from Human Feedback (RLHF) aligns language models with human preferences but is computationally expensive. We explore two approaches that leverage human gaze modeling to enhance RLHF: (1) gaze-aware reward models and (2) gaze-based distribution of sparse rewards at token level. Our experiments demonstate that gaze-informed RLHF achieves faster convergence while maintaining or slightly improving performance, thus, reducing computational costs during policy optimization. These results show that human gaze provides a valuable and under-used signal for policy optimization, pointing to a promising direction for improving RLHF efficiency.

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) has emerged as a powerful paradigm for aligning language models with human values and preferences (Ziegler et al., 2019), typically involving training a reward model on human preference data, followed by policy optimization algorithms such as PPO (Schulman et al., 2017). The process demands substantial computational resources and time, as it typically requires a large number of optimization steps to reach convergence. This inefficiency is driven in part by a fundamental challenge in RLHF: the sparsity of feedback signals (Chan et al., 2024). Reward models typically provide a single scalar value for an entire sequence, offering limited guidance on which specific parts of a model's output contribute positively or negatively to human preferences.

Human visual attention – as captured by eye tracking and gaze patterns – offers fine-grained insights into where cognitive effort is directed during text processing (Kleinke, 1986). When skimming through text, humans naturally fixate on segments that are particularly informative, challenging, or salient. Incorporating gaze information into RLHF may help identify which parts of a sequence drive human judgments, thus improving the efficiency policy training.

We leverage a gaze prediction model trained on existing eye-tracking corpora and investigate two alternative approaches for integrating its predictions into the RLHF pipeline. To our knowledge, this is the first work to explore the use of human gaze modeling within the RLHF framework:

- First, we evaluate the application of gaze-informed reward models, introduced in López-Cardona et al. (2025). These models have demonstrated improved accuracy in preference ranking – i.e. ranking outputs according to human preferences – but have not yet been evaluated within RLHF pipelines. Our contribution is to show that incorporating these reward models into PPO and GRPO frameworks (Shao et al., 2024) leads to faster convergence. We refer to this approach as *GazeRM*.

- Second, we propose *GazeDistrib*. Inspired by prior work that leverages reward model attention for reward attribution (Chan et al., 2024), our approach instead relies on predicted human gaze to assign token-level credit in RLHF. By aligning rewards with human attention, we aim to provide more targeted feedback, accelerating convergence and potentially improving performance. Notably, *GazeDistrib* requires no changes to the reward model itself – only to how its scores are distributed.

Our experiments with both PPO and GRPO demonstrate that gaze-informed RLHF consistently achieves faster convergence—by a factor of 1.5-2 — compared to standard approaches. This acceleration occurs while maintaining or slightly improving final performance, suggesting that human visual attention patterns provide valuable signals for more

(a) Validation scores gained by policy in PPO


(b) Training rewards gained by policy in PPO

Figure 1: Performance of LLaMa-7B-open-instruct models trained with and without gaze information on HH-RLHF

efficient learning.

The primary contributions of this work are:

- integrating human gaze prediction models into RLHF pipelines through two alternative methods;
- empirical evidence showing significant convergence acceleration across two online-RL algorithms;
- analysis of gaze-informed reward distribution effects on the learning process.[1]

2 Background and Related Work

2.1 Gaze Prediction Models

Collecting human gaze data is time-consuming and costly. A practical alternative is to train a model on existing eye-tracking corpora (Cop et al., 2016; Hollenstein et al., 2018) and use it to predict gaze patterns on new data. Such models predict various token-level features including First Fixation Duration (how long a reader initially fixates on a token), Go-Past Time (total time spent from first encountering a token until moving past it), Total Reading Time (cumulative duration of all fixations on a token), and number of fixations (nFix). The predictions serve as proxies for actual human visual attention patterns.

---

[1]The code is available at https://github.com/KGallyamov/Gaze_RLHF

2.2 Gaze-Informed Reward Models

López-Cardona et al. (2025) proposed integrating gaze features into reward modeling by projecting them into the model's hidden representation space using a trainable feedforward neural network. These gaze feature projections are either added to the first layer's hidden activations or concatenated.

Their experiments demonstrated substantial improvements in preference prediction accuracy with both integration methods. However, their work did not extend to evaluating how these gaze-enhanced reward models perform when incorporated into complete RLHF training pipelines.

2.3 Dense Reward Distribution from Model Attention

Previous work by Chan et al. (2024) explored an alternative to sparse sequence-level rewards by proposing a method to distribute rewards at the token level using the attention patterns from a reward model's final layer. Their approach transforms the single scalar reward into a dense, token-specific signal that provides more granular feedback during training. Their experiments demonstrated that such dense reward distribution leads to faster convergence compared to traditional sparse reward approaches, suggesting that more fine-grained feedback signals can improve the efficiency of policy optimization. However, this method has no explicit connection to human attention.

3 Our approach

3.1 Human attention model analysis

To get an intuition for why modeling human gaze may improve training dynamics, we analyze how gaze is distributed across different parts of speech (PoS). We applied PoS tagging using FLAIR (Akbik et al., 2018) to obtain grammatical categories for each word. Since our gaze prediction model operates on subword tokens while PoS tags are assigned at the word level, we computed word-level attention scores by summing the predicted attention values across all subtokens comprising each word.

In Table 1, we report the mean human attention scores by part of speech (PoS), using coarser categories derived from the Penn Treebank tags used by FLAIR. We computed word-level human attention scores for each instruction in OASST2, applied PoS tagging, and aggregated the scores by grammatical category to obtain the mean attention value for each PoS across all word instances.

Overall, the human gaze model clearly attends to content-rich words (especially nouns and verbs), as well as punctuation. One possible explanation for *GazeDistrib*'s faster convergence is its strategy of assigning higher rewards to tokens that carry more information. By prioritizing these key elements, the model can more efficiently capture underlying patterns and structures, leading to faster and more effective training.

POS Tags	Meaning	Gaze
TO	to	0.0122
NFP	Superfluous punctuation	0.0246
AFX	Affix	0.0282
SYM	Symbol	0.0291
CC	Coordinating conjunction	0.0318
EX	Existential there	0.0331
XX	Unknown	0.0369
DT	Determiner	0.0376
IN	Preposition or conjunction	0.0386
FW	Foreign word	0.0397
PRP	Personal pronoun	0.0402
LS	List item marker	0.0419
MD	Modal	0.0422
WDT	Wh-determiner	0.0437
CD	Cardinal number	0.0442
UH	Interjection	0.0578
WRB	Wh-adverb	0.0620
WP	Wh-pronoun	0.0748
WP$, PRP$	Possessive (wh-)pronouns	0.0854
. , | , | HYPH	Punctuation marks	0.1316
JJ*	Adjectives	0.1355
RB*	Adverbs	0.1466
NN*	Nouns	0.2295
VB*	Verbs and gerunds	0.2697
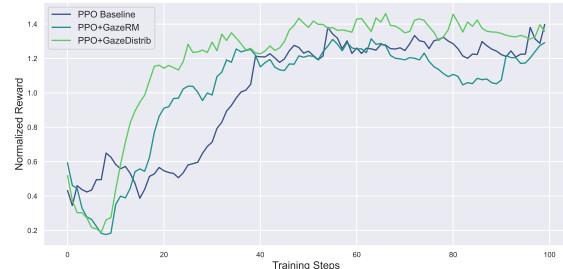
Table 1: Part-of-Speech Tags Gaze Scores

We investigate two alternative approaches to integrating human gaze information into RLHF pipelines: enhancing reward models with explicit gaze features and using human gaze patterns to redistribute sparse sequence-level rewards across individual tokens.
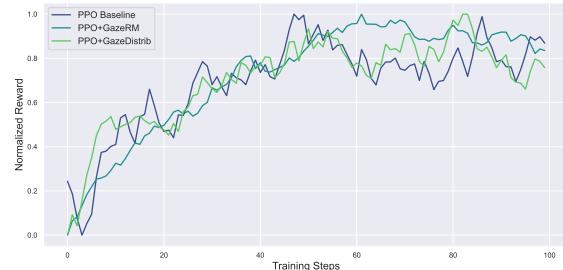
3.2 Gaze-Informed Reward Models

Building upon previous work by López-Cardona et al. (2025), the human eye tracking data is accomplished by projecting the predicted gaze features into the hidden space of a reward model. These projections are either added or concatenated directly into the first layer token embeddings of the reward model.

López-Cardona et al. (2025) only evaluated their models on reward benchmarks and validation preference datasets. While they observed improved performance compared to gaze-agnostic baseline reward models, they did not integrate their reward models into the RLHF pipeline. Our work examines whether these improvements translate into

(a) Validation scores gained by policy in PPO

(b) Training rewards gained by policy in PPO

Figure 2: Performance of Mistral-7B trained with and without gaze information

more efficient and effective policy optimization when deployed within the complete RLHF framework. Our central hypothesis is that a more precise and informative reward signal derived from human attention patterns will lead to more targeted and efficient updates during the policy optimization phase, thereby accelerating convergence and potentially improving the quality of the final policy. Samples of predicted human gaze could be found in Appendix C.

3.3 Dense reward from human gaze

Traditional RLHF methods commonly assign a single scalar reward to an entire generated sequence, providing a very sparse signal that offers limited guidance on which specific tokens or spans contributed positively or negatively to the overall human judgement. We address this sparsity problem by leveraging human gaze predictions to infer token-level importance and distribute rewards accordingly.

In our implementation, we use Total Reading Time (TRT) as the primary gaze metric for distributing rewards at the token level. TRT measures the cumulative duration a reader fixates on each token, including both initial reading and any subsequent rereadings. This makes it a strong proxy for a token's processing load and its importance in human evaluation. By levereging TRT, we can derive token-level dense rewards using readily available gaze prediction models, offering a more fine-

grained signal for training.

Given a predicted TRT distribution $T = \{t_1, t_2, ..., t_n\}$ for a sequence of n tokens and a scalar reward R for the entire sequence (obtained from standard human preference judgments), we compute token-level rewards r_i using a softmax-like distribution over the gaze values:

$$r_i = R \cdot \frac{e^{t_i}}{\sum_{j=1}^{n} e^{t_j}} \quad (1)$$

This approach (*GazeDistrib*) ensures that tokens receiving more visual attention (higher TRT) from humans are assigned proportionally larger portions of the overall sequence reward R. This process creates a dense reward signal that is aligned with human attention and cognitive effort, providing more precise and granular guidance to the model optimization process compared to a single sequence-level scalar.

Crucially, since the used gaze predicting models are significantly smaller (up to 125M parameters) compared to the backbone models used for both the reward model and the trained policy (7B parameters), their inference cost and storage requirements are negligible and do not represent a bottleneck in either of the proposed integration scenarios.

4 Experiments

We train our models on HH-RLHF (Bai et al., 2022). This is a dataset provided by Anthropic, totaling 161K human annotated chosen-rejected dialog pairs. We also run our experiments on OASST-2 (Köpf et al., 2023) - another preference pairs dataset, with 64K prompts. For RLHF alignment, only instruction prompts are required — not full dialogues — so we truncate each example before the model's final response. Following López-Cardona et al. (2025), we filter out all non-English examples from the dataset, as the gaze predicting model - and thus the gaze-informed reward model - was only trained on English corpora. To demonstrate the robustness of our approach, we integrate both our methods with both PPO and GRPO methods. The implementation of these methods is taken from TRL (von Werra et al., 2020).

4.1 Gaze Modeling

In our work, we used gaze models from Cop et al. (2016) and Hollenstein et al. (2018). The models are trained on large corpora with gaze annotations; the token-level regression tasks involved predicting

several variables such as First Fixation Duration, Go-Past Time, and Total Reading Time. After training, the models are frozen. During our experiments, we use these models to infer gaze features, which are then used as scores for dense reward on policy model's output generations. Other way of gaze features utilization is to train the reward models after having these gaze features computed for the preference corpora as in (López-Cardona et al., 2025).

4.2 Reward models

We reproduce the work of López-Cardona et al. (2025) by training Mistral 7B (Jiang, 2024) for reward modeling with concatenation of gaze projected features to the input token embeddings (setup referred to as $f_{comb2.5}$ in the original paper).

4.3 Evaluation

Since our experiments involve policies trained with different reward models, directly comparing their final reward values would be misleading. To enable meaningful cross-training comparisons, we utilized a "hold-out" reward model - the one not used in any policy training process - to evaluate all policies on a consistent benchmark. We use OpenAssistant's deberta-large-v2 (Köpf et al., 2023) reward model as this validating model. The selected reward model achieves accuracy of 69.25 on HH-RLHF validation split thus making it quite a strong model for comparing models on that dataset. This hold-out model serves two purposes: it provides comparable scores across different training runs and validates that any observed improvements in convergence speed reflect genuine learning rather than simply exploiting weaknesses in specific reward schemes, i.e. reward hacking. We define a policy's validation score as the difference between its hold-out reward score and the score of the initial SFT model. For further analysis, we also plot min-max normalized plots.

All plots and tables show averages over 3 runs with different seeds to mitigate the stochasticity in policy optimization.

5 Results

Figure 1 shows the convergence behavior of PPO with and without gaze-based rewards across both model architectures. The results consistently demonstrate significant speed-up in convergence: validation scores grow much faster for gaze-aware approaches in both LLaMa and Mistral experi-

Method	LLaMa (HH-RLHF)		Mistral (OASST2)	
	Val. Score	Steps to Conv.	Val. Score	Steps to Conv.
Baseline	1.09 ± 0.08	41.40 ± 0.21	1.21 ± 0.15	53.40 ± 0.29
GazeRM	1.19 ± 0.10	28.33 ± 0.54	1.10 ± 0.31	29.20 ± 0.95
GazeDistrib	1.22 ± 0.12	22.33 ± 1.02	1.16 ± 0.25	30.60 ± 0.73

Table 2: Convergence and quality comparison on PPO across different model architectures and datasets

ments. However, they ultimately reach similar performance levels, suggesting that gaze-aware methods primarily accelerate training rather than improve final model quality.

The training reward behavior varies across different SFT initializations. For one policy initialization, training rewards closely resemble validation scores 1b, while for another initialization method 2b, there is no significant difference in training rewards between methods. This inconsistency in min-max normalized reward model comparisons confirms that tracking validation scores provides more reliable insights.

To demonstrate that our approaches are invariant to policy optimization algorithms, we repeated the PPO experiments using GRPO (Shao et al., 2024), observing similar convergence speed improvements of approximately 1.3 times faster with gaze-guided methods. The consistency across different model architectures (LLaMa and Mistral) and datasets (HH-RLHF and OASST2) validates the generalizability of our approach. When using Mistral, we aligned the reward model backbone with the corresponding architecture to ensure tokenizer compatibility, which is crucial for dense reward training. For additional details on GRPO experiments, refer to Appendix B.

6 Conclusion

In this paper, we have presented a novel method of integrating human gaze data into RLHF through dense reward attribution guided by human attention patterns (GazeDistrib). Additionally, we validated that existing gaze-informed reward models from prior work, when incorporated into PPO and GRPO frameworks, provide substantial benefits to the training process. Our experiments demonstrate that both approaches accelerate convergence during policy optimization by a factor of 1.5-2 times without compromising final model performance.

The results indicate that human visual attention patterns provide valuable signals that enhance alignment efficiency. By addressing the sparse reward problem through more targeted feedback, our gaze-based approaches offer a promising direction for improving computational efficiency in RLHF.

Limitations

In this work, we have experimented only with on-policy RLHF methods. Further experiments with off-policy methods, such as DPO, are required. Our study was limited to two datasets initialization and performed only on English. Future work should extend these experiments to additional language, datasets, and language models.

Acknowledgment

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

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, and 12 others. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv e-prints*, arXiv:2204.05862.

Alex J Chan, Hao Sun, Samuel Holt, and Mihaela Van Der Schaar. 2024. Dense reward for free in reinforcement learning from human feedback. *arXiv preprint arXiv:2402.00782*.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2016. Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific Data*, 5.

Fengqing Jiang. 2024. Identifying and mitigating vulnerabilities in llm-integrated applications. Master's thesis, University of Washington.

Chris Kleinke. 1986. Gaze and eye contact. a research review. *Psychological bulletin*, 100:78–100.

Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, and 1 others. 2023. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36:47669–47681.

Ángela López-Cardona, Carlos Segura, Alexandros Karatzoglou, Sergi Abadal, and Ioannis Arapakis. 2025. Seeing eye to AI: Human alignment via gaze-based response rewards for large language models. In *The Thirteenth International Conference on Learning Representations*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

## A Implementation details and hparams tuning
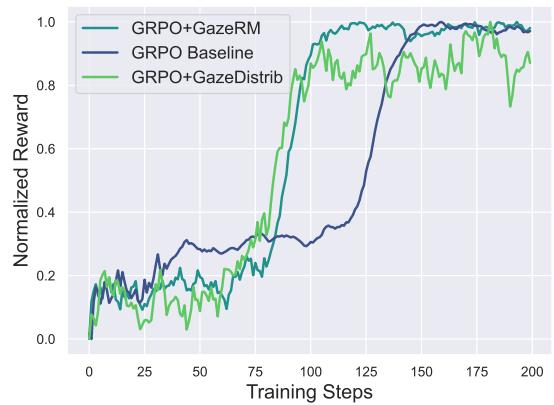
In this section, we list the setups used for PPO and GRPO experiments: SFT initialization checkpoints, hold-out reward models and datasets.

A.1

Dataset: Anthropic/hh-rlhf

Holdout Reward Model: OpenAssistant/reward-model-deberta-v3-large-v2

Policy Model Initialization: VMware/open-llama-7b-open-instruct

(a) Training rewards gained by policy in GRPO

A.2

Dataset: OASST2

Holdout Reward Model: OpenAssistant/reward-model-deberta-v3-large-v2

Policy Model Initialization: Mistral-7B-v0.1

We run a sweep for learning rate and batch size identical to that described in (Chan et al., 2024).

B Transferring to GRPO

On 3a one may see that the convergence speed boost observed on PPO also transfers to GRPO - however, the gained reward graph differs from that observed in PPO

C Sample human gaze patterns
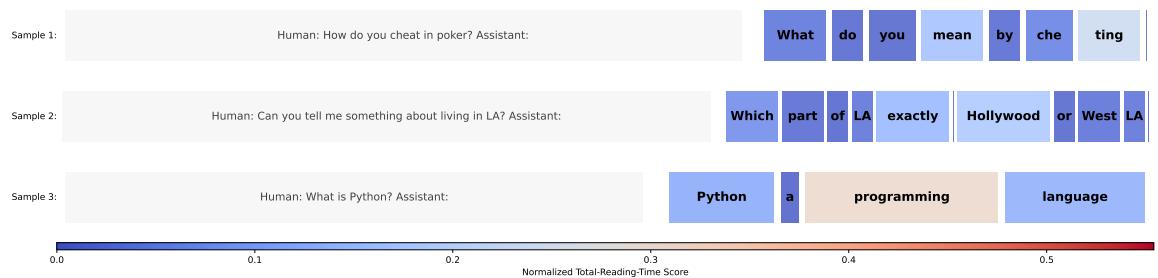
Figure 4: Samples of human gaze modeling